Bachelor Thesis

Estimates on the Chance of Further Improvement in Bayesian Optimization

Author:	Marc Philip Kaebel (381256)
Supervisor:	Prof. Dr. Hanno Gottschalk
Co-Supervisor:	Dr. Tobias Riedlinger

Technische Universität Berlin
Fakultät II
Institut für Mathematik

2024

Eidesstattliche Versicherung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Berlin, den 21.03.2024

Marc Philip Kaebel
Matrikelnummer: 381256

Kurzzusammenfassung

In dieser Arbeit beschäftigen wir uns mit dem Bayesian Optimization Verfahren. Dazu legen wir zunächst einige Grundlagen aus der Wahrscheinlichkeitstheorie. Wir leiten die Gaussian Process Regression her und zeigen einige ihrer bekannten Eigenschaften. Anschließend erkären wir, was Bayesian Optimization ist und leiten eine obere Schranke für die Wahrscheinlichkeit einer weiteren Verbesserung her. Mithilfe dieser Abschätzung schlagen wir eine Erweiterung der Bayesian Optimization vor, die auf dem Branch and Bound Algorithmus basiert.

Contents

Chapter 1

Introduction

1.1 Topic of this Thesis

Let us start by looking at a generic optimization problem. That is, we have some function $f\colon \mathcal{X} \to \mathbb{R}$ on a suitable space $\mathcal{X}$, and we want to find a minimizer

$$x^* \in \operatorname*{argmin}_{x \in \mathcal{X}} f(x).$$

An example of such a function is shown in Figure 1.1. One simple way to approach this problem would be to start checking points randomly and remembering the lowest point we encountered. For finite spaces $\mathcal{X}$ this is sometimes feasible, but quickly runs into limitations if the number of elements gets too large. An example of a more refined strategy would be to use information of a lower bound on the function, if we have one. This lets us rule out areas of search, where the lower bound is already higher than a value we have previously encountered, and subsequently only look in areas, in which improvement is possible. This method is called branch and bound [22]. Although this is an improvement on the previous method, it still does not guarantee, that we will find the global minimum. But it does show a common theme in optimization of using all the availaible information about the space and the objective function to our benefit. Is the function differentiable? Linear? Is the space discrete? Is it continous, but still finite dimensional? These, among many others, are restrictions with specific approaches taylored to them. The approach discussed in this thesis is no different.

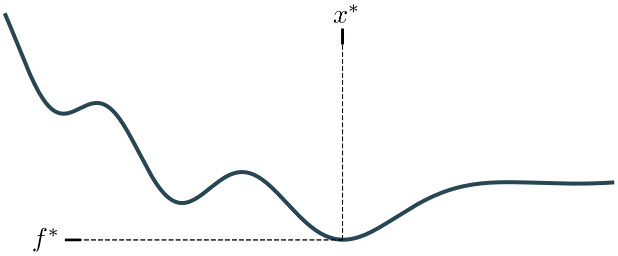

Figure 1.1 | A first optimization example.

The setup is as follows: We have a function that is possibly expensive to evaluate and behaves randomly, but the structure of that randomness can be well quantified locally. "Structure of randomness?" This is most easily explained with a picture. The data for the three pictures in Figure 1.2 was randomly generated by three different random processes. They are all random, but don't share the same structure. One could generate many pictures from the first process resembling the first picture in structure, but none would look like any, that were created by the second or third process. A small spoiler of Section 2.3 reveals that the difference in structure is determined by how much points in space are influenced by their neighboring points.

Gaussian Process Regression

Say we now have a random process to model the structure of our optimization problem, more specifically a Gaussian process. It is characterized by a mean and covariance function quantifying the random structure. Possible realizations of such a process are shown in Figure 1.3. In Bayesian terms, this model is the prior, so the likelihood of samples before incorporating any measurement data. How do we ensure, that the random process meets the measured values at the points of observation? For those points, how do we make it not random? This is where the regression framework comes in. We can construct another random process, one that is an optimal predictor of the underlying process, conditional on the measured values. It is constructed as a weighted linear combination of the measured values, with a covariance function that becomes zero at points of measurement. In Bayesian terms, this is the posterior. In this way we can frame our beliefs as a prior distribution and update our model by conditioning on observations, thus arriving at a posterior distribution like Figure 1.4. This model is often referred

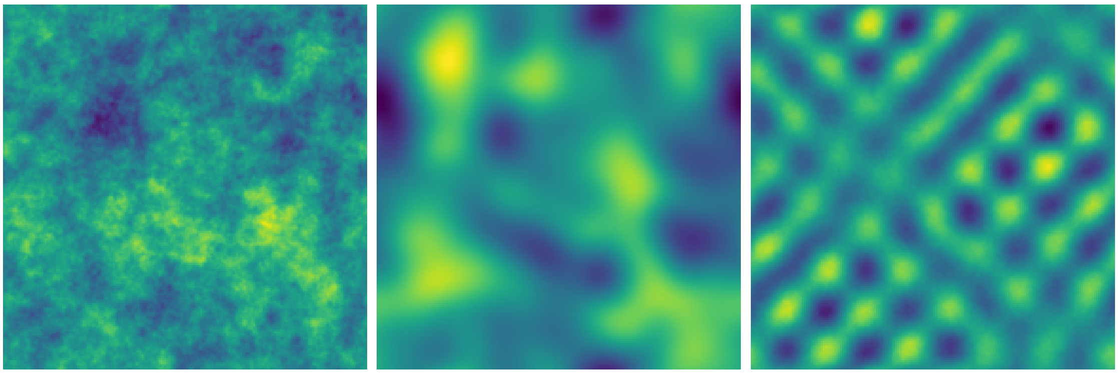

Figure 1.2 | Samples drawn from random fields with three different covariance functions.

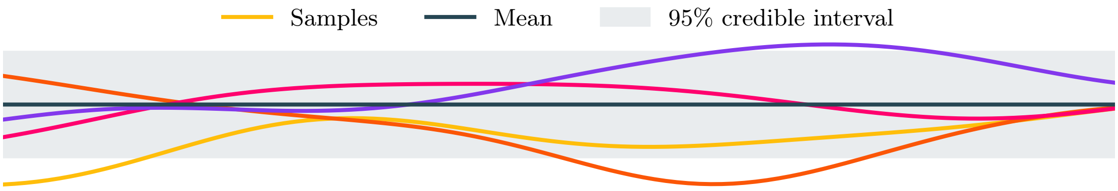

Figure 1.3 | Gaussian process regression prior.

to as a surrogate model. One benefit of choosing Gaussian processes is that such a construction is then also a Gaussian process.

Problems like this first arose in the area of geostatistics [26]. When mining gold it is of great interest to find out where the greatest probability of a dense deposit of gold is. Due to the geological structure, i.e. porosity of the ground and other variables, the spatial distibution looks inherently random. In this setting, the random density of gold at points is partially determined by the density of gold in the proximity. The governing principle is, that things that are close together are similar. Given a measurment of high gold concentration at a certain point, there is a high probability of having a similar concentration 5 meters away. But with 50 meters distance there is less certainty. It is therefore reasonable to search for an optimal place to mine by not just relying on the observations, but also utilizing the mean and covariance functions generated by the Gaussian process regression, like in Figure 1.4.

Beyond geostatistics, this model of Gaussian process regression allows for enough abstraction to be useful in many applications. A more recent example is design problems. Designing molecules for a new drug [19], designing structural metal parts for a new airplane wing [29], or setting hyperparameters in machine learning applications [40]. In these contexts, the spaces of optimization are parameter spaces, with parameters like material thickness and diameter. An observation amounts to running a simulation with a set of parameters and seeing how well they perform given a certain objective function accounting for things like structural stability, weight, material cost. Whenever observation is expensive and we have information about how things that are close together are similar, Gaussian processes become a viable option.

Bayesian Optimization

An observant reader might have already noticed that the observation points just appeared without much consideration. Realistically, one can often pick them sequentially in an optimal manner. This leads to a scheme called Bayesian optimization. Instead of calculating one Gaussian process regression, Bayesian optimization involves building a new Gaussian process surrogate model after each measurement.

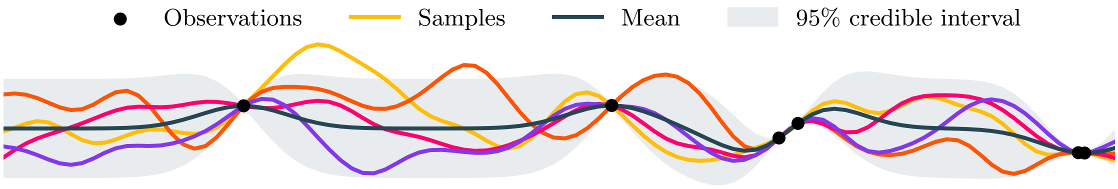

Figure 1.4 | Gaussian process regression with observations.

And then using this model to pick the next observation point. This is repeated until some stopping criterion is met. When picking an observation point, there is a trade-off between exploring areas of big uncertainty and exploiting a known low value and searching in its immediate neighborhood for a lower value. This trade-off can be weighted differently with different objective functions, also called acquisition functions. Thus leading to different optimization schemes, some of which we will discuss in Section 3.2.

Chance of Further Improvement in Bayesian Optimization

The primary contribution of this thesis is the derivation of an upper bound on the chance of further improvement in a given area during Bayesian optimization. In order to derive such a bound, we will use chaining tools. Roughly speaking these are tools that use the metric entropy of the data generated by the random process to show that it has certain regularity properties. With this bound on the chance of further improvement, one can stop searching for solutions in areas where further improvement is unlikely. We will show an example of an implementation in an algorithm similar to the branch and bound algorithm briefly mentioned earlier.

1.2 Related work

The review papers [14, 30] give a good overview of recent work in the area of Bayesian optimization. Some worthwhile mentions are the works of [3, 32], which use an upper confidence bound as an acquisition function for Bayesian optimization. The works by [27, 38] build on this to create an optimistic optimization algorithm, which can be extended into a branch and bound scheme [12]. Also worth mentioning is [10], where chaining techniques were used for Bayesian optimization.

1.3 Structure of this Thesis

The rest of the thesis is structured as follows. In Chapter 2 we review the necessary probability theory foundations. We define the basics of measure theory and introduce random variables. Then we go on to define stochastic processes and random fields. We explain concentration and chaining, which are tools that help us uniformly bound stochastic processes. In Chapter 3 we explain Gaussian process regression and Bayesian optimization in detail. For the former, we show that it is the best linear unbiased predictor. For the latter, we give a brief overview and talk about different quantities one can optimize for. In Chapter 4 we then derive a bound of the metric entropy of a stochastic process obtained from Bayesian optimization. This bound is then used to derive an upper bound on the chance of further improvement by two different approaches. Firstly using Dudley's inequality combined with a concentration inequality. And secondly using Talagrand's inequality. This bound is then implemented in an algorithm. We end the thesis with a quick conclusion of the results and an outlook in Chapter 5. If any notation is not self evident from the context, we refer to Appendix A.1.

Chapter 2

Probability Theory Foundations

This part of the thesis mainly serves to introduce the necessary fundamentals to describe the later algorithms and to establish consistent notation. The results of Sections 2.3 and 2.5 especially will become important in Chapter 4. For a more in depth treatment of the subjects of this chapter, we refer the reader to literature such as [6, 7].

2.1 Basic Probability

Probability theory often concerns itself with the probability of drawing from a certain subset instead of drawing one specific result. Given the task of guessing a random number between 0 and 1, the probability of any single guess being correct is zero (without prior knowledge about how was randomly picked). But the probability of drawing from a certain subset of $[0, 1]$, for example $[0, 0.3]$ has a non-zero probability. This gives rise to a more meaningful way to describe the random behavior. For that reason, we will start this chapter not with probability theory, but with a treatment of measuring subsets. Specifically we start with measure theory.

As it turns out, finding a measure for all subsets of $\mathbb{R}$ that coincides with the measure induced by the Euclidean distance on all open intervals is not possible, assuming the axiom of choice, due to the Vitali set [36]. As a next best thing, we will not work with the set of all subsets, but with σ-algebras instead.

Definition 2.1 (σ-Algebra)**.** Let Ω be a set. The collection $\mathcal{A} \subseteq \mathcal{P}(\Omega)$ is called a *σ-algebra* if

(i) Ω is in $\mathcal{A}$.

(ii) $\mathcal{A}$ is closed under complementation, for $A \in \mathcal{A}$ we have $A^C \in \mathcal{A}$.

(iii) $\mathcal{A}$ is closed under countable unions, for $A_1, A_2, \dots \in \mathcal{A}$ we have $\bigcup_{k\in\mathbb{N}} A_k \in \mathcal{A}$

We refer to the elements of the σ-algebra as events. The tuple $(\Omega, \mathcal{A})$ is called a *measurable space*.

Some trivial examples of σ-algebras for any set Ω include the collection $\{\emptyset, \Omega\}$ as well as the power set $\mathcal{P}(\Omega)$. But the useful cases often live in between the two extremes. One commonly used construction is that of the smallest σ-algebra that contains a certain set of events $\mathcal{M}$. To obtain it we can take the intersection of all σ-algebras containing X and are guaranteed a unique σ-algebra [7, p.52]. This σ-algebra is called the *σ-algebra generated by $\mathcal{M}$*, we write $\sigma(\mathcal{M})$. The σ-algebra generated by the open sets of the topology of a space has a special name, it is called the Borel σ-algebra $\mathcal{B}$. In $\mathbb{R}^n$ the Borel σ-algebra $\mathcal{B}(\mathbb{R}^n)$ is also generated by the set of open balls with radius $\varepsilon > 0$.

Definition 2.2 (Measure). Let $(\Omega, \mathcal{A})$ be a measurable space. A function $\mu\colon \mathcal{A} \to [0, \infty]$ is called a *measure* on $(\Omega, \mathcal{A})$ if $\mu(\emptyset) = 0$ and if for any sequence $\{A_n\}_{n\geq 0}$ of pairwise disjoint sets in $\mathcal{A}$, the following property (σ-additivity) is satisfied

$$\mu(\sum_{n=0}^{\infty} A_n) = \sum_{n=0}^{\infty} \mu(A_n).$$

The triple $(\Omega, \mathcal{A}, \mu)$ is called a *measure space*. If $\mu(\Omega) = 1$, we call it a *probability measure* and $(\Omega, \mathcal{A}, \mu)$ a *probability space*. As an example of a measure let us look at the Lebesgue measure λ^n on $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$. We will not explicitly define the measure for all sets. Instead, we can define the measure on a generator of the Borel σ-algebra, and are guaranteed uniqueness [7, p.61] by an application of Carathéodory's Theorem. Thus we define

$$\lambda^n \left(\prod_{i=1}^{n} (a_i, b_i] \right) := \prod_{i=1}^{n} (b_i - a_i)$$

for all $a_i, b_i \in \mathbb{R}$. Another canonical example is the *Dirac measure.* For $a \in \Omega$, it is defined as $\delta_a(C) = 1_C(a)$.

Now that we have established the notions of σ-algebras, measurable spaces and measures, we can talk about measureable functions. We call a function $f\colon \Omega \to E$ between two measurable spaces $(\Omega, \mathcal{A})$ and $(E, \mathcal{E})$ a *measureable function* with respect to $\mathcal{A}$ and $\mathcal{E}$ if

$$f^{-1}(C) \in \mathcal{A} \text{ for all } C \in \mathcal{E}.$$

Note that the definition of measureability does not require a measure, but only a measurable space. A measurable function $f\colon (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ satisfying

$$\int_{\Omega} |f| \, \mathrm{d}\mu < \infty$$

is called a *μ-integrable function.*

Definition 2.3 ($\mathcal{L}^p$ spaces, L^p spaces)**.** Let $(\Omega, \mathcal{A}, \mu)$ be a measure space, E a Banach space with norm $\|\cdot\|$. For $1 \leq p \leq \infty$ we define the vector space

$$\mathcal{L}^p(\Omega, \mathcal{A}, \mu; E) \coloneqq \{f \colon \Omega \to E \mid \|f\|_p < \infty\} .$$

with the seminorm

$$\|f\|_p \coloneqq \left(\int_\Omega \|f\|^p \,\mathrm{d}\mu\right)^{1/p}$$

for $p < \infty$ and

$$\|f\|_\infty = \inf\{C \geq 0 : |f(x)| \leq C \text{ for almost every x}\}.$$

Define the equivalence relation

$$f \sim g \colon \iff \mu(\{f(x) \neq g(x)\}) = 0.$$

Then factorizing by $\sim$, we arrive at the Banach space

$$L^p \coloneqq \mathcal{L}^p / \sim$$

with norm

$$\|[f]\|_p \coloneqq \|f\|_p.$$

Whenever obvious from the context, we will not mention the underlying sets, instead writing L^p. We will also write f instead of $[f]$ for an element $[f] \in L^p$.

Now let us shift the notation from the analytic and measure theoretic view to the probabilistic view. Instead of dealing with functions f, we now have random variables X. Instead of integrating f, we take the expected value $E[X]$.

Definition 2.4 (Random variables)**.** Let $(\Omega, \mathcal{A}, \mathrm{P})$ be a probability space and let $(E, \mathcal{E})$ be a measurable space. A measurable function

$$X \colon (\Omega, \mathcal{A}) \to (E, \mathcal{E})$$

is called a *random element* with values in E. If $E = \mathbb{R}$ and $\mathcal{A} = \mathcal{B}(\mathbb{R})$, it is called a *random variable.* If $E = \mathbb{R}^n$ and $\mathcal{A} = \mathcal{B}(\mathbb{R}^n)$ for some $n \in \mathbb{N}$, it is called a *random vector*.

Dealing with the one dimensional case first, the expected value of a random variable X is defined as

$$\mathrm{E}[X] \coloneqq \int_\Omega X \,\mathrm{dP}$$

and its variance as

$$\mathrm{Var}[X] \coloneqq \mathrm{E}\left[(X - \mathrm{E}[X])^2\right] .$$

The covariance between two random variables X and Y is given by

$$\mathrm{Cov}(X,Y) = \mathrm{E}[(X - \mathrm{E}[X])(Y - \mathrm{E}[Y])].$$

For a random vector $X = (X_1, \dots, X_n)^T$, such that $X_1, \dots X_n$ are square integrable random variables, we define the *mean*

$$\mu_X := \mathrm{E}[X] = (\mathrm{E}[X_1], \dots, \mathrm{E}[X_n])^T$$

and the *covariance matrix* of X

$$\Sigma_X := \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^T\right] = \{\sigma_{X_i,X_j}\}_{1 \le i,j \le n}.$$

Take any $y \in \mathbb{R}^n$, The calculation

$$\begin{aligned} y^T \Sigma_X y &= y^T \mathrm{E}\left[(X - \mathrm{E}[X])(X - \mathrm{E}[X])^T\right] y \\ &= \mathrm{E}\left[y^T (X - \mathrm{E}[X])(X - \mathrm{E}[X])^T y\right] \\ &= \mathrm{E}\left[\left\|(X - \mathrm{E}[X])^T y\right\|^2\right] \ge 0. \end{aligned}$$

shows that the covariance matrix is always positive semidefinite. Now we have some tools to describe random variables globally. But we do not have the structure to describe them locally yet. We say a random element X with values in a measurable space $(E, \mathcal{E})$ has a *distribution* P_X on $(E, \mathcal{E})$, which is given by the image of the probability measure P mapping X from $(\Omega, \mathcal{A})$ to $(E, \mathcal{E})$, that is, for all $C \in \mathcal{E}$ we have

$$\mathrm{P}_X(C) = \mathrm{P}(X \in C).$$

Note that P_X is a measure on $(E, \mathcal{E})$. The *cumulative distribution function* (CDF) F_X of a random variable is given by

$$F_X(x) = \mathrm{P}_X((-\infty, x]) = \mathrm{P}(X \le x)$$

for any x. For two random variables X and Y we define the *joint cumulative distribution function*

$$F_{X,Y}(x,y) = \mathrm{P}(X \le x, Y \le y).$$

Two random variables X and Y are called *independent* of each other if the joint cumulative distribition function can be decomposed into

$$F_{X,Y}(x,y) = F_X(x) F_Y(y)$$

for all x, y. If, for a random variable X, there exists a real valued function f_X, such that

$$\mathrm{P}(X \in A) = \int_A f_X \,\mathrm{dP}$$

for all $A \in \mathcal{A}$, we say f_X is the *probability density function*(PDF) of X. If a random variable X has a density f_X, then its expected value is also given by

$$\mathrm{E}[X] = \int_{-\infty}^{\infty} x f_X(x)\,\mathrm{d}x,$$

the expected value of a random variable mapped by a function g is given by

$$\mathrm{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x)\,\mathrm{d}x.$$

It is often useful to give an alternative characterization of a distribution. Two common approaches are the characteristic function and the moment generating function. Both give a unique and complete alternative characterization of a distribution. The *characteristic function* φ_X of a random variable X is defined by the Fourier transformation

$$\varphi_X(u) = \mathrm{E}\left[e^{iuX}\right].$$

Similarly, the characteristic function of a real random vector X is defined by

$$\varphi_X(u) = \mathrm{E}\left[e^{iu^T X}\right].$$

The *moment generating function* (MGF) of a random variable X is

$$M_X(u) = \mathrm{E}\left[e^{uX}\right],$$

provided that this expectation exists in a neighbourhood of $u = 0$. Similarly we define

$$M_X(u) = \mathrm{E}\left[e^{u^T X}\right]$$

for a random vector X. Characteristic functions have some useful properties. For independent variables X and Y, the characteristic function of $X + Y$ has the property

$$\varphi(X + Y) = \varphi(X)\varphi(Y).$$

Later, when we take measurements of a stochastic process, the measurments will be incorporated into the model by conditioning the prior probability on the measurements. Before we do that in the context of stochastic processes, let us look at how random variables can be conditioned. Similar to the conditional probability $\mathrm{P}(A \mid B) = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(B)}$ for events A and B, there also is a *conditional probability distribution*

$$f_{X|Y}(x, y) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} \text{ if } f_Y(y) > 0 \\ 0 \text{ otherwise} \end{cases}$$

for random variables X, Y with densities f_X, f_Y. If these are jointly distributed random variables, f_Y is often computed as a *marginal density* $f_Y(y) = \int f_{X,Y}(x,y)\,\mathrm{d}x$. We further define the *conditional (cumulative) distribution* of X given $Y = y$ by

$$\mathrm{P}(X \leq x \mid Y = y) = \int_{-\infty}^{x} f_{X|Y}(\vartheta, y)\,\mathrm{d}\vartheta,$$

as well as the *conditional expectation*

$$\mathrm{E}[X \mid Y = y] = \int_{-\infty}^{\infty} x f_{X|Y}(x,y)\,\mathrm{d}x.$$

More generally, consider a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ with a sub-algebra $\mathcal{H} \subset \mathcal{A}$ and a random variable X. A $\mathcal{H}$-measurable function $\mathrm{E}[X \mid \mathcal{H}]$ satisfying

$$\int_H \mathrm{E}[X \mid \mathcal{H}]\,\mathrm{dP} = \int_H X\,\mathrm{dP}$$

for all $H \in \mathcal{H}$ is called the *conditional expectation* of X given $\mathcal{H}$. The existence and uniqueness of this object is not trivial, we refer to [7] for a treatment of that matter. One important example of a sub-algebra often used in this context, is that of a σ-algebra generated by a random element $Y\colon (\Omega, \mathcal{A}) \to (E, \mathcal{E})$, denoted by $\sigma(Y)$. It is the smallest σ-algebra containing all the pre-images of open sets, so $\sigma(Y) = \{Y^{-1}(A) \mid A \in \mathcal{E}\}$. When working with random variables, one can pick any generating set of the Borel σ-algebra $\mathcal{B}$, like the right bounded intervals $\{(-\infty, a] \mid a \in \mathbb{R}\}$ which yields $\sigma(Y) = \{\{Y \leq a\} \mid a \in \mathbb{R}\}$. If we are conditioning with a σ-algebra generated by a random variable Y, we write

$$\mathrm{E}[X \mid Y] = \mathrm{E}[X \mid \sigma(Y)].$$

The variance also lends itself to conditioning

$$\mathrm{Var}[X \mid Y] = \mathrm{E}\left[(X - \mathrm{E}[X \mid Y])^2 \mid Y\right].$$

Example 2.5 (Uniform distribution)**.** Let $a, b \in \mathbb{R}$. A real random variable X with probability density function

$$f(x) = \frac{1}{b-a} 1_{[a,b]}$$

is called a *uniform* random variable on $[a,b]$. This is denoted by $X \sim \mathcal{U}([a,b])$. The CDF of X is given by

$$F_X(x) = \begin{cases} 0 \text{ for } x < a, \\ \frac{x-a}{x-b} \text{ for } a \leq x \leq b, \\ 1 \text{ for } x > b. \end{cases}$$

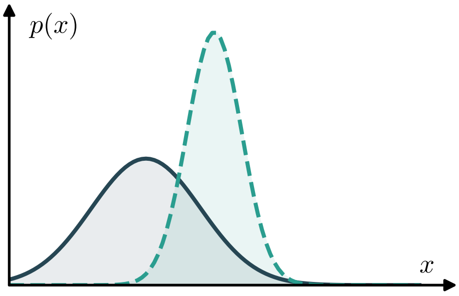

Figure 2.1 | Two normal distributions with different σ and μ.

Example 2.6 (Normal distribution). A random variable X with PDF

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}},$$

where $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}_+$, is called a *Gaussian* random variable. We also say X is normally distributed. This is denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$. Two examples of normal distributions are shown in Figure 2.1. A random vector X with PDF

$$f(x) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma)}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right),$$

where x is a vector of input variables and Σ is the covariance matrix, is called a *multivariate Gaussian* random vector. In this case we also say that its components x_i are jointly Gaussian.

Gaussian random variables have some properties, that make them easier to work with. To mention one, uncorrelated Gaussian random variables are automatically independent [7, Thm. 3.2.7]. Now that we have seen two first examples of probability distributions, let us take a closer look at how to distributions. Under some weak assumptions, the laws of large numbers assert that the sums of independent random variables are likely to be near their expected values. These sums serve as fundamental examples of random variables that are concentrated around their mean. As we will see on a few examples, this behavior is common among many functions of independent random variables. This leads us to concentration inequalities. They usually take the form of bounds for the tails $X - \mu$, such as

$$\mathrm{P}(|X-\mu| \geq x) \leq \text{something small.}$$

We can see different tail behaviors of two distributions shown in Figure 2.2. A first example of a concentration inequality is *Markov's inequality*. Note that for any nonnegative random variable X and any $t > 0$,

$$X \geq t\mathbf{1}_{\{X \geq t\}},$$

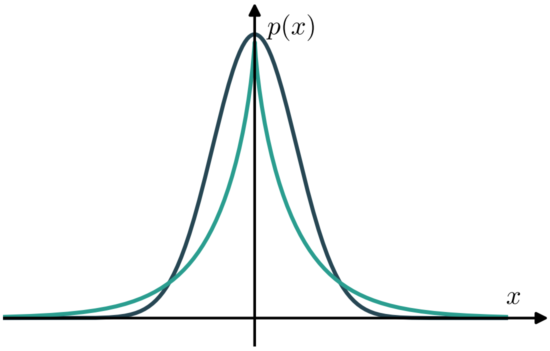

Figure 2.2 | Two distributions with different tail behaviors.

where $\mathbf{1}_{\{X \geq t\}}$ is the indicator function of the event $\{X \geq t\}$. Taking the expectation of both sides yields Markov's inequality

$$\mathrm{E}[X] \geq t\mathrm{P}(X \geq t).$$

It follows from Markov's inequality that if φ is a strictly monotonically increasing nonnegative function, then for any random variable X and any real number t

$$\mathrm{P}(X \geq t) = \mathrm{P}(\varphi(X) \geq \varphi(t)) \leq \frac{\mathrm{E}[\varphi(X)]}{\varphi(t)}.$$

The application of $\varphi(x) = x^2$ leads to *Chebyshev's inequality*. For an arbitrary random variable X and $t > 0$

$$\mathrm{P}(|X - \mathrm{E}[X]| \geq t) = \mathrm{P}(|X - \mathrm{E}[X]|^2 \geq t^2) \leq \frac{E[|X - \mathrm{E}[X]|^2]}{t^2} = \frac{\mathrm{Var}(X)}{t^2}.$$

Another application of this technique are *Chernoff bounds*. For that, we take $\varphi(x) = e^{\lambda x}$ for some $\lambda > 0$ and obtain

$$\mathrm{P}(X \geq t) = \mathrm{P}(e^{\lambda X} \geq e^{\lambda t}) \leq \frac{\mathrm{E}[e^{\lambda X}]}{e^{\lambda t}}.$$

Going back to Gaussian random variables once more, we can categorize random variables by their tail behavior in relation to the Gaussian distribution. We call a random variable X *sub-Gaussian*, if there exists a positive constant C, such that for every $x \geq 0$,

$$\mathrm{P}(|X| \geq x) \leq 2\exp(-x^2/C^2).$$

Equivalently, we can define sub-Gaussian random variables X using the bound on the moment generating function $\mathrm{E}[e^{\lambda(X-\mu)}] \leq e^{\sigma^2 \frac{\lambda^2}{2}}$.

2.2 Stochastic Processes

Definition 2.7 (Stochastic process)**.** Take a probability space $(\Omega, \mathcal{A}, \mathrm{P})$ and a set of random variables $\{X_t\}_{t \in T}$, which all take values in the same measurable sample space $(E, \mathcal{E})$, indexed by some set T. We call $\{X_t\}_{t \in T}$ a *stochastic process*.

In this thesis, we will always assume that T is a metric space with a group structure. From the point of view of mappings, we have for every $t \in T$ a measurable map

$$X_t \colon \Omega \to E,$$

whose inverse maps $\mathcal{E}$ into $\mathcal{A}$. Fixing a point $\omega \in \Omega$ in the probability space instead, we can take a different perspective and define sample paths

$$\begin{aligned} X(\omega) \colon T &\to E \\ t &\mapsto X_t(\omega). \end{aligned}$$

But ultimately, we want to treat the stochastic process as a random variable with values in the space of functions $E^T := \{f \mid f \colon T \to E\}$. That is,

$$\begin{aligned} X \colon \Omega &\to E^T \\ w &\mapsto X(\omega). \end{aligned}$$

Figure 2.3 shows an example of such random functions next to random variables and vectors. The question arises if X, viewed in this way, is measurable as a map from $(\Omega, \mathcal{A})$ to $(E^T, \mathcal{E}^T)$. As our σ-algebra $\mathcal{E}^T$ we pick the σ-algebra generated by all the sets of the form $\{x \in E^T \mid x_t \in C\}$ with $t \in T$ and $C \in \mathcal{E}$. This is the smallest σ-algebra that makes X and the projections $\pi_t \colon E^T \to E$ with $\pi_t(x) = x_t$ measurable [6, Lemma 3.2]. By the measurability of the X_t, we have

$$X^{-1}(\{x \mid x_t \in C\}) = \{\omega \mid X_t(\omega) \in C\} \in \mathcal{A}$$

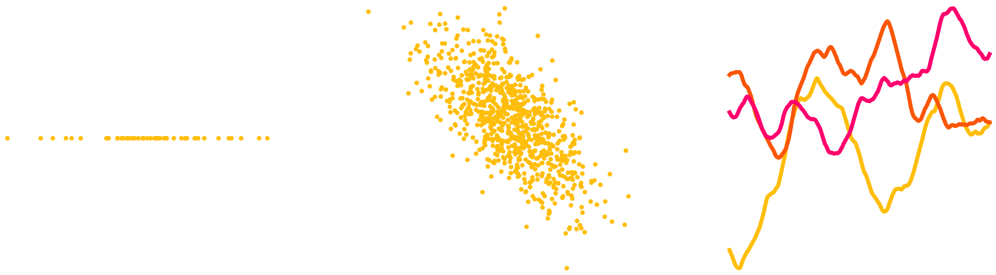

Figure 2.3 | Samples drawn from a random variable, a random vector, and a random process.

for all $t \in T$ and $C \in \mathcal{E}$. Therefore $X^{-1}(S) \in \mathcal{A}$ for all S from our generating set and X is measurable in the desired sense.

The collection of probability distributions of the random vectors

$$(X_{t_1}, \dots, X_{t_k})$$

for all $k \geq 1, t_1, \dots, t_k \in T$ is called the *finite dimensional distribution* of $\{X_t\}_{t \in T}$. We define the projections onto a finite subset $H \subset T$ as the projections onto the vector with components in H

$$\begin{aligned} \pi_H \colon E^T &\to E^H \\ x &\mapsto (x_h)_{h \in H}. \end{aligned}$$

Suppose that for each finite $H \subset T$ there is a probability measure μ_H on E^H. We say that the collection of measures μ_H is consistent if

$$H \subset T \text{ finite } \Leftrightarrow \mu_H = \mu_T \circ \pi_H^{-1}.$$

Not every measurable stochastic process is consistent. This is where the Kolmogorov extension theorem comes in.

Theorem 2.8 (Kolmogorov extension theorem)**.** Given a consistent collection $\{\mu_H \mid H \subset T \text{ finite}\}$ of probability measures on $\mathbb{R}^H$ respectively, there exists a unique probability measure P on the Borel subsets of $\mathbb{R}^T$ such that $\mathrm{P} \circ \pi_H^{-1} = \mu_H$, for all finite $H \subset T$.

We will omit the proof of this theorem and refer to standard literature such as [7, 34]. vTwo stochastic processes $\{X(t)\}_{t \in T}$ and $\{Y(t)\}_{t \in T'}$, defined on the same probability space, with values in $(E, \mathcal{E})$ and $(E', \mathcal{E}')$ respectively, are called *independent* if their σ-algebras $\sigma(X(t); t \in T)$ and $\sigma(Y(t); t \in T')$ are independent, so if all pairs of events, one from each of the two σ-algebras, are independent. The probability P_X on $(E^T, \mathcal{E}^T)$, that is the image of P by $X \colon \Omega \to E^T$, is called the *distribution* of $\{X(t)\}_{t \in T}$.

Proposition 2.9 ([7])**.** For the stochatic processes $\{X(t)\}_{t \in T}$ and $\{Y(t)\}_{t \in T}$ to be independent, it suffices that for all $t_1, \dots, t_k \in T$ and $s_1, \dots, s_l \in T'$, the vectors $(X_{t_1}, \dots, X_{t_k})$ and $(Y_{s_1}, \dots, Y_{s_l})$ to be independent.

Two processes X and Y are said to be a *version* of each other if they have the same finite dimensional distributions. They are said to be a *modification* of each other if they are equal almost surely, that is, if $\mathrm{P}(X_t = Y_t) = 1$. A measurable stochastic process $\{X(t)\}_{t \in T}$ satisfying the condition

$$\mathrm{E}\left[|X(t)|^2\right] < \infty$$

is called a *second-order* stochastic process For such a process we can define the *mean*

$$m(t) \coloneqq \mathrm{E}[X(t)]$$

and *covariance*

$$\Gamma(t,s) \coloneqq \mathrm{Cov}(X_t, X_s) = \mathrm{E}[X(t)X(s)] - m(t)m(s).$$

Definition 2.10 (Gaussian Process)**.** Let T be an arbitrary index set. A real-valued stochastic process $\{X_t\}_{t\in T}$ is called a *Gaussian process* if for all finite $t_1, \dots, t_n \in T$, the random vector $(X_{t_1}, \dots, X_{t_n})$ is Gaussian.

Theorem 2.11 (Gaussian Process)**.** Let T be a set and let $C\colon T \times T \to \mathbb{R}$ define a positive-definite function, that is,

$$\sum_{i=1}^{n}\sum_{j=1}^{n} u_i u_j C(t_i, t_j) > 0$$

for all $n \in \mathbb{N}$, $t_1, \dots, t_n \in T$ and $u_1, \dots, u_n \in \mathbb{R}$. Then there exists a unique zero mean Gaussian process with covariance C.

Proof. Take an arbitrary $n \in \mathbb{N}$ and take $t_1, \dots, t_n \in T$. Let us construct a Gaussian vector X with zero mean and covariance matrix $\Gamma = \{C(t_i, t_j)\}_{1\le i,j\le n}$. Since Γ defines a positive-definite matrix, it has a Cholesky decomposition $\Gamma = AA^T$. Let $X = AZ$, where Z is a vector of independent standard normal random variables. Then X is a Gaussian vector with zero mean and covariance matrix Γ, as we can see from computing the characteristic function

$$\varphi_X(u) = \mathrm{E}[e^{i\langle u, AZ\rangle}] = e^{-\frac{1}{2}\|u^T A\|^2} = e^{-\frac{1}{2}\langle u, \Gamma u\rangle}.$$

To construct a Gaussian process with covariance C, we will use these vectors and apply Theorem 2.8. For that we need to show that the set of finite dimensional distributions constructed in this way are consistent. As pointed out in [9], the projection of the Gaussian distribution on $\mathbb{R}^n$ with covariance matrix $\Gamma = [C(t_i, t_j)]_{1\le i,j\le n} \in \mathbb{R}^{n\times n}$ to the first first $n-1$ coordinates is a Gaussian distribution with covariance matrix $\Gamma = [C(t_i, t_j)]_{1\le i,j\le n-1} \in \mathbb{R}^{(n-1)\times(n-1)}$. Therefore the set of finite dimensional distributions for all $n \in \mathbb{N}$ and $t_1, \dots, t_n \in T$ is consistent. And by Theorem 2.8 a zero mean Gaussian process with covariance C exists and is unique. □

We have seen how the Kolmogorov extension theorem can be used to construct a Gaussian process. Of course it is also possible to construct stochastic processes with other distributions. Markov chains would be another example. But Gaussian processes will suffice for our purposes, we will go into more detail about why Gaussian processes are so useful at the beginning of Chapter 3. Figure 2.4 shows

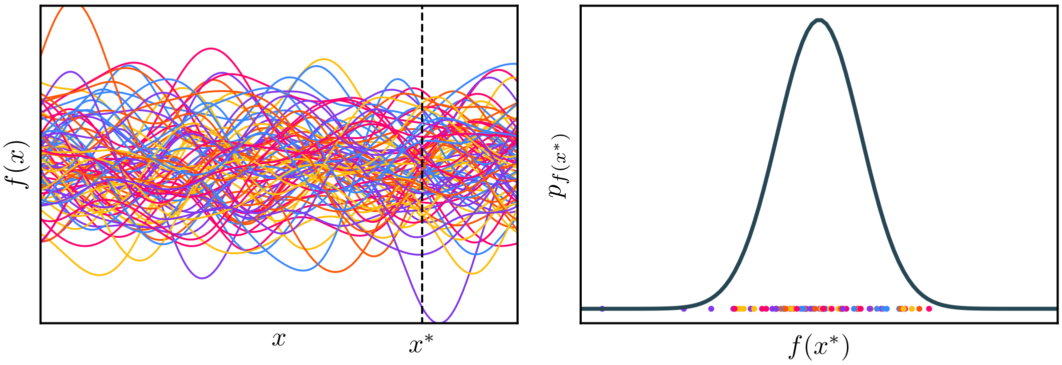

Figure 2.4 | Sample paths of a Gaussian process evaluated at $x^* \in T$.

samples drawn from a Gaussian process and a cross-section for one $x^* \in T$. This cross section illustrates, that for any given point, the paths are normally distributed.

Lemma 2.12 (Conditioned Gaussian Random Vector)**.** Let $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ be a Gaussian random vector with mean

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$$

Then the conditioned random vector $X_1 \mid X_2$ is also Gaussian with mean

$$\bar{\mu} = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(a - \mu_2)$$

and covariance

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Proof. This is a well known result. We will follow the proof outlined in [4]. Let n_1 and n_2 be the dimensions of X_1 and X_2 respectively. Further, let

$$\Sigma^{-1} = \Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}$$

be the inverse of the covariance matrix. We note that Λ is symmetric, since the covariance matrix is also symmetric. One possible path to the solution is to compute the conditional distribution $p(X_1 \mid X_2)$ explicitly. This is done in [31]. We will take a different approach. Note that due to the chain rule of probability, the conditional distribution is given by

$$p(X_1 \mid X_2) = \frac{p(X_1, X_2)}{p(X_2)} = \frac{p(X)}{p(X_2)}.$$

If we fix X_2, this distribution is a scaled version of the joint distribution $p(X)$. We will use this fact, and the fact that Gaussian distributions are uniquely determined by the quatratic form in the exponent of the density function

$$-\frac{1}{2}(X-\mu)^T\Sigma^{-1}(X-\mu) \tag{2.1}$$

to derive the conditional distribution. For that purpose, we want to show that the exponent of the conditional distribution is a quadratic form in X_1. Spreading the multiplication into the block matrix components (2.1) yields

$$\begin{aligned}-\frac{1}{2}&(X-\mu)^T\Sigma^{-1}(X-\mu) = \\ &-\frac{1}{2}(X_1-\mu_1)^T\Lambda_{11}(X_1-\mu_1) - \frac{1}{2}(X_1-\mu_1)^T\Lambda_{12}(X_2-\mu_2) \\ &-\frac{1}{2}(X_2-\mu_2)^T\Lambda_{21}(X_1-\mu_1) - \frac{1}{2}(X_2-\mu_2)^T\Lambda_{22}(X_2-\mu_2)\end{aligned} \tag{2.2}$$

To show that this is a quadratic form in X_1, we will use a technique called *completing the square*. A general Gaussian distribution with mean μ and covariance Σ has an exponent of the form

$$-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) = -\frac{1}{2}x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu - \frac{1}{2}\mu^T\Sigma^{-1}\mu. \tag{2.3}$$

Here, we have a part that is quadratic in x, a part that is linear in x and a part that is constant in x. To bring (2.2) into this form, we can thus collect the second order terms in X_1 from (2.2) and compare them to the second order terms of x in (2.3) to derive the covariance matrix. And collect the first order terms in x_1 to calculate the mean.

Collecting all the second order terms in X_1 from (2.2), we get

$$-\frac{1}{2}X_1^T\Lambda_{11}X_1.$$

Thus we can immediately conclude $\bar{\Sigma} = \Lambda_{11}^{-1}$. Similarly we can collect all the first order terms in X_1 from (2.2) and get

$$X_1^T(\Lambda_{11}\mu_1 - \Lambda_{12}(X_2-\mu_2)),$$

where we have used the symmetry of Λ. Due to (2.3), the coefficient of this expression must be equal to $\bar{\Sigma}^{-1}\bar{\mu}$. So, multiplying by $\bar{\Sigma}$, we get

$$\begin{aligned}\bar{\mu} &= \bar{\Sigma}(\Lambda_{11}\mu_1 - \Lambda_{12}(X_2 - \mu_2)) \\ &= \mu_1 - \Lambda_{11}^{-1}\Lambda_{12}(X_2 - \mu_2)\end{aligned}$$

We can ignore the constant terms in (2.2) and (2.3), since they are only a scaling factor for the density function. We have thus shown, that the exponent of the conditional distribution is a quadratic form in X_1. Therefore the conditional distribution is Gaussian.

The only thing missing now is to calculate the inverses of our block matrices in Λ, in order to get an explicit formula for $\bar{\mu}$ and $\bar{\Sigma}$. A Matrix inversion Lemma [24] for 2 × 2 block matrices states that

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} (A - BD^{-1}C)^{-1} & -(A - BD^{-1}C)^{-1}BD^{-1} \\ -D^{-1}C(A - BD^{-1}C)^{-1} & D^{-1} + D^{-1}C(A - BD^{-1}C)^{-1}BD^{-1} \end{bmatrix},$$

given, that all the inverses exist. Applying this to our case, we get

$$\begin{aligned}\Lambda_{11}^{-1} &= \Sigma_{11}^{-1} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \\ \Lambda_{12} &= -(\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\Sigma_{12}\Sigma_{22}^{-1}.\end{aligned}$$

Plugging these into the expressions for $\bar{\mu}$ and $\bar{\Sigma}$ gives the formulas

$$\begin{aligned}\bar{\Sigma} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}, \\ \bar{\mu} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2).\end{aligned}$$ □

Proposition 2.13 (Conditional Gaussian Processes)**.** Take a Gaussian process $\{X(t)\}_{t\in T}$ with mean function $m(t)$ and covariance function $\Gamma(t,s)$. Conditioning on the event $X(t_1) = x_1, \dots, X(t_n) = x_n$ gives another Gaussian process.

Proof. This is just an application of Lemma 2.12 and Theorem 2.8. □

2.3 Random Fields

A lot of the study of stochastic processes is concerned with random variables indexed by time. *Random fields* are stochastic processes on a more general topologial index set T. For our purposes, we will just look at n-dimensional real vectors. An example of a random field on $[0,1]^2$ is given in Figure 2.5. But other index sets, like sets of functions are possible.

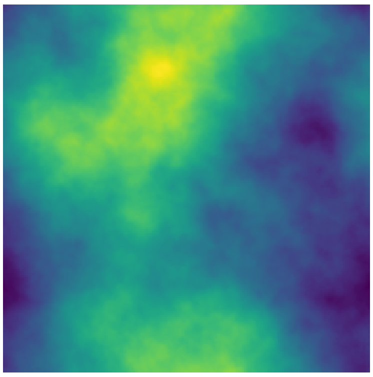

Figure 2.5 | A random field.

To capture the spacially correlated structure of random fields, we need to define a *covariance function*, often also referred to as a covariance kernel. Such a function gives the covariance of the values of a random field X at points x, y

$$\begin{aligned} C(x,y) &\coloneqq \operatorname{Cov}(X_x, X_y) \\ &= \operatorname{E}\left[(X_x - \operatorname{E}[X_x])(X_y - \operatorname{E}[X_y])\right]. \end{aligned}$$

Arbitrary functions are not admissable, a function is a valid covariance function if and only if it is positive semi-definite, i.e.

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_i C(x_i, x_j) w_j \geq 0$$

for all $x, y \in T$, $n \in \mathbb{N}$ and weights $w \in R^n$ It is clear to see, that this is a necessary condition, since $\operatorname{Var}(\sum_{i=1}^n w_i X_{x_i}) = \sum_{i=1}^n \sum_{j=1}^n w_i C(x_i, x_j) w_j$ needs to be non-negative. The fact that every positive semidefinite function defines a valid covariance function is less obvious and is a conclusion the following Theorem 2.14.

Theorem 2.14 (Bochner Theorem). [5, Thm. 19] A continous stationary function $\Gamma(s,t) = C(|s-t|)$ is positive definite (i.e. a covariance function) if and only if it can be represented as

$$C(t) = \int e^{2\pi i \omega t}\, \mathrm{d}\mu(\omega),$$

where μ is a finite positive measure. If μ has a density $S(s)$, then S is called the *spectral density* or *power spectrum* of C, and C and S are Fourier duals, that is

$$C(t) = \int S(s) e^{2\pi i \langle t,s\rangle}\, \mathrm{d}s$$

and

$$S(s) = \int C(t) e^{-2\pi i \langle t,s\rangle}\, \mathrm{d}t.$$

This form of a covariance function can then be used to construct a random field. For the details we refer to [11, p.84]. The following examples of covariance functions are not given in terms of two variables from the index set T, but in terms of the distance between two points $s, t \in T$. A covariance Γ for example might be charactercized by a function C with $\Gamma(s,t) = C(|s-t|)$. The edge cases of perfect covariance and zero covariance everywhere but the origin lead to random fields that are the same everywhere and white noise respectively. Given a correlation length scale ρ and variance σ, the *cosine covariance function* is defined as

$$C_{\cos}(d) = \sigma^2 \cos\left(2\pi \frac{d}{\rho}\right).$$

This function was used to generate the random field on the right in Figure 1.2. The *squared exponential covariance function* is defined as

$$C_{SE}(d) = \sigma^2 \exp\left(-\frac{d^2}{\rho^2}\right).$$

This covariance function is infinitely differentiable, therefore generates very smooth random fields [2]. A less smooth, and hence often more useful covariance function is the *Matérn covariance function*, defined as

$$C_\nu(d) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{d}{\rho}\right)^\nu K_\nu(\sqrt{2\nu}\frac{d}{\rho}),$$

where Γ is the gamma function, K_ν is the modified Bessel function of the second kind and ν is a positive smoothness parameter of the covariance function. Its Fourier transformation is given by

$$\widehat{C_\nu}(z) = \frac{\Gamma(\nu + d/2)}{\pi^{d/2}\Gamma(\nu)} \frac{\alpha^d}{(1+\alpha^2 z^2)^{\nu+d/2}},$$

for $z \in \mathbb{R}$ [1, Eq. 11.4.44].

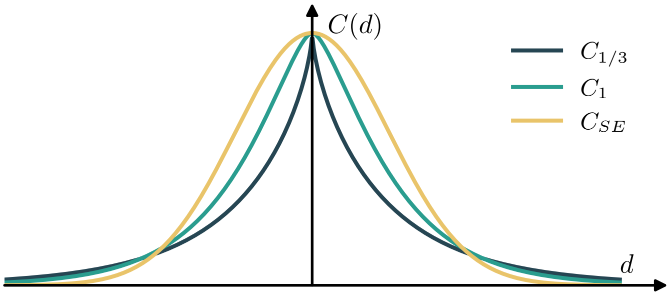

Figure 2.6 | Different Matérn and squared exponential covariance kernels.

Definition 2.15 (Strict Stationarity). A stochastic process $\{X(t)\}_{t\in T}$ is called *strictly stationary* if for all $k \geq 1$, all $(t_1, \dots t_k)$ in T^k, the probability distribution of the random vector

$$(X(t_1 + h), \dots, X(t_k + h))$$

is independent of $h \in T$, with $h \in T$ such that $t_1 + h, \dots, t_k + h \in T$.

Definition 2.16 (Second Order Stochastic Process). A measureable stochastic process $\{X_t\}$ satisfying the condition

$$\mathrm{E}\left[|X_t|^2\right] < \infty$$

for all $t \in T$ is called a *second order* stochastic process

Definition 2.17 (Weak Stationarity)**.** A second order stochastic process $\{X_t\}$ with mean $\mu_X(t)$ and covariance $\Gamma_X(t,s)$ at $t,s \in T$ is called *weakly stationary*, if μ_X is constant and $\Gamma_X(t,s)$ is a function of $t-s$ only.

A process is called *isotropic* if its covariance function $\Gamma(t,s)$ is only a function of the distance $|t-s|$. An example of when this condition is not met is shown in Figure 2.7

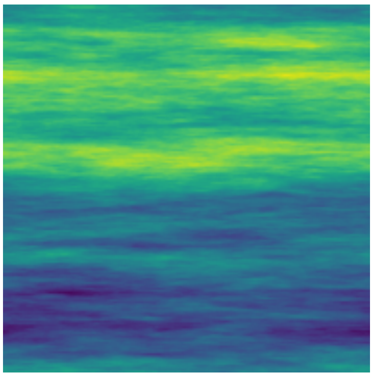

Figure 2.7 | An anisotropic random field.

Theorem 2.18 (Stationarity Equivalence)**.** For Gaussian processes on $T=\mathbb{R}^d$, $d \in \mathbb{N}$ the concepts of weak and strict stationarity coincide.

Proof. Strict stationarity implies weak stationarity trivially. For the other direction note that the mean and covariance functions completely characterize the finite dimensional distributions, since these are normal distributions. □

Proposition 2.19. The Matérn covariance function with length parameter ω and smoothness parameter α is positive definite. Thereby it is a valid covariance function.

Proof. We can use the Fourier transformation of the Matérn covariance function, which is given by $\xi \mapsto \frac{1}{(|\xi|^2+\omega^2)^\alpha}$. For $n \in \mathbb{N}$ and $z_1,\dots,z_n \in \mathbb{C}$ we have

$$\begin{aligned}\sum_{j=1}^{n}\sum_{k=1}^{n} z_j\overline{z_k}C(t_j-t_k) &= \sum_{j=1}^{n}\sum_{k=1}^{n} z_j\overline{z_k}\int_{\mathbb{R}^d} e^{i\xi t_j}e^{-i\xi t_j}\frac{1}{(|\xi|^2+\omega^2)^\alpha}\,\mathrm{d}\xi \\ &= \int_{\mathbb{R}^d}\sum_{j=1}^{n}\sum_{k=1}^{n} z_j e^{i\xi t_j}\overline{z_k e^{i\xi t_j}}\frac{1}{(|\xi|^2+\omega^2)^\alpha}\,\mathrm{d}\xi\end{aligned}$$

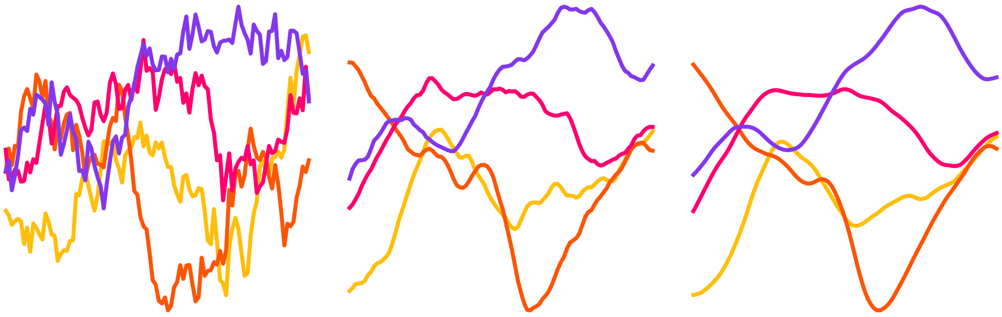

Figure 2.8 | Samples drawn from Matérn processes with different smoothness parameters.

$$= \int_{\mathbb{R}^d} \left| \sum_{j=1}^{n} z_j e^{i\xi t_j} \right|^2 \frac{1}{(|\xi|^2 + \omega^2)^\alpha} \, \mathrm{d}\xi \geq 0.$$

Therefore the Matérn covariance function is positive definite and thereby a valid covariance function. □

The smoothness of the Gaussian process f is determined by the smoothness of the covariance function C. A process f is continuous in the mean square sense at $x^* \in T$, i.e. $\mathrm{E}[|f(x^*) - f(x_k)|^2] \to 0$ as $x_k \to x^*$, if and only if the covariance function $C(x, x')$ is continuous at the point $x = x' = x^*$ [39]. If for a stationary processes f the $2n$-th order partial derivative of the covariance function $\partial^{2n} C(x)/\partial^2_{x_{i1},\ldots,x_{in}}$ exists and is finite at $x = 0$, then the n-th order partial derivative $\partial^n f(x)/\partial_{x_{i1},\ldots,x_{in}}$ exists for all $x \in \mathbb{R}^d$ as a mean square limit [39]. As we can see in Figure 2.6, the Matérn covariance kernels for different parameters are not differentiable at the origin. But we still get Hölder continuity to varying degrees depending on the parameter ν, as will be explored closer in Chapter 4. Figure 2.8 shows samples drawn from Matérn processes with different smoothness parameters. The bigger the parameter ν, the smoother the random field. The limit $\nu \to \infty$ yields the squared exponential covariance function [28]. Another result dealing with the continuity of random fields is the following proposition.

Proposition 2.20 (Multiparameter Kolmogorov Continuity Theorem)**.** Let $X = \{X_t \mid t \in [0,1]^N\}$ be a real-valued stochastic process. Suppose there exist positive constants β, K and δ such that

$$\mathrm{E}[|X_s - X_t|^\beta] \leq K \|s - t\|^{N+\delta}$$

for all $s, t \in [0,1]^N$. Then X has a modification which is uniformly Hölder continuous on $[0,1]^N$ of all orders $< \frac{\delta}{\beta}$

For a proof of this statement, we refer to [41].

2.4 Concentration

This section deals with the concentration of stochastic processes around their mean. We are looking to generalize the concept of concentration of measure for random variables introduced in Section 2.1 to stochastic processes.

Definition 2.21 (Sub-Gaussian Process)**.** A process $\{X_t\}t \in T$ is *sub-Gaussian* with respect to a metric$d$ on T if

$$\mathrm{E}\left[e^{\lambda(X_t X_{\widetilde{t}})}\right] \leq e^{\frac{\lambda^2 d(t,\widetilde{t})^2}{2}}$$

for all $t, \widetilde{t} \in T$ and $\lambda \in \mathbb{R}$.

Similar to sub-Gaussian random variables, this statement is equivalent [37, p.134] to the tail bound

$$\mathrm{P}(|X_\theta - X_{\widetilde{\theta}}| \geq t) \leq 2 \exp\left(-\frac{t^2}{2d_X(\theta, \widetilde{\theta})^2}\right).$$

For the following theorem we need one property.

Definition 2.22 (Separable Process)**.** A stochastic process $\{X_t\}_{t \in T}$ is called *separable* if there exists a countable subset $T_0 \subset T$ and a subset $\Omega_0 \subset \Omega$ with $\mathrm{P}(\Omega_0) = 0$, such that for all $\omega \in \Omega_0$, $t \in T$ and $\varepsilon > 0$,

$$X_t(\omega) \in \overline{\{X_s(\omega) \mid s \in T_0 \cap B_d(t, \varepsilon)\}},$$

where $B_d(t, \varepsilon)$ is the ball of radius ε around t with respect to the metric d.

Proposition 2.23. [16, Prop. 2.1.12] A process $\{X_t\}_{t \in T}$ on (T, d_X), where $d_X(s,t)^2 = \mathrm{E}\left[(X_s - X_t)^2\right]$ is the canonical distance, has a seperable version if and only if the pseudo-metric space (T, d_X) is separable.

Proposition 2.24. [16, Thm. 2.1.20] Let $\{X_t\}_{t \in T}$ be a seperable centered Gaussian process such that

$$\mathrm{P}(\sup_{t \in T} |X(t)| < \infty) > 0.$$

Let Ψ be an even, convex, measurable function, nondecreasing on $[0, \infty)$. Let g be $\mathcal{N}(0,1)$. Then $\sigma(X) := \sup_{t \in T} \mathrm{E}[X_t^2]^{1/2} < \infty$ and $\mathrm{E}[\sup_{t \in T} |X_t|] < \infty$. Further, the following inequalites hold:

$$\mathrm{E}\left[\Psi\left(\sup_{t \in T} |X_t| - \mathrm{E}\left[\sup_{t \in T} |X_t|\right]\right)\right] \leq \mathrm{E}\left[\Psi\left(\frac{\pi}{2}\sigma g\right)\right]$$

and

$$\mathrm{P}\left(\left|\sup_{t \in T} |X_t| - \mathrm{E}\left[\sup_{t \in T} |X_t|\right]\right| > u\right) \leq 2e^{-(Ku^2/2\sigma^2)},$$

where $K = \frac{1}{\pi^2}$.

2.5 Chaining

Another important characterization of a stochastic process is its magnitute

$$\mathrm{E}\left[\sup_{t \in T} X_t\right].$$

We will derive bounds on this quantity by using a technique called chaining. Directly analyzing the stochastic process for uniform bounds can be challenging. The process of chaining involves using information about the geometry of the space (T, d_X) to derive uniform bounds on the stochastic process, where

$$d_X(s,t) = \mathrm{E}\left[(X_s - X_t)^2\right]^{1/2}$$

is the *canonical metric* induced by the stochastic process. We already briefly used this metric in the last section. It is in general not a metric but only a pseudometric, since it is not necessarily positive definite. The canonical metrix can also be reformulated in terms of covariances, for stationary processes in terms of the covariance function C. Take arbitrary $s, t \in T$. Then

$$\begin{aligned} d(s,t)^2 &= \mathrm{Var}(X_s - X_t) \\ &= \mathrm{Var}(X_s) + \mathrm{Var}(X_t) - 2\,\mathrm{Cov}(X_s, X_t) \\ &= 2(C(0) - C(|s-t|)). \end{aligned}$$

The process of chaining is usually done by constructing a sequence of finite sets T_n such that each set T_{n+1} refines the previous set T_n. The elements of each set T_n are chosen such that they cover the index set T as closely as possible with respect to the metric d_X. As n increases, the set T_n becomes a better and better approximation of T. Once we have our sequence of chains T_n, we can then study the behavior of the stochastic process over each chain and ultimately derive bounds of the process over the original index set T.

Definition 2.25 (Covering number)**.** A δ-cover of a set T with respect to a metric d is a set $\{t_1, \ldots, t_N\} \subset T$, such that for each $t \in T$, there exists some $i \in \{1, \ldots, N\}$ such that $d(t, t_i) \leq \delta$. The δ-covering number $\mathcal{N}(T, d, \delta)$ is the cardinality of the smallest δ-cover.

With this first notion of geometric complexity established, we will get to an important result of the chaining technique, Dudley's integral inequality.

Theorem 2.26 (Dudley's integral inequality)**.** Let $(X_t)_{t \in T}$ be a mean zero sub-Gaussian process with respect to the canonical distance d_X. Define $D = \sup_{t,\tilde{t}} d_X(t, \tilde{t})$. Then for any $\delta \in (0, D]$, we have

$$\mathrm{E}\left[\sup_{s,t \in T}(X_s - X_t)\right] \leq 2\,\mathrm{E}\left[\sup_{\substack{\gamma,\gamma' \in T \\ d_X(\gamma,\gamma') \leq \delta}} (X_\gamma - X_{\gamma'})\right] + 32 \int_{\delta/4}^{D} \sqrt{\log \mathcal{N}(T, d_X, u)}\,\mathrm{d}u.$$

Proof. We follow the proof outlined in [37, p. 140]. Let $U = \{t_1, \ldots, t_N\}$ be a minimal δ-covering set of T and for each integer $m = 1, 2, \ldots, L$, let U_m be a minimal $\varepsilon_m = D2^{-m}$ covering set of U in the metric d_X, where we allow any element of T to be used. Here we define L as the smallest integer with $U_L = U$. For the chaining setup, define the projections $\pi_m \colon U \to U_m$ by

$$\pi_m(t) = \operatorname*{argmin}_{\beta \in U_m} d_X(t, \beta)$$

for $m = 1, \dots, L$, so that $\pi_m(t)$ is the best approximation of $t \in U$ from the set U_m. We start from the finest covering $U_L = U$ and recursively construct a sequence with $\gamma^L = t$ and $\gamma^{m-1} = \pi_{m-1}(\gamma^m)$ for $m = L, L-1, \dots, 2$. This lets us construct the telescopic sum

$$X_t - X_{\gamma^1} = \sum_{m=2}^{L} (X_{y^m} - X_{\gamma^{m-1}}),$$

therefore

$$|X_t - X_{\gamma^1}| \le \sum_{m=2}^{L} \max_{\beta \in U_m} |X_\beta - X_{\pi_{m-1}(\beta)}|.$$

Given any other $\widetilde{t}$, we can define the sequence $\{\widetilde{\gamma}^1, \dots, \widetilde{\gamma}^L\}$ and derive an analogous bound for $|X_{\widetilde{t}} - X_{\widetilde{\gamma}^1}|$. Combining the two, we have

$$\begin{aligned} |X_t - X_{\widetilde{t}}| &= |X_{\gamma^1} - X_{\widetilde{\gamma}^1} + (X_t - X_{\gamma^1}) + (X_{\widetilde{\gamma}^1} - X_{\widetilde{t}})| \\ &\le |X_{\gamma^1} - X_{\widetilde{\gamma}^1}| + |X_t - X_{\gamma^1}| + |X_{\widetilde{\gamma}^1} - X_{\widetilde{t}}|. \end{aligned}$$

Taking maxima over all $\widetilde{t}$ and t yields

$$\max_{t, \widetilde{t} \in U} |X_t - X_{\widetilde{t}}| \le \max_{\gamma, \widetilde{\gamma} \in U_1} |X_\gamma - X_{\widetilde{\gamma}}| + 2 \sum_{m=2}^{L} \max_{\beta \in U_m} |X_\beta - X_{\pi_{m-1}(\beta)}|.$$

We first upper bound the finite maximum over U_1 , which has $N(\frac{D}{2}) := \mathcal{N}(T, d_X, \frac{D}{2})$ elements. Because the process is sub-Gaussian, so are the increments $X_t - X_{\widetilde{t}}$ with parameter at most $d_X(t, \widetilde{t}) \le D$. Now we can apply a known bound for the maxima of $N(\frac{D}{2})$ sub-Gaussian random variables [37, p. 53], namely

$$\mathrm{E}\left[\max_{\gamma, \widetilde{\gamma} \in U_1} |X_\gamma - X_{\widetilde{\gamma}}|\right] \le 2D\sqrt{\log N(D/2)}.$$

Similarly, for each $m = 2, 3, \dots, L$, the set U_m has $N(D2^{-m})$ elements, and $\max_{\beta \in U_m} d_x(\beta, \pi_{m-1}(\beta)) \le D2^{-(m-1)}$. Therefore

$$\mathrm{E}\left[\max_{\beta \in U_m} |X_\beta - X_{\pi_{m-1}(\beta)}|\right] \le 2D2^{-(m-1)}\sqrt{\log N(D2^{-m})}.$$

Combining the two pieces, we conclude

$$\mathrm{E}\left[\max_{t, \widetilde{t} \in U} |X_t - X_{\widetilde{t}}|\right] \le 4 \sum_{m=1}^{L} D2^{-(m-1)}\sqrt{\log N(D2^{-m})}.$$

Since the covering number $N(t)$ is non-increasing in t, we have

$$D2^{-(m-1)}\sqrt{\log N(D2^{-m})} \le 4 \int_{D2^{-(m+1)}}^{D2^{-m}} \sqrt{\log N(u)}\, \mathrm{d}u.$$

Here we used the fact, that for non-increasing f, we have

$$\frac{x}{2} = \int_{x/2}^{x} 1 \,\mathrm{d}u \le \int_{x/2}^{x} \frac{f(u)}{f(x)} \,\mathrm{d}u,$$

implying

$$2xf(x) \le 4 \int_{x/2}^{x} f(u) \,\mathrm{d}u.$$

Therefore we can put the all integral boundaries together and get

$$2\,\mathrm{E}\left[\max_{t,\widetilde{t}\in U} |X_t - X_{\widetilde{t}}|\right] \le 32 \int_{\delta/4}^{D} \sqrt{\log \mathcal{N}(T, d_X, u)} \,\mathrm{d}u. \tag{2.4}$$

We are close to the desired result, but we still need to get rid of the restriction to the finite set U. Because U is a δ-cover of T, we can find for any $t \in T$ some $t_i \in U$, such that $d(t, t_i) \le \delta$. Therefore, fixing any $t_1 \in U$,

$$\begin{aligned} X_t - X_{t_1} &= (X_t - X_{t_i}) + (X_{t_i} - X_{t_1}) \\ &\le \sup_{\substack{\gamma,\gamma' \in T \\ d(\gamma,\gamma') \le \delta}} (X_\gamma - X_{\gamma'}) + \max_{i \in \{1,\dots,N\}} |X_{t_i} - X_{t_1}| \\ &\le \sup_{\substack{\gamma,\gamma' \in T \\ d(\gamma,\gamma') \le \delta}} (X_\gamma - X_{\gamma'}) + \max_{t_i,t_j \in U} |X_{t_i} - X_{t_j}|. \end{aligned}$$

The same bound holds given any other $\widetilde{t} \in T$. Adding together yields

$$\sup_{t,\widetilde{t}\in T} (X_t - X_{\widetilde{t}}) \le 2 \sup_{\substack{\gamma,\gamma' \in T \\ d(\gamma,\gamma') \le \delta}} (X_\gamma - X_{\gamma'}) + 2 \max_{t_i,t_j \in U} |X_{t_i} - X_{t_j}|.$$

We can now take the expected value and plug in (2.4). This concludes the proof. □

Remark 2.27. The standard version of Dudley's inequality is stated in terms of the magnitute $\mathrm{E}[\sup_{t\in T} X_t]$ instead of increments. It can be recovered from our version, noting that because of the mean zero assumption we can choose an arbitrary $t_0 \in T$ and write

$$\mathrm{E}\left[\sup_{t\in T} X_t\right] = \mathrm{E}\left[\sup_{t\in T} (X_t - X_{t_0})\right] \le \mathrm{E}\left[\sup_{s,t\in T} (X_t - X_s)\right]$$

Then we take the limit $\delta \to 0$ and get

$$\mathrm{E}\left[\sup_{t\in T} X_t\right] \le 32 \int_0^\infty \sqrt{\log \mathcal{N}(\varepsilon, T, d_X)} \,\mathrm{d}\varepsilon.$$

The factor of 32 is not sharp and could be further improved by modifying the proof [37, p. 140].

The following is a lower bound counterpart to Dudley's inequality.

Theorem 2.28 (Sudakov's minoration inequality)**.** Let $(X_t)_{t\in T}$ be a mean zero Gaussian process. Then for any $\varepsilon \geq 0$, we have

$$\mathrm{E}\sup_{t\in T} X_t \geq c\varepsilon\sqrt{\log \mathcal{N}(\varepsilon, T, d_X)}$$

where d_X is the canonical distance induced by the process and c is a constant.

The general idea of the proof is to compare the increments of the process $\{X_t\}_{t\in T}$ to the increments of a simpler Gaussian process $\{Y_t\}_{t\in T'}$ defined by $Y_t := \frac{\varepsilon}{\sqrt{2}} g_t$, where T' is a certain finite subset of T and g_t are independent $\mathcal{N}(0,1)$ random variables. And then apply the Sudakov-Fernique inequality. For a proof of both Sudakov's minoration inequality and the Sudakov-Fernique inequality we refer the reader to [35].

Theorem 2.29 (Sudakov-Fernique inequality)**.** Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean zero Gaussian processes. Assume that for all $t, s \in T$, we have

$$\mathrm{E}\left[(X_t - X_s)^2\right] \leq \mathrm{E}\left[(Y_t - Y_s)^2\right].$$

Then

$$\mathrm{E}\left[\sup_{t\in T} X_t\right] \leq \mathrm{E}\left[\sup_{t\in T} Y_t\right].$$

Theorem 2.30 (Talagrand inequality)**.** [33, Thm. 2.4] Consider a Gaussian process $\{X_t\}_{t\in T}$. Let $\sigma^2 = \sup_{t\in T}\mathrm{E}[X_t^2]$. Consider the canonical distance on T given by $d_X(s,t)^2 = \mathrm{E}\left[(X_s - X_t)^2\right]$. Assume that for some constant $A > \sigma$, some $v > 0$ and some $0 \leq \varepsilon_0 \leq \sigma$ we have

$$\mathcal{N}(T, d, \varepsilon) \leq (A/\varepsilon)^v$$

whenever $\varepsilon < \varepsilon_0$. Then for $u \geq \sigma^2[(1+\sqrt{v})/\varepsilon_0]$ we have

$$\mathrm{P}\left(\sup_{t\in T} X_t \geq u\right) \leq \left(\frac{KAu}{\sqrt{v}\sigma^2}\right)^v \Phi(\frac{u}{\sigma}) \leq \left(\frac{KAu}{\sqrt{v}\sigma^2}\right)^v e^{\frac{-u^2}{2\sigma^2}},$$

where Φ denotes the CDF of the standard normal distribution. If $\varepsilon_0 = \sigma$, the condition on g is $g \geq \sigma[1+\sqrt{v}]$.

Chapter 3

Gaussian Process Regression and Bayesian Optimization

This chapter is focused on Gaussian process regression, as well as Bayesian optimization. Gaussian process regression is a method for interpolating data points with a Gaussian process governed by prior covariances. It originated in geostatistics, where it was used to predict the distribution of ore. It is also known as Kriging, named after the South African mining engineer D. G. Krige [20], who developed the first ideas in the 1950s. These were then picked up and worked on further in the seminal works of Georges Matheron in the 1960s [26], who coined the term Kriging. Figure 3.1 shows a Gaussian random field regression with a Matérn kernel.

Bayesian optimization is a method for minimizing functions that are expensive to evaluate. The method consists of two steps that are performed sequentially. First, a measurement point is picked by maximizing a utility function, making a trade-off between exploring areas of high uncertainty and exploiting areas of low predicted values. Secondly, a Gaussian process regression is computed based on the previous measurements and the random structure of the problem. This Gaussian process regression is then used to pick the next measurement point. The two steps are repeated until we have either reached a global minimum or a stopping criterion is met. Bayesian optimization gets used in many fields, including but not limited to machine learning hyperparamter tuning[40], molecule design[17, 19] and climate model calibration[25]. As mentioned in the introduction, we will only work with Gaussian processes from here on. Gaussian processes are expressive enough to be useful and simple enough to be worked with. They have a number of unique features that make them nice to work with. Namely, Gaussian processes are closed under addition, Bayesian conditioning (measurements) and linear operations. That means if X is a Gaussian process with mean $m(\cdot)$ and covariance $C(\cdot,\cdot)$, $\mathcal{L}$ is a linear operator, then $\mathcal{L} \circ X$ is a Gaussian process with mean $\mathcal{L} \circ m(\cdot)$ and covariance $\mathcal{L}^2 \circ C(\cdot,\cdot)$. In particular multiplications with matrices, differentiation and integration all return Gaussian processes. One noteworthy drawback when modeling with Gaussian processes is their conditional homoskedasticity. That is, the variance at a point $t \in T$, $\operatorname{Var}(X_t \mid X)$ does not depend on X [11, p. 110].

3.1 Kriging

Let $\{X_t\}_{t \in T}$ be a second order Gaussian random field. And let $t_1, \ldots, t_n \in T$ be measurement points. Kriging predicts the value of the random field at a points $t \in T$ as a weighted average of the measurements $X_{t_1}, \ldots, X_{t_n}$, so

$$\widehat{X}_t = \sum_{i=1}^{n} w_i X_{t_i} + c,$$

where w_i, c are certain weights, chosen, so that the linear predictor is unbiased and minimizes the mean squared error. In this section we will deal with simple Kriging. Its model assumptions are, that the mean μ is known and constant and that the random field is second order stationary with known covariance function C. Other variants of Kriging, like ordinary Kriging and universal Kriging [11], deal with situations, where these assumptions are not met.

First we deal with the unbiasedness condition. We can write the linear predictor as

$$\widehat{X}_t = w^T y + c$$

with $w = [w_1, \ldots, w_n]$ and $y = [X_{t_1}, \ldots, X_{t_n}]$. If we require $\widehat{X_t}$ to be unbiased,

$$\mu = \mathrm{E}[\widehat{X}_t] = \mathrm{E}[c + w^T y] = c + w^T(\mathbf{1}_n \mu).$$

Here we denote by $\mathbf{1}_n$ the vector of length n with all entries equal to one. The last equality is due to the observations y being distributed with mean μ. Thus the constant term must be

$$c = \mu - w^T(\mathbf{1}_n \mu)$$

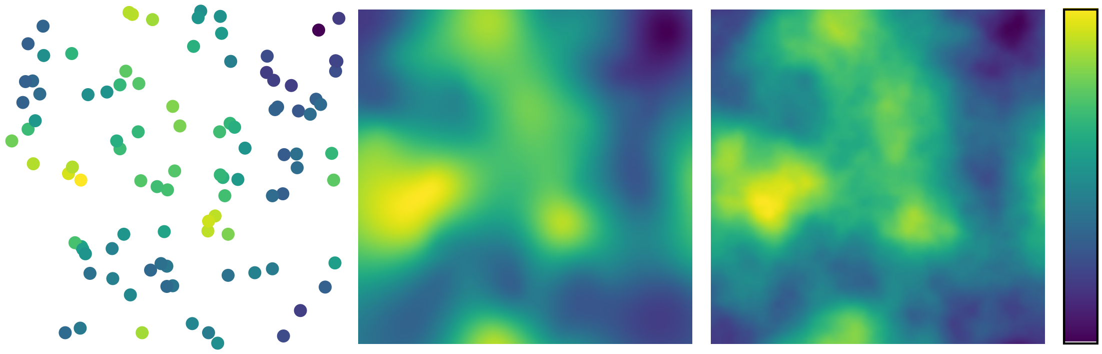

Figure 3.1 | Gaussian random field regression: Sampled points, regression, ground truth and color bar.

and the unbiased predictor is

$$\widehat{X}_t = \mu + w^T(y - (\mathbf{1}_n\mu)).$$

Before we deal with the minimization of the mean squared error, we will introduce some notation. Let us define the vector

$$\begin{bmatrix} X_t \\ \widehat{X}_M \end{bmatrix},$$

where X_t is the random variable obtained from X at a fixed $t \in T$ and $\widehat{X}_M$ is the random vector containing measurements $[X_{t_1}, \dots, X_{t_n}]$ at points $M = \{t_1, \dots, t_n\} \subset T$. This vector is Gaussian with mean

$$\mu = \begin{bmatrix} \mu_t \\ \mu_M \end{bmatrix}$$

and covariance

$$\Sigma = \begin{bmatrix} \Sigma_{t,t} & \Sigma_{t,M} \\ \Sigma_{M,t} & \Sigma_{M,M} \end{bmatrix},$$

where $\mu_M := [\mu_{t_1}, \dots, \mu_{t_n}]^T$ and the covariance block matrices are given by

$$\begin{aligned} \Sigma_{t,t} &= C(t,t) \\ \Sigma_{t,M} &= [C(t,t_i)]_{1 \le i \le n} \\ \Sigma_{M,t} &= (\Sigma_{t,M})^T \\ \Sigma_{M,M} &= [C(t_i,t_j)]_{1 \le i,j \le n}. \end{aligned}$$

Now we come to the minimization of the mean squared error. We can reformulate it as

$$\begin{aligned} \mathrm{MSE}(\widehat{X}_t) &= \mathrm{E}\left[\left(\widehat{X}_t - X_t\right)^2\right] \\ &= \mathrm{E}\left[(w^T(y-\mu) + (\mu - X_t))^2\right] \\ &= w^T\,\mathrm{E}[(y-\mu)(y-\mu)^T]w + \mathrm{E}[(\mu - X_t)^2] - 2\,\mathrm{E}[w^T(y-\mu)(X_t-\mu)] \\ &= w^T\Sigma_{M,M}w + \Sigma_{t,t} - 2w^T\Sigma_{M,t}. \end{aligned}$$

Differentiating with respect to w and setting the gradient to zero yields

$$\begin{aligned} 0 &= \frac{\partial}{\partial w}\left(w^T\Sigma_{M,M}w + \Sigma_{t,t} - 2w^T\Sigma_{M,t}\right) \\ &= 2\Sigma_{M,M}w - 2\Sigma_{M,t}, \end{aligned}$$

which is equivalent to

$$\Sigma_{M,M} w = \Sigma_{M,t}.$$

From this we derive the optimal weights $w = \Sigma_{M,M}^{-1}\Sigma_{M,t}$. And thus we have the Kriging predictor

$$\widehat{X}_t = \mu + \Sigma_{t,M}\Sigma_{M,M}^{-1}(y - (\mathbf{1}_n\mu)),$$

as well as the resulting mean squared error

$$\mathrm{MSE}(\widehat{X}_t) = \Sigma_{t,t} - \Sigma_{t,M}\Sigma_{M,M}^{-1}\Sigma_{M,t}.$$

Hence, the Kriging predictor is distributed akin to the conditional distribution of a Gaussian random vector, given the measurements, as described in Lemma 2.12. That is,

$$\widehat{X}_t \sim \mathcal{N}\left(\mu + \Sigma_{t,M}\Sigma_{M,M}^{-1}(y - (\mathbf{1}_n\mu)),\ \Sigma_{t,t} - \Sigma_{t,M}\Sigma_{M,M}^{-1}\Sigma_{M,t}\right). \tag{3.1}$$

As we have seen, Kriging provides the *best linear unbiased predictor* (BLUP). It is also an exact interpolator [11, p.359], meaning that the predictor is exact at the measurement points. One special property of Gaussian processes is, that the optimal predictor and optimal linear predictor are the same under squared error loss [11, p.110].

3.2 Bayesian Optimization

Bayesian optimization has its roots in the 60s and 70s with the works of Kushner [21] and Mockus [43]. But due to practical considerations, like the computational cost of inverting ill-conditioned $n \times n$ covariance matrices [42], it was not until the 90s that it gained popularity [18]. As mentioned in the introduction of the chapter, Bayesian optimization is a minimization scheme that consists of two key ingredients. The first is a probabilistic surrogate model, which captures our beliefs about the behavior of the unknown objective function and an observation model that describes the data generation mechanism. The second is a utility function, that describes, how optimal a sequence of observations is. The expected utility is then maximized to select an optimal sequence of observations, while after each observation the surrogate model is again updated. Computing the expected utility is often intractable, so heuristics are used to approximate it [30]. These heuristics are often called *acquisition functions*. This can be formalized as in Algorithm 1. The objective function f is usually expensive to evaluate and does not necessarily have a closed form expression. It is often nonconvex and multimodal. If gradient information is available, this can be incorporated into the algorithm [23, Sec. 4.2.1], but is beyond the scope of this thesis.

Algorithm 1 Bayesian Optimization

```
for $n = 1, 2, \dots$ do
    select new point $x_n$ to evaluate by maximizing the acquisition function $\alpha$
        $x_n = \operatorname{argmax}_{x \in \mathcal{X}} \alpha(x, \mathcal{D}_{n-1})$
    evaluate $f(x_n)$
    extend the data set $\mathcal{D}_n = \mathcal{D}_{n-1} \cup \{(x_n, f(x_n))\}$
    update the surrogate model
end for
```

With the previous section in mind, we already have a structure for a surrogate model, so let us look at the acquisition functions. We are looking to build a function α, such that we can choose the next evaluation point with

$$x_{n+1} = \operatorname*{argmax}_{x \in T} \alpha(x, \mathcal{D}_n),$$

where n is the number of previous observations, $\mathcal{D}_n$ is the data gained from the previous observations and T is the space over which we optimize. For a Gaussian process surrogate model $\{G_x\}_{x \in T}$ we define the improvement at $x \in T$ as

$$I(x) = \max(G_x - g^*, 0),$$

where g^* is the best solution encountered thus far. This is a random variable, so for one sample ω

$$I(x)(\omega) = \max(G_x(\omega) - g^*, 0).$$

We denote by Φ the CDF of the standard normal distribution. Then the probability of improvement is

$$\begin{aligned} \mathrm{P}(I(x) > 0) &= \mathrm{P}(G_x > g^*) \\ &= 1 - \mathrm{P}(G_x \leq g^*) \\ &= 1 - \Phi\left(\frac{g^* - \mu(x)}{\sigma(x)}\right) \\ &= \Phi\left(\frac{\mu(x) - g^*}{\sigma(x)}\right) \end{aligned}$$

since G_x is normally distributed. This gives us the probability of improvement acquisition function

$$\alpha_{\mathrm{PI}}(x) := \Phi\left(\frac{\mu(x) - g^*}{\sigma(x)}\right).$$

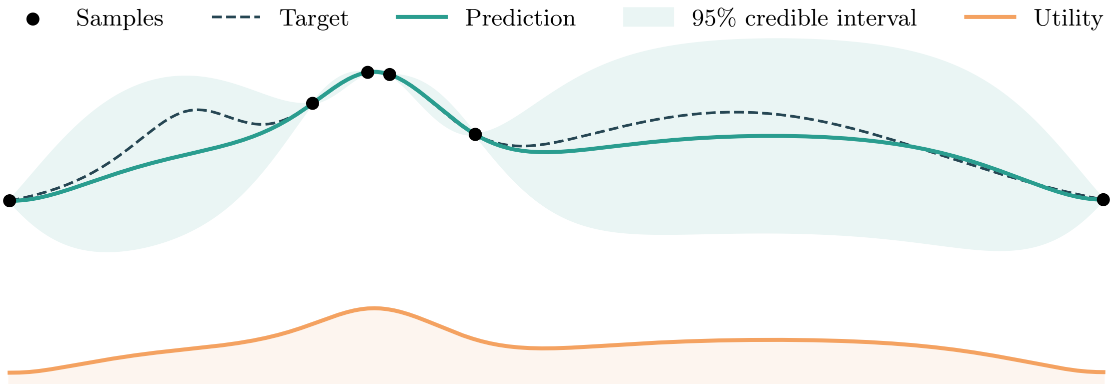

Figure 3.2 | Bayesian Optimization with Expected Improvement.

One drawback of this acquisition function is, that it does not account for the size of the improvement. This leads to the expected improvement, shown in Figure 3.2,

$$\alpha_{\mathrm{EI}} \coloneqq \mathrm{E}[I] = \mathrm{E}[\max(G_x - g^*, 0)].$$

To bring this into an explicit form, recall that for non-negative random variables X we have

$$\mathrm{E}[X] = \int_0^\infty \mathrm{P}(X > t)\,\mathrm{d}t.$$

Hence

$$\begin{aligned}\mathrm{E}[I] &= \int_0^\infty \mathrm{P}(I > t)\,\mathrm{d}t \\ &= \int_0^\infty \mathrm{P}(G_x > g^* + t)\,\mathrm{d}t \\ &= \int_0^\infty \Phi\left(\frac{\mu(x) - g^* - t}{\sigma(x)}\right)\,\mathrm{d}t.\end{aligned}$$

To solve this integral, we substitute $z(t) = \frac{\mu(x)-g^*-t}{\sigma(x)}$ with $z'(t) = -\frac{1}{\sigma(x)}$. Then

$$\mathrm{E}[I] = -\sigma \int_{-\infty}^0 \Phi(z)dz.$$

Integrating by parts yields

$$\mathrm{E}[I] = \sigma z \Phi(z(t)) + \varphi(z(t))\Big|_{-\infty}^0$$

, where φ is the PDF of the standard normal distribution. Plugging in φ and Φ and taking care of the limit, we arrive at the explicit form of the expected improvement

$$\mathrm{E}[I] = (\mu(x) - g^*)\Phi(\frac{\mu(x) - g^*}{\sigma(x)}) + \sigma\varphi(\frac{\mu(x) - g^*}{\sigma(x)}).$$

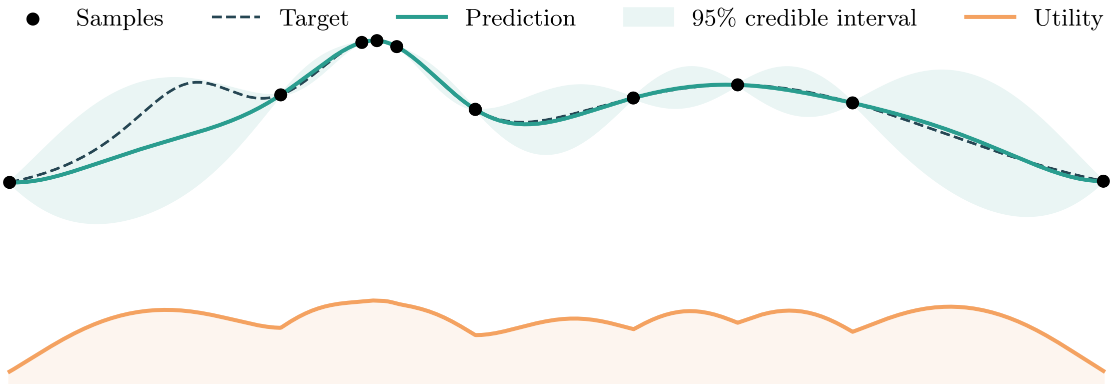

Figure 3.3 | Bayesian Optimization with Upper Confidence Bounds.

The expected improvement has monotonicity properties [18]

$$\frac{\partial \, \mathrm{EI}}{\partial u} = -\Phi(u) \left(\frac{f_n^* - u}{\sigma_n} \right) < 0$$

and

$$\frac{\partial \, \mathrm{EI}}{\partial \sigma_n} = \varphi(u) \left(\frac{f_n^* - u}{\sigma_n} \right) > 0.$$

This means, that the EI function is monotonely increasing in Kriging uncertainty and monotonely decreasing in Kriging prediction. We can nicely see, that the expected improvement is quantifying the trade-off between exploration and exploitation. Some convergence rates for expected improvement are established in [8]. One more acquisition function, shown in Figure 3.3, is the Upper Confidence Bound method [3]. It is given by

$$\alpha_{\mathrm{UCB}}(x, D) = -\mu_n(x) + \beta_n \sigma_n(x)$$

with a parameter β_n that controls the confidence level, μ_n the mean and σ_n the standard deviation of the surrogate model.

Having talked about the possible acquisition functions, let us briefly look at the practical considerations of picking the surrogoate model. Namely picking the kernel and its hyperparameters. This can be done in multiple ways. Take the Matérn kernel as an example. It has three hyperparameters, the amplitude α, the length scale ν and the smoothness μ, which we will denote by

$$\eta = (\alpha, \nu, \mu).$$

The first approach is to use maximum likelihood estimation (MLE) to find the hyperparameters that maximize the likelihood of the data.

$$\widehat{\eta} = \underset{\eta}{\operatorname{argmax}} \, \mathrm{P}(X_{x_1,\ldots,x_n} \mid \eta)$$

The second is to use a Bayesian approach and compute the posterior distribution (MAP) of the hyperparameters.

$$\begin{aligned}\widehat{\eta} &= \underset{\eta}{\operatorname{argmax}} \, \mathrm{P}(\eta \mid X_{x_1,\dots,x_n}) \\ &= \underset{\eta}{\operatorname{argmax}} \frac{\mathrm{P}(X_{x_1,\dots,x_n} \mid \eta)\mathrm{P}(\eta)}{\int P(X_{x_1,\dots,x_n} \mid \eta')\mathrm{P}(\eta') \, \mathrm{d}x} \\ &= \underset{\eta}{\operatorname{argmax}} \, \mathrm{P}(X_{x_1,\dots,x_n} \mid \eta)\mathrm{P}(\eta)\end{aligned}$$

The prior distribution can be picked uniformly or with expert knowledge in mind, depending on the application. Alternatively, one can marginalize over the hyperparameters for a fully Bayesian approach, but this is often intractable. This leads to the use of Markov Chain Monte Carlo (MCMC) methods to approximate the posterior distribution.

Chapter 4

Deriving a Uniform Bound

In this chapter we will derive an upper bound on the probability of further improvement of a Kriging process. For that we will use two different inequalities from Section 2.5. The first is Dudley's inequality, the second is Talagrand's inequality. In both cases we use the metric entropy of the space to find a bound on the variance of the process, then use this bound to derive a bound on the probability of further improvement. We limit ourselves to the case of Gaussian processes with Matérn covariances on underlying d-dimensional spaces $T = [0,1]^d$.

Suppose our underlying stationary covariance function is C. We write $C(x,y) = C(|x-y|)$ for $x, y \in T$. Further we write the Kriging covariance as $C(x, y \mid x_1, \dots, x_n)$ for x, y and $x_1, , \dot{x}_n \in T$. We know from Equation (3.1) that the Kriging process has a covariance matrix

$$\Sigma_{(x,y)|x_1,\dots,x_n} = \Sigma_{x,y} - \Sigma^T_{(x,y),x_1,\dots,x_n} \Sigma^{-1}_{(x_1,\dots,x_n),(x_1,\dots,x_n)} \Sigma_{(x,y),x_1,\dots,x_n}.$$

Writing out the matrices gives us

$$\begin{aligned}\Sigma_{(x,y)|x_1,\dots,x_n} = &\begin{bmatrix} C(x,x) & C(x,y) \\ C(x,y) & C(y,y) \end{bmatrix} \\ &- \begin{bmatrix} C(x,x_1) & \dots & C(x,x_n) \\ C(y,x_1) & \dots & C(y,x_n) \end{bmatrix} \Sigma^{-1} \begin{bmatrix} C(x,x_1) & C(y,x_1) \\ \vdots & \vdots \\ C(x,x_n) & C(y,x_n) \end{bmatrix},\end{aligned}$$

with $\Sigma^{-1} := \Sigma^{-1}_{(x_1,\dots,x_n),(x_1,\dots,x_n)} = [C(x_i), C(x_j)]^{-1}_{1 \le i,j \le n}$. And hence we can extract formulas for the Kriging covariance functions

$$C(x, y \mid x_1, \dots, x_n) = C(x,y) - \begin{bmatrix} C(x,x_1) & \dots & C(x,x_n) \end{bmatrix} \Sigma^{-1} \begin{bmatrix} C(y,x_1) \\ \vdots \\ C(y,x_n) \end{bmatrix}$$

and

$$C(x, x \mid x_1, \dots, x_n) = C(x,x) - \begin{bmatrix} C(x,x_1) & \dots & C(x,x_n) \end{bmatrix} \Sigma^{-1} \begin{bmatrix} C(x,x_1) \\ \vdots \\ C(x,x_n) \end{bmatrix}.$$

4.1 Preparations

For both the Dudley and Talagrand approaches we will need to bound the covering numbers as well as the variance of the Kriging process $\{X_t\}_{t\in T}$, so a bound on $\bar{\sigma}^2 := \sup_{t\in B} C(t,t) = \sup_{t\in B} \mathrm{E}[(X_t - \mathrm{E}[X_t])^2]$ for certain parts of the index set $B \subseteq T$. We start by bounding the covering numbers.

Lemma 4.1 (Hölder continuity of Matérn covariance)**.** Let C denote the Matérn covariance function with parameters $\nu > d/2$ and $m > 0$. For $0 < \eta < 2\nu - d$ there exists a constant $k > 0$ such that for all $s, t \in \mathbb{R}^d$

$$|C(s) - C(t)| \leq k|s-t|^\eta.$$

Proof. We will follow the proof from [13, Lemma 4.4]. Fix $\eta \in (0,1)$. Note that for $z, w \in \mathbb{C}$ with $|z-w| \leq 2$ we have

$$|z-w| = |z-w|^{1-\eta}|z-w|^\eta \leq 2^{1-\eta}|z-w|^\eta.$$

Further note that we have

$$|e^{-i\xi x} - e^{-i\xi y}| \leq 2$$

for all $x, y, \xi \in \mathbb{R}^d$ since e^{-it} is on the unit circle for $t \in \mathbb{R}^d$. By an application of the mean value theorem to $f(t) = e^{-i\xi t}$ we have

$$|e^{-i\xi x} - e^{-i\xi y}| \leq |\xi(x-y)|.$$

We can combine these facts and obtain

$$\begin{aligned} |C(s) - C(t)| &\leq \frac{1}{(2\pi)^d} \left| \int_{\mathbb{R}^d} \frac{e^{-i\xi x} - e^{-i\xi y}}{(|\xi|^2 + m^2)^\nu} \,\mathrm{d}\xi \right| \\ &\leq \frac{2^{1-\eta}}{(2\pi)^d} |x-y|^\eta \int_{R^d} \frac{|\xi|^\eta}{(|\xi|^2 + m^2)^\nu} \,\mathrm{d}\xi. \end{aligned}$$

To answer the remaining question of when this integral is finite, we bound

$$\int_{\mathbb{R}^d} \frac{|\xi|^\eta}{(|\xi|^2 + m^2)^\nu} \,\mathrm{d}\xi \leq \int_{\mathbb{R}^d} \frac{1}{|\xi|^{2\nu - \eta}} \,\mathrm{d}\xi.$$

This integral is finite for $2\nu - \eta > d$, which proves the desired inequality. □

Proposition 4.2 (Bounding the covering numbers)**.** Let C be a Matérn covariance function C with parameters ν and m. Let $\{G_t\}_{t\in T}$ be a Kriging process with measurements $x_1, \ldots, x_n$. We again write $C(x,y) := C(|x-y|)$ for the unconditioned process as well as $C(x, y \mid x_1, \ldots, x_n) := \mathrm{Cov}(G_x, G_y)$ for the Kriging process. There exist constants A and α, such that the covering numbers of any subset $B \subseteq T$ for $\varepsilon > 0$ with regards to the canonical distance are bounded by

$$\mathcal{N}(B, d_G, \varepsilon) \leq (\frac{A}{\varepsilon})^\alpha.$$

Proof. We adapt a proof from [13] to the setting of Kriging processes. We will first show Hölder continuity of the canonical distance of the Kriging process. This will then help us to cover any ε-balls of the canonical distance with certain $\widetilde{\varepsilon}$-balls of the euclidean norm. Then we can bound the covering number $\mathcal{N}(B, d_G, \varepsilon)$ by the covering number $\mathcal{N}(B, |\cdot|, \widetilde{\varepsilon})$, which is easy to bound as a function of the diameter of B. Let us begin with proving the Hölder continuity of the canonical distance d_G. As a convenience we define $c_x = \begin{bmatrix} C(x, x_1) & \dots & C(x, x_n) \end{bmatrix}^T$ for $x \in T$. We can write the canonical distance for $x, y \in T$ as

$$\begin{aligned} d_G(x,y)^2 &= \operatorname{Var}(G_x - G_y) \\ &= \operatorname{Var}(G_x) + \operatorname{Var}(G_y) - 2\operatorname{Cov}(G_x, G_y) \\ &= C(x, x \mid x_1, \dots, x_n) + C(y, y \mid x_1, \dots, x_n) - 2C(x, y \mid x_1, \dots, x_n) \\ &= C(0) - c_x^T \Sigma^{-1} c_x + C(0) - c_y^T \Sigma^{-1} c_y - 2C(|x-y|) + 2c_x^T \Sigma^{-1} c_y \\ &= 2(C(0) - C(|x-y|)) - (c_x - c_y)^T \Sigma^{-1} (c_x - c_y), \end{aligned}$$

where we used the fact that Σ^{-1} is symmetric to transpose $c_x^T \Sigma^{-1} c^y = c_y^T \Sigma^{-1} c^x$. Now since Σ^{-1} is also positive definite, $(c_x - c_y)^T \Sigma^{-1} (c_x - c_y) \geq 0$ and the $d_G(x,y)^2$ can be bounded by

$$d_G(x,y)^2 \leq 2(C(0) - C(|x-y|)).$$

Applying the Hölder continuity of Lemma 4.1, we have

$$d_G(x,y) \leq \sqrt{2k}|x-y|^{\eta/2}$$

for $0 < \eta < 2\nu - d$ with a constant k.

We are ready to deal with the covering numbers. Due to the bound on the canonical distance by the Euclidean norm, we can cover any $\widetilde{\varepsilon}$-balls of the Euclidean distance with ε-balls of the canonical distance. To be precise, we have

$$B_{|\cdot|,\widetilde{\varepsilon}}(x) \subseteq B_{d_G,\varepsilon}(x)$$

with $\widetilde{\varepsilon} = \left(\frac{\varepsilon^2}{2k}\right)^{1/\nu}$ Therefore any covering of B with a set of $B_{|\cdot|,\widetilde{\varepsilon}}(x)$ balls gives us a covering of B with the same number of $B_{d_G,\varepsilon}(x)$ balls. Hence

$$\mathcal{N}(B, d_G, \varepsilon) \leq \mathcal{N}(B, |\cdot|, \widetilde{\varepsilon}).$$

As is shown in [35, Cor. 4.2.13], the latter can be bounded by

$$\mathcal{N}(B, |\cdot|, \widetilde{\varepsilon}) \leq \left(\frac{2\operatorname{diam}(B)\sqrt{d}}{\widetilde{\varepsilon}}\right)^d,$$

where d is the dimension of T. Putting everything together we have

$$\begin{aligned}\mathcal{N}(B, d_G, \varepsilon) &\leq \left(\frac{2\,\mathrm{diam}(B)\sqrt{d}}{\widetilde{\varepsilon}}\right)^d \\ &= \left(\frac{(2\,\mathrm{diam}(B)\sqrt{d})^\eta 2k}{\varepsilon^2}\right)^{\frac{d}{\eta}} \\ &= \left(\frac{\sqrt{(2\,\mathrm{diam}(B)\sqrt{d})^\eta 2k}}{\varepsilon}\right)^{\frac{2d}{\eta}}.\end{aligned}$$

Therefore the bound

$$\mathcal{N}(B, d_G, \varepsilon) \leq (\frac{A}{\varepsilon})^\alpha$$

holds, with $\alpha = \frac{2d}{\eta}$ and $A = \Big((2\,\mathrm{diam}(B)\sqrt{d})^\eta 2k\Big)^{1/2}$. □

Having derived a bound on the covering numbers, let us turn to bounding the process variance $\bar{\sigma}^2 = \sup_{x \in B} C(x, x)$.

Lemma 4.3. Let $C(x, y \mid x_1, \ldots, x_n)$ with $n \in \mathbb{N}$ and $x, y \in T$ be the covariance function of a Kriging process with an underlying Matérn covariance and measurements $x_1, \ldots, x_n \in T$. Then the Kriging variance is monotonely decreasing with added measurements, i.e.

$$C(x, x \mid x_1, \ldots, x_n) \leq C(x, x \mid x_2, \ldots x_n).$$

Proof. Before we look at the Kriging covariance functions, we will need a small matrix computation. We define the covariance matrix

$$\Sigma = \begin{bmatrix} C(x_1, x_1) & \ldots & C(x_1, x_n) \\ \vdots & \ddots & \vdots \\ C(x_n, x_1) & \ldots & C(x_n, x_n) \end{bmatrix} = \begin{bmatrix} C(x_1, x_1) & \Sigma_B \\ \Sigma_B^T & \Sigma_C \end{bmatrix}$$

with $\Sigma_C \in \mathbb{R}^{n-1,n-1}$ and $\Sigma_B \in \mathbb{R}^{n-1}$. Since C is a Matérn covariance function, the matrices Σ, Σ_C are positive definite and $C(x_1, x_1) > 0$. Hence Σ and Σ_C are invertible. We write

$$\Sigma^{-1} = \Lambda = \begin{bmatrix} \Lambda_A & \Lambda_B \\ \Lambda_B^T & \Lambda_C \end{bmatrix}.$$

with $\Lambda_A \in \mathbb{R}$ and all other dimensions to match. Computing the blocks of this inverse matrix gives us

$$\Lambda_A = \left(C(x_1, x_1) - \Sigma_B \Sigma_C^{-1} \Sigma_B^T\right)^{-1}$$

$$\begin{aligned}\Lambda_B &= -\left(C(x_1,x_1) - \Sigma_B\Sigma_C^{-1}\Sigma_B^T\right)^{-1}\Sigma_B\Sigma_C^{-1}\\ \Lambda_C &= \Sigma_C^{-1} + \Sigma_C^{-1}\Sigma_B^T\left(C(x_1,x_1) - \Sigma_B\Sigma_C^{-1}\Sigma_B^T\right)^{-1}\Sigma_B\Sigma_C^{-1}.\end{aligned}$$

Now according to [15, Prop. 2.2] applied to Σ the Schur complement $\Sigma/\Sigma_C = C(x_1,x_1) - \Sigma_B\Sigma_C^{-1}\Sigma_B^T$ is positive definite, therefore $C(x_1,x_1) - \Sigma_B\Sigma_C^{-1}\Sigma_B^T > 0$. With this established, we can turn our attention to the Kriging covariance functions. By definition we have

$$C(x, x \mid x_1, \ldots x_n) = C(x,x) - c_x^T\Lambda c_x$$

and

$$C(x, x \mid x_2, \ldots x_n) = C(x,x) - \widetilde{c}_x^T\Sigma_C^{-1}\widetilde{c}_x$$

with $c_x := \begin{bmatrix} C(x,x_1) & \ldots & C(x,x_n)\end{bmatrix}^T$ and $\widetilde{c}_x := \begin{bmatrix} C(x,x_2) & \ldots & C(x,x_n)\end{bmatrix}^T$. To show the desired monitonicity, it thus suffices to show

$$c_x^T\Lambda c_x \geq \widetilde{c}_x^T\Sigma_C^{-1}\widetilde{c}_x.$$

We start by multiplying out the left hand side

$$\begin{aligned}c_x^T\Lambda c_x &= C(x,x_1)\Lambda_A C(x,x_1) + C(x,x_1)\Lambda_B\widetilde{c}_x + \widetilde{c}_x^T\Lambda_B^T C(x,x_1) + \widetilde{c}_x^T\Lambda_C\widetilde{c}_x\\ &= C(x,x_1)^2\Lambda_A + 2C(x,x_1)\Lambda_B\widetilde{c}_x + \widetilde{c}_x^T\Lambda_C\widetilde{c}_x.\end{aligned}$$

Plugging in from the inverse block matrix Λ we have

$$\begin{aligned}c_x^T\Lambda c_x = &\frac{1}{C(x_1,x_1) - \Sigma_B\Sigma_C\Sigma_B^T}(C(x,x_1)^2 - 2C(x,x_1)\Sigma_B\Sigma_C^{-1}\widetilde{c}_x\\ &+ \widetilde{c}_x\Sigma_C^{-1}\Sigma_B^T\Sigma_B\Sigma_C^{-1}\widetilde{c}_x) + \widetilde{c}_x^T\Sigma_C^{-1}\widetilde{c}_x\\ = &\frac{1}{C(x_1,x_1) - \Sigma_B\Sigma_C\Sigma_B^T}\left(C(x,x_1) - \Sigma_B\Sigma_C^{-1}\widetilde{c}_x\right)^2 + \widetilde{c}_x^T\Sigma_C^{-1}\widetilde{c}_x\\ \geq &\,\widetilde{c}_x^T\Sigma_C^{-1}\widetilde{c}_x,\end{aligned}$$

which proves the monotonicity. □

Now let us look specifically at the case with only one measurement x_1. We can write the covariance matrix as

$$\Sigma_{(x,y)|x_1} = \begin{bmatrix} C(x,x \mid x_1) & C(x,y \mid x_1)\\ C(x,y \mid x_1) & C(x,y \mid x_1)\end{bmatrix}.$$

As before,

$$\Sigma_{(x,y)|x_1} = \Sigma_{x,y} - \Sigma_{(x,y),x_1}^T\Sigma_{x_1,x_1}^{-1}\Sigma_{(x,y),x_1}$$

$$= \begin{bmatrix} C(x,x) & C(x,y) \\ C(x,y) & C(y,y) \end{bmatrix} - \frac{1}{C(x_1,x_1)} \begin{bmatrix} C(x,x_1)^2 & C(x,x_1)C(y,x_1) \\ C(x,x_1)C(y,x_1) & C(y,x_1)^2 \end{bmatrix}.$$

Thus for a stationary covariance function C, we have the formula

$$C(x, x \mid x_1) = C(0) - \frac{C(x,x_1)^2}{C(0)}.$$

Lemma 4.4 (Bounding the variance)**.** Take a Matérn covariance Kriging process with measurements $x_1, \ldots, x_n \in T$. Let $B \subseteq T$ be a set, such that for every $x \in B$ there exists a measurement point x_i with $|x - x_i| \leq \varepsilon$. The variance of the process is bounded by

$$C(x, x \mid x_1, \ldots x_n) \leq 2k\varepsilon^{\eta}$$

with the constants k, η from Lemma 4.1 for all $x \in B$

Proof. To prove this bound we will combine the monotonicity of Lemma 4.3 with the bound of Lemma 4.1. For any $x \in B$ pick the closest measurement x_i. Then

$$\begin{aligned} C(x, x \mid x_1, \ldots x_n) &\leq C(x, x \mid x_i) \\ &= \frac{C(0)^2 - C(x,x_i)^2}{C(0)} \\ &= \frac{(C(0) - C(x,x_i))(C(0) + C(x,x_i))}{C(0)} \\ &\leq 2(C(0) - C(x,x_i)). \end{aligned}$$

The last inequality holds because of the decreasing nature of the Matérn covariance $C(0) \geq C(x, x_i)$. We can now apply Lemma 4.1 and have

$$C(x, x \mid x_1, \ldots x_n) \leq 2k|x - x_i|^{\eta}$$

with constants k, η from Lemma 4.1. Knowing that our measurement covers the space well enough, we can therefore uniformly bound the Kriging variance with

$$C(x, x \mid x_1, \ldots x_n) \leq 2k\varepsilon^{\eta}. \qquad \square$$

4.2 Bounds on the chance of further improvement

Theorem 4.5 (Chance of further improvement with Dudley)**.** Let $\{X_t\}_{t \in T}$ be a Kriging process with measurements $x_1, \ldots, x_n$ and Matérn covariance C with parameters ν, m. Let $B \subseteq T$ be a set, such that for every $x \in B$ there exists a measurement point x_i with $|x - x_i| \leq \varepsilon$ for some ε. Define $\bar{\sigma}^2 = \sup_{x \in B} C(x, x \mid x_1, \ldots, x_n)$, as well as the constant A and α as in Proposition 4.2 applied to $B \subseteq T$.

Then $\bar{\sigma}^2 \leq 2k\varepsilon^\eta$ and the probability of finding a value greater than $h \geq 12A\sqrt{\pi}\sqrt{\alpha}$ is bounded by

$$\mathrm{P}(\sup_{t\in B} X_t > h) \leq 2\exp\left(\frac{(h-12A\sqrt{\pi}\sqrt{\alpha})^2}{4\pi^2 k\varepsilon^\eta}\right).$$

Proof. We can apply Theorem 2.26 to $B \subseteq T$ to get

$$\mathrm{E}\left[\sup_{t\in B}|X_t|\right] \leq 24\int_0^\infty \sqrt{\log(\mathcal{N}(B, d_X, \varepsilon))}\,\mathrm{d}\varepsilon.$$

The covering numbers for $\varepsilon \geq 0$ are bounded due Proposition 4.2 with $\mathcal{N}(B, d_X, \varepsilon) \leq \left(\frac{A}{\varepsilon}\right)^\alpha$. For $\varepsilon \geq A$ this is bounded by $\left(\frac{A}{\varepsilon}\right)^\alpha \leq 1$, which means the covering number must be 1 and we can estimate

$$\mathrm{E}\left[\sup_{t\in B}|X_t|\right] \leq 24\int_0^A \sqrt{\log\left(\frac{A}{\varepsilon}\right)^\alpha}\,\mathrm{d}\varepsilon.$$

To solve this integral, we first note that

$$f(\varepsilon) = \sqrt{\log(\frac{A}{\varepsilon})^\alpha}$$

is injective with

$$f^{-1}(y) = Ae^{-\frac{y^2}{\alpha}}$$

Then we apply Fubini's theorem in order to integrate over the y-axis

$$\begin{aligned}\int_0^A \sqrt{\log(\frac{A}{\varepsilon})^\alpha}\,\mathrm{d}\varepsilon &= \int_0^\infty \int_0^A \chi_{y\leq f(\varepsilon)}\,\mathrm{d}\varepsilon\,\mathrm{d}y \\ &= \int_0^\infty \int_0^A \chi_{f^{-1}(y)\leq\varepsilon}\,\mathrm{d}\varepsilon\,\mathrm{d}y \\ &= \int_0^\infty f^{-1}(y)\,\mathrm{d}y \\ &= \int_0^\infty Ae^{-\frac{y^2}{\alpha}}\,\mathrm{d}y\end{aligned}$$

where χ is the indicator function. This integral is a scaled Gaussian integral that can be computed by polar coordinates, leading to

$$\int_0^\infty Ae^{-\frac{y^2}{\alpha}}\,\mathrm{d}y = A\frac{\sqrt{\pi}\sqrt{\alpha}}{2}.$$

Now we can simplify the estimate to

$$\mathrm{E}\left[\sup_{t\in B}|X_t|\right] \leq 12A\sqrt{\pi}\sqrt{\alpha}.$$

Next we will us use the concentration inequality from Proposition 2.23. The first consideration is the seperability of $\{X_t\}_{t\in B}$. Here we can use the fact, that the seperability the process is equivalent to the seperability of (B, d_X). Which is given due to the seperability of $B \subseteq T$ with the Euclidean norm, as well as the Hölder continuity of the canonical distance $d_X(x,y) \leq 2k|x-y|^\eta$. The second consideration is the centeredness. If $\{X_t\}_{t\in B}$ is not centered, we can subtract the Kriging mean function μ_t and work with $\{X_t - \mu_t\}_{t\in T}$ instead. Due to Lemma 4.4 the variance $\bar{\sigma}^2$ is finite and bounded by $\bar{\sigma}^2 \leq 2k\varepsilon^\eta$. Therefore by the concentration inequality

$$\mathrm{P}\left(\left|\sup_{t\in B}|X_t| - \mathrm{E}\left[\sup_{t\in B}|X_t|\right]\right| > u\right) \leq 2e^{-(u^2/2\pi^2\bar{\sigma}^2)}.$$

for $u > 0$. Further

$$\mathrm{P}\left(\sup_{t\in B} X_t > u + \mathrm{E}\left[\sup_{t\in B}|X_t|\right]\right) \leq \mathrm{P}\left(\left|\sup_{t\in B}|X_t| - \mathrm{E}\left[\sup_{t\in B}|X_t|\right]\right| > u\right).$$

We can substitute $h = u + \mathrm{E}\left[\sup_{t\in B}|X_t|\right]$. Putting the two inequalites together

$$\mathrm{P}(\sup_{t\in B} X_t > h) \leq 2\exp\left(\frac{(h - \mathrm{E}\left[\sup_{t\in B}|X_t|\right])^2}{2\pi^2\bar{\sigma}^2}\right) \leq 2\exp\left(\frac{(h - 12A\sqrt{\pi}\sqrt{\alpha})^2}{2\pi^2\bar{\sigma}^2}\right)$$

holds for $h \geq 12A\sqrt{\pi}\sqrt{\alpha}$. Plugging in the estimate for $\bar{\sigma}^2$ proves the desired inequality. □

Theorem 4.6 (Chance of further improvement with Talagrand)**.** Let $\{X_t\}_{t\in T}$ be a Kriging process with measurements $x_1, \ldots, x_n$ and Matérn covariance C with parameters ν, m. Let $B \subseteq T$ be a set, such that for every $x \in B$ there exists a measurement point x_i with $|x - x_i| \leq \varepsilon$ for some ε. Define $\bar{\sigma}^2 = \sup_{x\in B} C(x, x \mid x_1, \ldots, x_n)$, as well as the constant A and α as in Proposition 4.2. Further choose $\widetilde{A} > \max\{\sqrt{2k\varepsilon^\eta}, A\}$. Then the probability of finding a value greater than u is bounded by

$$\mathrm{P}\left(\sup_{t\in B} X_t \geq u\right) \leq \left(\frac{KAu}{\sqrt{\alpha}\sigma^2}\right)^\alpha e^{\frac{-u^2}{2\sigma^2}}$$

for $u \geq \bar{\sigma}(1 + \sqrt{\alpha})$.

Proof. We will check all the neccessary conditions needed to apply Theorem 2.30. By Lemma 4.4 the variance of the process is bounded with $\sigma^2 \leq 2k\delta^\eta$. By the choice of $\widetilde{A}$ we therefore have $\widetilde{A} > \sigma$. Further we have the bound on the covering numbers

$$\mathcal{N}(B, d_X, \varepsilon) \leq \left(\frac{A}{\varepsilon}\right)^{\alpha} \leq \left(\frac{\widetilde{A}}{\varepsilon}\right)^{\alpha}$$

from Proposition 4.2. We can now apply Lemma 4.4, which gives us

$$\mathrm{P}\left(\sup_{t \in B} X_t \geq u\right) \leq \left(\frac{KAu}{\sqrt{\alpha}\sigma^2}\right)^{\alpha} e^{\frac{-u^2}{2\sigma^2}},$$

for $u \geq \bar{\sigma}(1 + \sqrt{\alpha})$ □

4.3 ALGORITHM

The following is an example of how this bound could be implemented in an algorithm. We start like regular Bayesian optimization, by finding an optimal measurement point according to some acquisition function. But in between taking measurements we use the bound on the chance of further improvement to check, if there are areas where further improvement is unlikely. We then prune away these areas and only search in the remaining parts. If calculating these bounds is numerically cheaper than searching in the full area, then there is a performance gain in comparison to regular Bayesian optimization defined in Algorithm 1. In each iteration we perform a measurement and try to prune away parts of the search space T, where improvement is no longer likely. This can be done recursively for finer and finer subsets. The variables d_i control the depth of search after each measurement point. The value describes how many times we cut T in half when looking for areas to rule out. One could set $d_i = 0$ periodically if taking multiple measurements in a row is desired. This depends on the cost of measurement in comparison to the cost of computing the bounds on the chance of improvement. To illustrate the procedure, we look at an example. Take $T = [0,1]^d$. We can cut into $[0,0.5] \times [0,1]^d$ and $[0.5,1] \times [0,1]^{d-1}$ on the first level, then $[0,0.5] \times [0,0.5] \times [0,1]^{d-2}$ and so on until we have exhausted all the dimensions and arrive at blocks like form $[0,0.5]^d$, spread around T. Then we can start by cutting the first components in half again and repeat.

Algorithm 2 Branching augmented global optimization

```
Inputs:
    surrogate model $\{G_t\}_{t \in T}$
    acquisition function $\alpha$
    probability threshold $\gamma$
    pruning search depths $d_i$
    a function $f$ a set $T$ a scheme for cutting $T$ into half spaces recursively
Initialize:
    $T_{pr} \leftarrow \emptyset$
    $\mathcal{D}_0 \leftarrow \emptyset$
    $g^* \leftarrow -\infty$
for $i = 1, 2, \dots$ do
    select next point $x_i \leftarrow \operatorname{argmax}_{x \in T \setminus T_{pr}} \alpha(x, \mathcal{D}_{i-1})$
    evaluate $f(x_i)$
    if $f(x_i) > g^*$ then
        $g^* \leftarrow f(x_i)$
    end if
    extend the data set $\mathcal{D}_i \leftarrow \mathcal{D}_{i-1} \cup \{(x_i, f(x_i))\}$
    update the surrogate model $\{G_t\}_{t \in T}$
    for $j = 0, \dots, d_i$ do
        for all depth $j$ half spaces $T_k$ of $T$ with $T_k \cap T_{pr} = \emptyset$ do
            if $\mathrm{P}\left(\sup_{t \in T_k} G_t > g^*\right) < \frac{\mathrm{vol}(T_k)}{\mathrm{vol}(T)} \gamma$ then
                $T_{pr} \leftarrow T_{pr} \cup T_k$
            end if
        end for
    end for
    if $T_{pr} = T$ then
        terminate algorithm early
    end if
end for
```

Chapter 5

Conclusion and Outlook

5.1 Conclusion

In this thesis we derived a chaining based bound on the chance of further improvement during Bayesian optimization. We suggested two approaches, one based on Dudley's inequality and a concentration inequality. And one based on Talagrand's inequality. We proposed an algorithm to implemement this a check of this this bound, in order to improve computation times of Bayesian optimization by ruling out areas, where further improvement is unlikely.

5.2 Outlook

In future work one could implement Algorithm 2 and check its numerical performance, with emphasis on smoothness parameters and dimensionality of the data. One could compare how well the two different approaches with Dudley's and with Talagrand's inequality respectively rule out areas, where further improvement is unlikely. Another possible direction of research is to adapt this technique for other schemes, like gradient enhanced kriging.

Bibliography

[1] Milton Abramowitz and Irene A Stegun. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Vol. 55. US Government printing office, 1968.

[2] Robert J Adler and Jonathan E Taylor. *Random fields and geometry.* Springer Science & Business Media, 2007.

[3] Peter Auer. “Using confidence bounds for exploitation-exploration trade-offs”. In: *Journal of Machine Learning Research* 3.Nov (2002), pp. 397–422.

[4] C Bishop. *Pattern recognition and machine learning.* Springer, 2006.

[5] Salomon Bochner. “Monotone funktionen, stieltjessche integrale und harmonische analyse”. In: *Mathematische Annalen* 108.1 (1933), pp. 378–410.

[6] Anton Bovier. *Stochastic Processes.* Lecture notes, Universität Bonn, 2022.

[7] Pierre Brémaud. *Probability theory and stochastic processes.* Springer, 2020.

[8] Adam D Bull. “Convergence rates of efficient global optimization algorithms.” In: *Journal of Machine Learning Research* 12.10 (2011).

[9] Zexun Chen, Jun Fan, and Kuo Wang. “Remarks on multivariate Gaussian process”. In: *arXiv preprint arXiv:2010.09830* (2020).

[10] Emile Contal, Cédric Malherbe, and Nicolas Vayatis. “Optimization for gaussian processes via chaining”. In: *arXiv preprint arXiv:1510.05576* (2015).

[11] Noel Cressie. *Statistics for spatial data.* John Wiley & Sons, 1993.

[12] Nando De Freitas, Alex Smola, and Masrour Zoghi. “Exponential regret bounds for Gaussian process bandits with deterministic observations”. In: *arXiv preprint arXiv:1206.6457* (2012).

[13] Oliver G Ernst et al. “Integrability and Approximability of Solutions to the Stationary Diffusion Equation with Lévy Coefficient”. In: *arXiv preprint arXiv:2010.14912* (2020).

[14] Peter I Frazier. “A tutorial on Bayesian optimization”. In: *arXiv preprint arXiv:1807.02811* (2018).

[15] Jean H Gallier. “The schur complement and symmetric positive semidefinite (and definite) matrice”. In: (2019).

[16] Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[17] Ryan-Rhys Griffiths and José Miguel Hernández-Lobato. “Constrained Bayesian optimization for automatic chemical design using variational autoencoders”. In: *Chemical science* 11.2 (2020), pp. 577–586.

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

[19] Ksenia Korovina et al. “Chembo: Bayesian optimization of small organic molecules with synthesizable recommendations”. In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 3393–3403.

[20] Daniel G Krige. “A statistical approach to some basic mine valuation problems on the Witwatersrand”. In: *Journal of the Southern African Institute of Mining and Metallurgy* 52.6 (1951), pp. 119–139.

[21] Harold J Kushner. “A versatile stochastic model of a function of unknown and time varying form”. In: *Journal of Mathematical Analysis and Applications* 5.1 (1962), pp. 150–167.

[22] Eugene L Lawler and David E Wood. “Branch-and-bound methods: A survey”. In: *Operations research* 14.4 (1966), pp. 699–719.

[23] Daniel James Lizotte. “Practical bayesian optimization”. In: (2008).

[24] Tzon-Tzer Lu and Sheng-Hua Shiou. “Inverses of 2× 2 block matrices”. In: *Computers & Mathematics with Applications* 43.1-2 (2002), pp. 119–129.

[25] Jinfeng Ma et al. “Using Bayesian optimization to automate the calibration of complex hydrological models: Framework and application”. In: *Environmental Modelling & Software* 147 (2022), p. 105235.

[26] Georges Matheron. “Principles of geostatistics”. In: *Economic geology* 58.8 (1963), pp. 1246–1266.

[27] Rémi Munos. “Optimistic optimization of a deterministic function without the knowledge of its smoothness”. In: *Advances in neural information processing systems* 24 (2011).

[28] Emilio Porcu et al. “The Matérn Model: A Journey through Statistics, Numerical Analysis and Machine Learning”. In: *arXiv preprint arXiv:2303.02759* (2023).

[29] S. Sakata, F. Ashida, and M. Zako. “Structural optimization using Kriging approximation”. In: *Computer methods in applied mechanics and engineering* 192.7-8 (2003), pp. 923–939.

[30] Bobak Shahriari et al. “Taking the human out of the loop: A review of Bayesian optimization”. In: *Proceedings of the IEEE* 104.1 (2015), pp. 148–175.

[31] Joram Soch et al. *StatProofBook/StatProofBook.github.io: StatProofBook 2023.* Version 2023. Jan. 2024.

[32] Niranjan Srinivas et al. “Gaussian process optimization in the bandit setting: No regret and experimental design”. In: *arXiv preprint arXiv:0912.3995* (2009).

[33] Michel Talagrand. “Sharper bounds for Gaussian and empirical processes”. In: *The Annals of Probability* (1994), pp. 28–76.

[34] Terence Tao. *An introduction to measure theory.* Vol. 126. American Mathematical Soc., 2011.

[35] Roman Vershynin. *High-dimensional probability.* 2020.

[36] Giuseppe Vitali. *Sul problema della misura dei Gruppi di punti di una retta: Nota.* Tip. Gamberini e Parmeggiani, 1905.

[37] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint.* Vol. 48. Cambridge university press, 2019.

[38] Ziyu Wang et al. “Bayesian multi-scale optimistic optimization”. In: *Artificial Intelligence and Statistics.* PMLR. 2014, pp. 1005–1014.

[39] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning.* Vol. 2. 3. MIT press Cambridge, MA, 2006.

[40] Jia Wu et al. “Hyperparameter optimization for machine learning models based on Bayesian optimization”. In: *Journal of Electronic Science and Technology* 17.1 (2019), pp. 26–40.

[41] Yimin Xiao. “Uniform modulus of continuity of random fields”. In: *Monatshefte für Mathematik* 159.1-2 (2010), pp. 163–184.

[42] Anatoly Zhigljavsky and Antanas Žilinskas. *Bayesian and high-dimensional global optimization.* Springer, 2021.

[43] A Žilinskas and J Mockus. “On a Bayes method for seeking an extremum”. In: *Automatika i vychislitelnaja tekhnika* 3 (1972).

Appendix

A.1 Notation

Throughout this thesis, the following notation is used.

Symbol	Meaning
$:=$, $:\Leftrightarrow$	defined to be equal to, defined to be equivalent to
$\mathbb{N}$	Set of natural numbers including zero
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^n$	n-dimensional real vector space, $n \in \mathbb{N} \setminus \{0\}$
$\mathbb{R}^{n,m}$	set of real $n \times m$ matrices, $n \in \mathbb{N} \setminus \{0\}$
i	imaginary unit, $i^2 = -1$
$\mathbb{C}$	set of complex numbers
$\overline{z}$	complex conjugate of $z \in \mathbb{C}$
$\det(A)$	determinant of $A \in \mathbb{C}^{n,n}$
$\mathrm{E}[f]$	expected value of f
$\mathrm{Var}(f)$	variance of f
P	probability measure
$L^p(I, \mathbb{R}^n)$	measurable functions f with $\int_I \lvert f(x)\rvert^p \, \mathrm{d}x < \infty$
L^p	shorthand for $L^p(I, \mathbb{R}^n)$
$\mathcal{N}(T, d, \varepsilon)$	covering number
$\mathcal{N}$	shorthand for $\mathcal{N}(T, d, \varepsilon)$
$\mathrm{diam}(T)$	diameter of a set T given a metric d
∂f	subdifferential of f, derivative of f if it contains only one element
$\mathcal{P}(\Omega)$	power set of a set Ω
$\sigma(A)$	σ-algebra generated by A
$\mathcal{B}$	Borel σ-algebra

A.2 List of Figures

A.3 Alphabetical Index