AutoML: Gaussian Processes

Gaussian Processes

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

Weight-Space View

- So far, we have considered a hypothesis space $\mathcal{H}$ of parameterized functions $f(\mathbf{x} \mid \boldsymbol{\theta})$ (in particular, the space of linear functions).

- Using Bayesian inference, we derived distributions for $\boldsymbol{\theta}$ after having observed data $\mathcal{D}_{\text{train}}$.

- Prior believes about the parameter are expressed via a prior distribution $q(\boldsymbol{\theta})$, which is updated according to Bayes' rule

$$\underbrace{p(\boldsymbol{\theta} \mid \mathbf{X}, \mathbf{y})}_{\text{posterior}} = \frac{\overbrace{p(\mathbf{y} \mid \mathbf{X}, \boldsymbol{\theta})}^{\text{likelihood}} \overbrace{q(\boldsymbol{\theta})}^{\text{prior}}}{\underbrace{p(\mathbf{y} \mid \mathbf{X})}_{\text{marginal}}}.$$

Function-Space View I

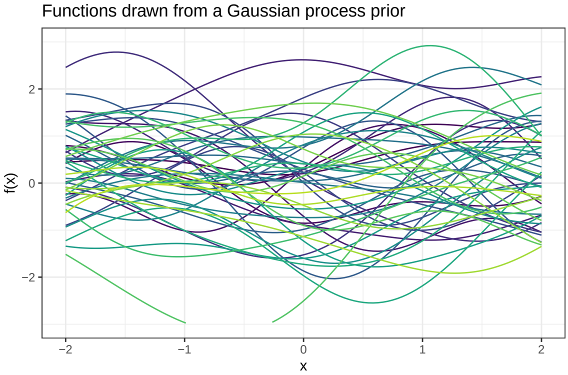
Let us change our point of view:

- Instead of “searching” for a parameter $\boldsymbol{\theta}$ in the parameter space, we directly search in a space of “allowed” functions $\mathcal{H}$.

- We will still use Bayesian inference, but instead of specifying a prior distribution over a parameter, we will specify a prior distribution **over functions** and will update it according to the data points that we observe.

Function-Space View II

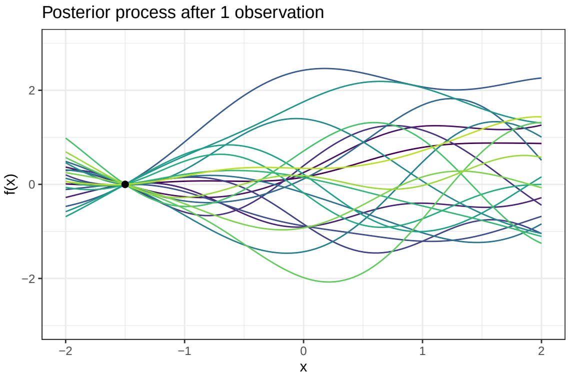
Intuitively, imagine we could draw a huge number of functions from some prior distribution over functions $^{(*)}$.



$^{(*)}$ We will see in a minute how distributions over functions can be specified.

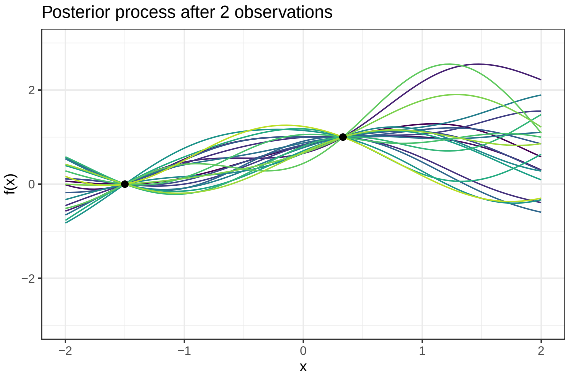
Function-Space View III

After observing some data points, we are allowed to sample only those functions that are consistent with the data.



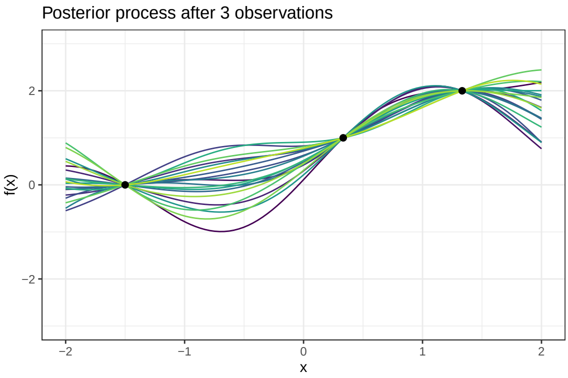
Function-Space View IV

After observing some data points, we are allowed to sample only those functions that are consistent with the data.



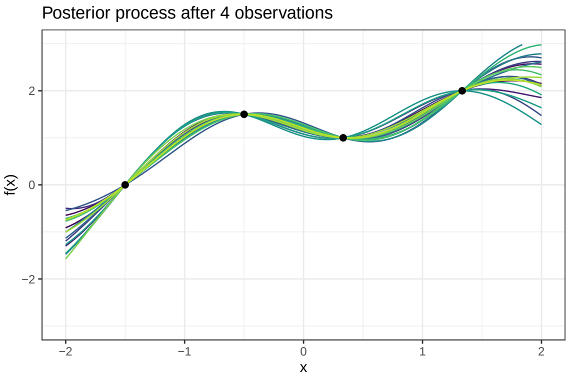
Function-Space View V

After observing some data points, we are allowed to sample only those functions that are consistent with the data.



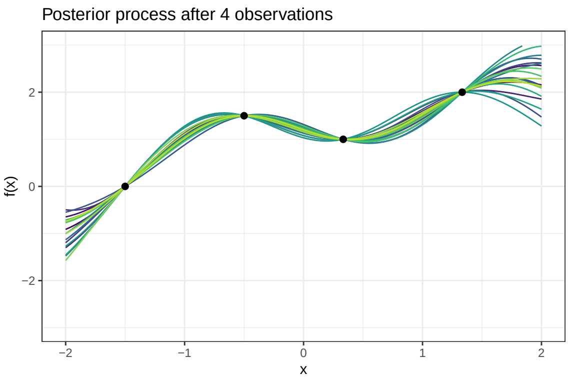
Function-Space View VI

As we observe more and more data points, the number of functions that consistent with the data shrinks.



Function-Space View VII

Intuitively, there is something like the “mean” and “variance” of a distribution over functions.



Weight-Space View vs. Function-Space View

Weight-Space View	**Function-Space View**
Parameterize functions Example: $f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}$	
Define distributions on $\boldsymbol{\theta}$	Define distributions on f
Inference in parameter space Θ	Inference in function space $\mathcal{H}$

Next, we will see how we can define distributions over functions mathematically.

Distributions on Functions

Discrete Functions I

For simplicity, we will firstly consider functions with finite domains.

- Let $\mathcal{X} = \left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$ be a finite set of elements and $\mathcal{H}$ the set of all functions $h : \mathcal{X} \to \mathbb{R}$.

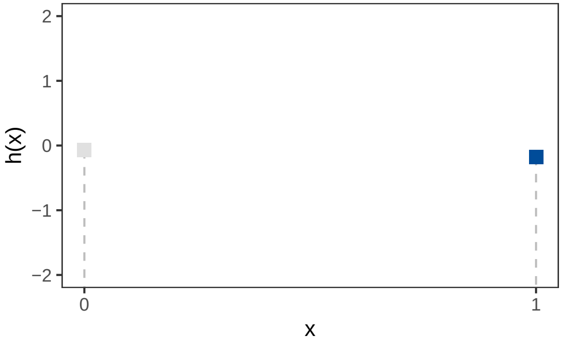
- Since the domain of any $h(\cdot) \in \mathcal{H}$ has only n elements, we can represent the function $h(\cdot)$ compactly as a n-dimensional vector

$$\boldsymbol{h} = \left[h\left(\mathbf{x}^{(1)}\right), \ldots, h\left(\mathbf{x}^{(n)}\right)\right].$$

Discrete Functions II

Example 1: Consider function $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **two** points $\mathcal{X} = \{0, 1\}$.

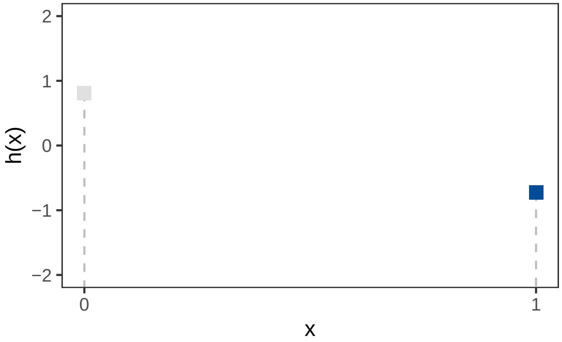
Examples for functions that live in this space:



Discrete Functions III

Example 1: Consider function $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **two** points $\mathcal{X} = \{0, 1\}$.

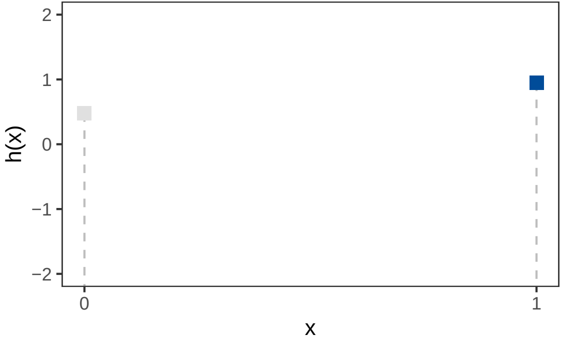
Examples for functions that live in this space:



Discrete Functions IV

Example 1: Consider function $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **two** points $\mathcal{X} = \{0, 1\}$.

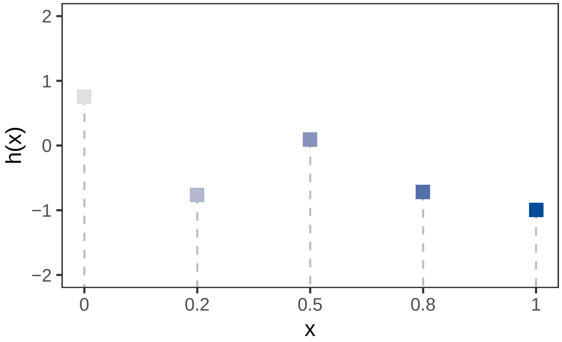
Examples for functions that live in this space:



Discrete Functions V

Example 2: Consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **five** points $\mathcal{X} = \{0, 0.25, 0.5, 0.75, 1\}$.

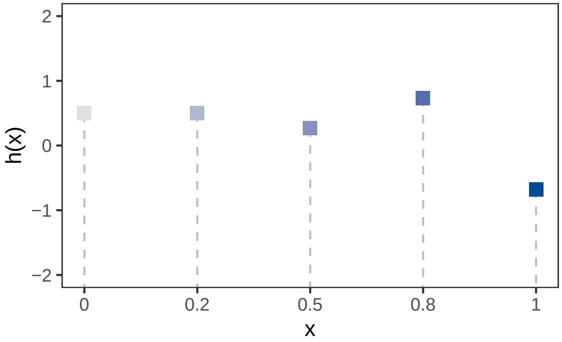
Examples for functions that live in this space:



Discrete Functions VI

Example 2: Consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **five** points $\mathcal{X} = \{0, 0.25, 0.5, 0.75, 1\}$.

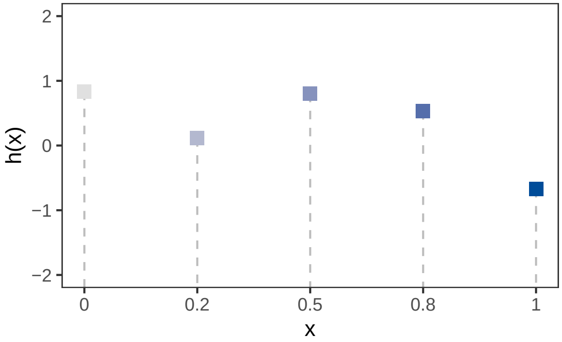
Examples for functions that live in this space:



Discrete Functions VII

Example 2: Consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **five** points $\mathcal{X} = \{0, 0.25, 0.5, 0.75, 1\}$.

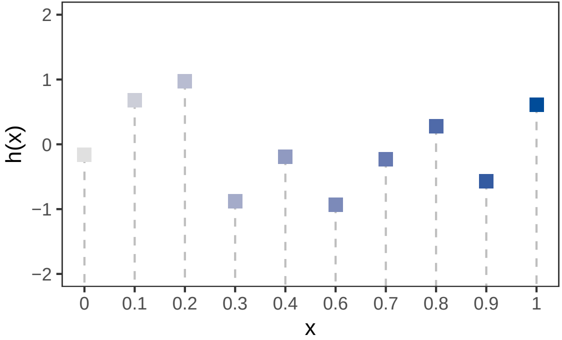
Examples for functions that live in this space:



Discrete Functions VIII

Example 3: Consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **ten** points.

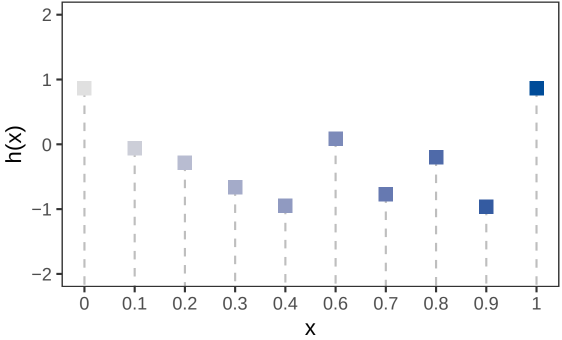
Examples for functions that live in this space:



Discrete Functions IX

Example 3: Consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **ten** points.

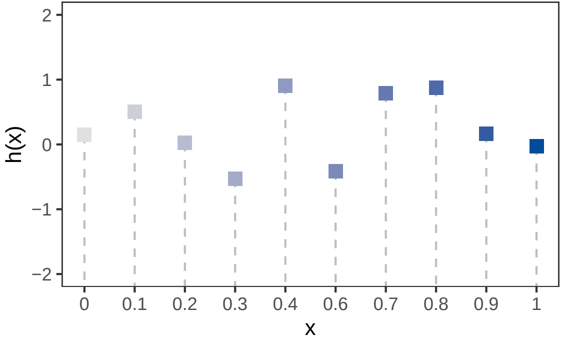
Examples for functions that live in this space:



Discrete Functions X

Example 3: Consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **ten** points.

Examples for functions that live in this space:



Distributions on Discrete Functions I

- One natural way to specify a probability distribution on a discrete function $h \in \mathcal{H}$ is to use the vector representation of the function:

$$\boldsymbol{h} = \left[h\left(\mathbf{x}^{(1)}\right), h\left(\mathbf{x}^{(2)}\right), \ldots, h\left(\mathbf{x}^{(n)}\right) \right].$$

- Let us consider $\boldsymbol{h}$ as a n-dimensional random variable. We will further assume the following normal distribution:

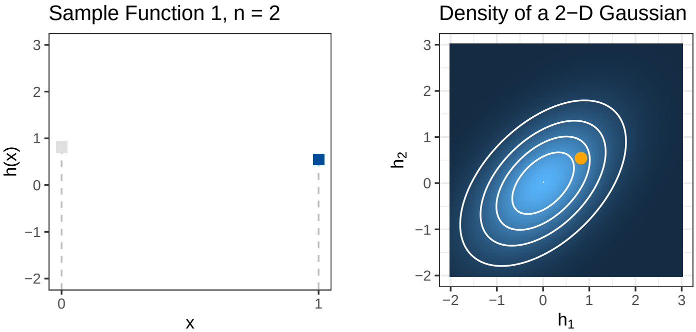
$$\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{K}\right).$$

Note: For now, we set $\boldsymbol{m} = \mathbf{0}$ and take the covariance matrix $\boldsymbol{K}$ as given. We will see later how they are chosen / estimated.

Distributions on Discrete Functions II

Example 1 (continued): Let $h : \mathcal{X} \to \mathcal{Y}$ be a function that is defined on **two** points $\mathcal{X}$. We sample functions by sampling from a two-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

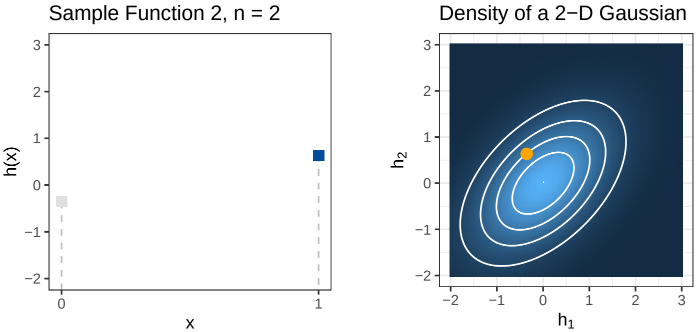


In this example, $m = (0, 0)$ and $K = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

Distributions on Discrete Functions III

Example 1 (continued): Let $h : \mathcal{X} \to \mathcal{Y}$ be a function that is defined on **two** points $\mathcal{X}$. We sample functions by sampling from a two-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

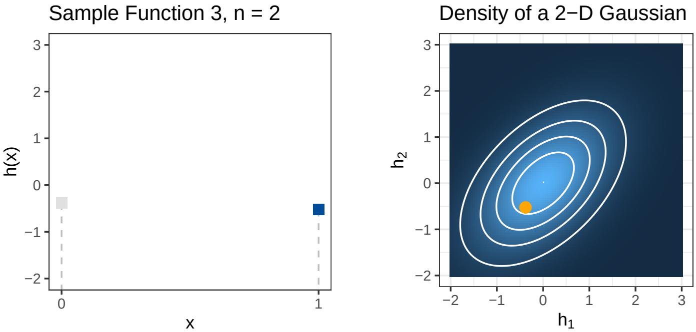


In this example, $m = (0, 0)$ and $K = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

Distributions on Discrete Functions IV

Example 1 (continued): Let $h : \mathcal{X} \to \mathcal{Y}$ be a function that is defined on **two** points $\mathcal{X}$. We sample functions by sampling from a two-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

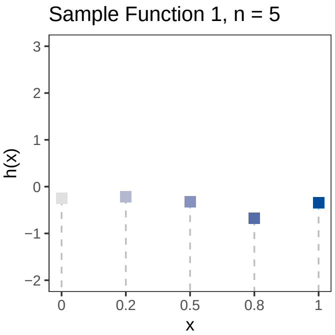


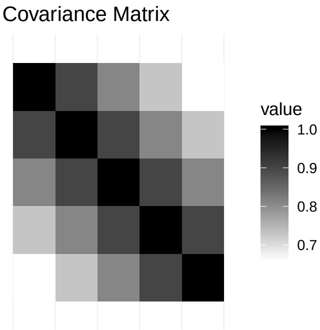
In this example, $m = (0, 0)$ and $K = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

Distributions on Discrete Functions V

Example 2 (continued): Let us consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **five** points. We sample functions by sampling from a five-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2), h(3), h(4), h(5)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

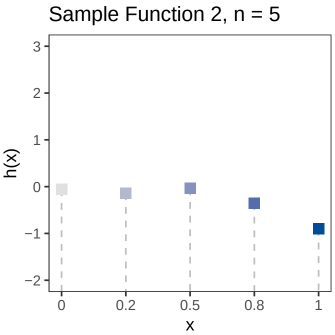


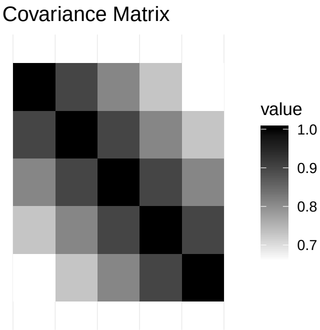


Distributions on Discrete Functions VI

Example 2 (continued): Let us consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **five** points. We sample functions by sampling from a five-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2), h(3), h(4), h(5)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

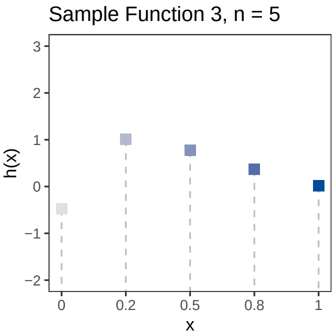


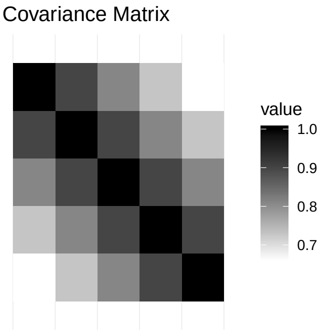


Distributions on Discrete Functions VII

Example 2 (continued): Let us consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **five** points. We sample functions by sampling from a five-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2), h(3), h(4), h(5)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

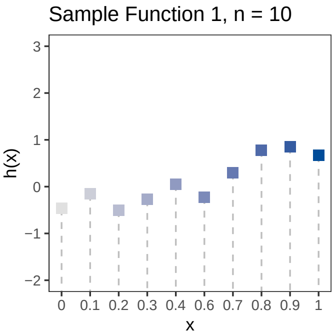


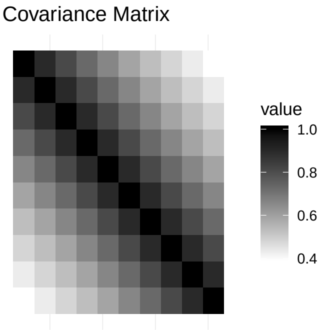


Distributions on Discrete Functions VIII

Example 3 (continued): Let us consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **ten** points. We sample functions by sampling from a ten-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2), \ldots, h(10)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

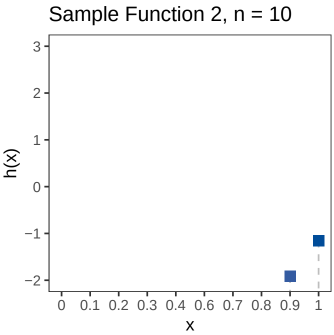


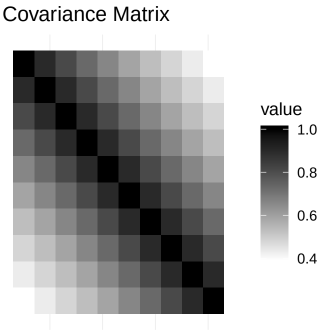


Distributions on Discrete Functions IX

Example 3 (continued): Let us consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **ten** points. We sample functions by sampling from a ten-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2), \ldots, h(10)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

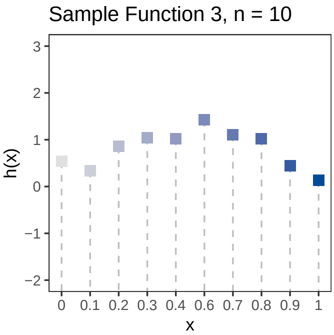


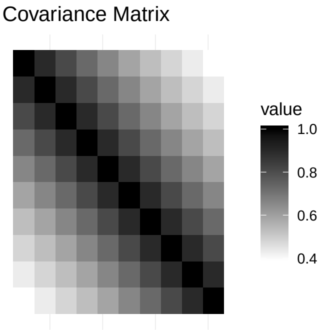


Distributions on Discrete Functions X

Example 3 (continued): Let us consider $h : \mathcal{X} \to \mathcal{Y}$ where the input space consists of **ten** points. We sample functions by sampling from a ten-dimensional normal variable

$$\boldsymbol{h} = [h(1), h(2), \dots, h(10)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$



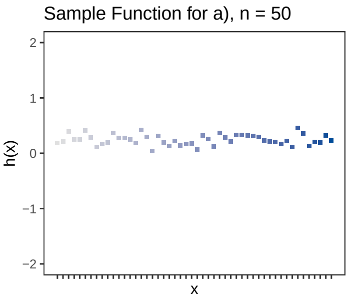


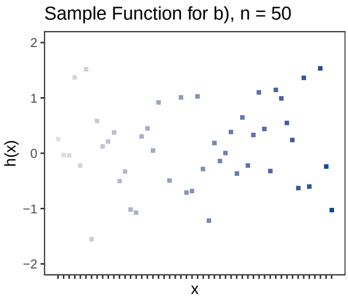
The Role of Covariance Function I

The covariance controls the "shape" of drawn functions. Consider two extreme cases where function values are:

a) strongly correlated: $\boldsymbol{K} = \begin{pmatrix} 1 & 0.99 & \dots & 0.99 \\ 0.99 & 1 & \dots & 0.99 \\ 0.99 & 0.99 & \ddots & 0.99 \\ 0.99 & \dots & 0.99 & 1 \end{pmatrix}$

b) uncorrelated: $\boldsymbol{K} = \boldsymbol{I}$.





The Role of Covariance Function II

- On a numeric space $\mathcal{X}$, "meaningful" functions may be characterized by the following spatial property:

  If $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ are close in the $\mathcal{X}$-space, their function values $f(\mathbf{x}^{(i)})$ and $f(\mathbf{x}^{(j)})$ should be close in $\mathcal{Y}$-space.

- In other words, if two data points are close in $\mathcal{X}$-space, their corresponding values should be **correlated**!

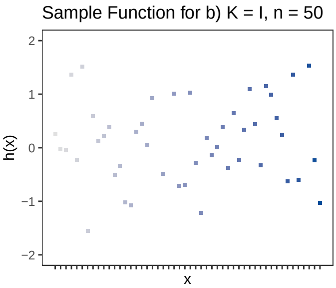
- We can enforce this condition by choosing a covariance function for which,

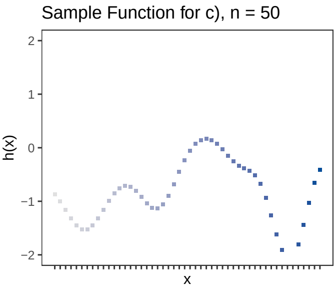
$$\boldsymbol{K}_{ij} \text{ is high, if } \mathbf{x}^{(i)} \text{ and } \mathbf{x}^{(j)} \text{ are close.}$$

The Role of Covariance Function III

We can compute the entries of the covariance matrix by a function that is based on the distance between $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$. For example:

c) spatial correlation: $K_{ij} = k(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}) = \exp\left(-\frac{1}{2}\left|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\right|^2\right)$





Note: $k(\cdot, \cdot)$ is known as the **covariance function** or **kernel**. It will be studied in more detail later on.

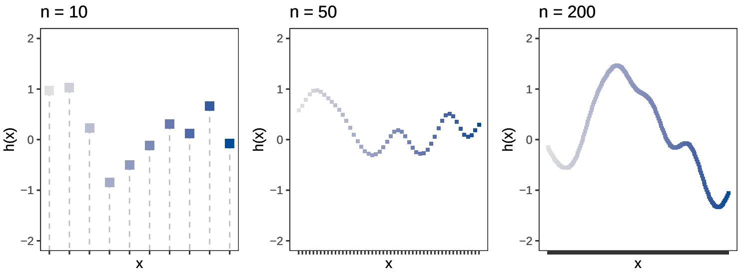
Gaussian Processes

From Discrete to Continuous Functions

- We have already considered distributions on functions with discrete domain. We did so, by defining Gaussian distributions on the vector of the respective function values

$$\mathbf{h} = [h(\mathbf{x}^{(1)}), h(\mathbf{x}^{(2)}), \dots, h(\mathbf{x}^{(n)})] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K}).$$

- We can generalize this idea for $n \to \infty$.



Gaussian Processes: Intuition I

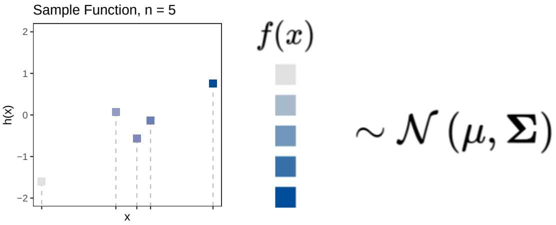
- No matter how large n is, we consider functions with discrete domains.
- But, how can we extend our definition to functions with **continuous** domains $\mathcal{X} \subset \mathbb{R}$?
- Intuitively, a function f drawn from a **Gaussian process** can be understood as an "infinite" long Gaussian random vector.
- It is unclear how to handle an "infinite" long Gaussian random vector!

Gaussian Processes: Intuition II

- Thus, it is required that for **any finite set** of inputs $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$, the vector $\mathbf{f}$ has a Gaussian distribution with $\boldsymbol{m}$ and $\boldsymbol{K}$ being calculated by a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$:

$$\boldsymbol{f} = \left[f\left(\mathbf{x}^{(1)}\right), \dots, f\left(\mathbf{x}^{(n)}\right)\right] \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{K}\right).$$

- This property is called the **Marginalization Property**.

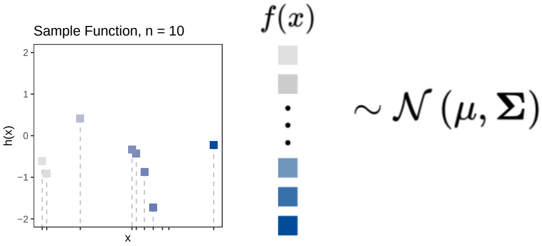


Gaussian Processes: Intuition III

- Thus, it is required that for **any finite set** of inputs $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$, the vector $\mathbf{f}$ has a Gaussian distribution with $\boldsymbol{m}$ and $\boldsymbol{K}$ being calculated by a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$:

$$\boldsymbol{f} = \left[f\left(\mathbf{x}^{(1)}\right), \dots, f\left(\mathbf{x}^{(n)}\right)\right] \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{K}\right).$$

- This property is called the **Marginalization Property**.

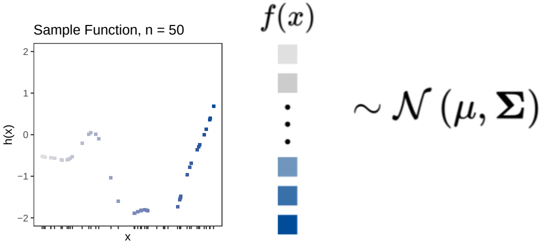


Gaussian Processes: Intuition IV

- Thus, it is required that for **any finite set** of inputs $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$, the vector $\mathbf{f}$ has a Gaussian distribution with $\boldsymbol{m}$ and $\boldsymbol{K}$ being calculated by a mean function $m(\cdot)$ and a covariance function $k(\cdot, \cdot)$:

$$\boldsymbol{f} = \left[f\left(\mathbf{x}^{(1)}\right), \ldots, f\left(\mathbf{x}^{(n)}\right)\right] \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{K}\right).$$

- This property is called the **Marginalization Property**.



Gaussian Processes: Formal Definitions I

- The above intuitive explanation is formally defined as follows.

A function $f(\mathbf{x})$ *is generated by a Gaussian process* $\mathcal{G}$ *if for* **any finite** *set of inputs* $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$*, the associated vector of function values has a Gaussian distribution:*

$$\boldsymbol{f} = \left(f(\mathbf{x}^{(1)}), \ldots, f(\mathbf{x}^{(n)})\right) \sim \mathcal{N}\left(\mathbf{m}, \mathbf{K}\right),$$

with

$$\mathbf{m} := \left(m\left(\mathbf{x}^{(i)}\right)\right)_i, \quad \mathbf{K} := \left(k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)_{i,j},$$

where $m(\mathbf{x})$ *is called mean function and* $k(\mathbf{x}, \mathbf{x}')$ *is called covariance function.*

Gaussian Processes: Formal Definitions II

- A GP is **completely specified** by its mean and covariance functions.
- The mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$ of a real process $f(\mathbf{x})$ are defined as:

$$\begin{aligned} m(\mathbf{x}) &= \mathbb{E}[f(\mathbf{x})] \\ k(\mathbf{x}, \mathbf{x}') &= \mathbb{E}\Big[(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]) \left(f(\mathbf{x}') - \mathbb{E}[f(\mathbf{x}')]\right)\Big] \end{aligned}$$

- We denote a GP by

$$f(\mathbf{x}) \sim \mathcal{G}\left(m(\mathbf{x}), k\left(\mathbf{x}, \mathbf{x}'\right)\right)$$

Note: For now, we assume $m(\mathbf{x}) \equiv 0$. This is not a drastic limitation. In fact, it is common to consider GPs with a zero mean function.

Sampling from a Gaussian Process Prior I

- We can draw functions from a Gaussian process prior. To do so, consider $f(\mathbf{x}) \sim \mathcal{G}\left(0, k(\mathbf{x}, \mathbf{x}')\right)$ with the squared exponential covariance function $^{(*)}$

$$k(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{1}{2\ell^2}\|\mathbf{x} - \mathbf{x}'\|^2\right), \quad \ell = 1.$$

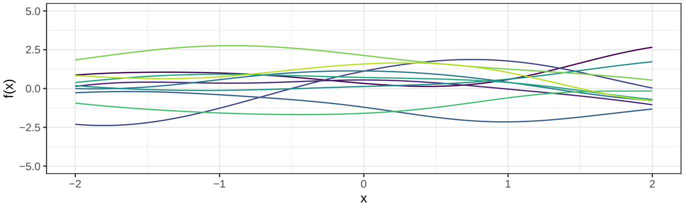
- This covariance function specifies the Gaussian process completely.

$^{(*)}$ We will talk later about different choices of covariance functions.

Sampling from a Gaussian Process Prior II

To visualize a sample function, we

- choose a large number of equidistant points: $\left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right\}$,
- compute their corresponding covariance matrix by plugging in all pairs of $\mathbf{x}^{(i)}$ and $\mathbf{x}^{(j)}$ in $\mathbf{K} = \left(k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)_{i,j}$,
- sample from a Gaussian $\boldsymbol{f} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{K})$.



We draw 10 times from the Gaussian, to get 10 different samples. Since we specified the mean function to be zero, the drawn functions have a zero mean.

Gaussian Processes as an Indexed Family

Gaussian Processes as an Indexed Family

- A Gaussian process is a special case of a **stochastic process** which is defined as a collection of random variables indexed by some index set (also called an **indexed family**).

- What does it mean?

- An **indexed family** is a mathematical function (or "rule") that maps indices $t \in T$ to objects in $\mathcal{S}$.

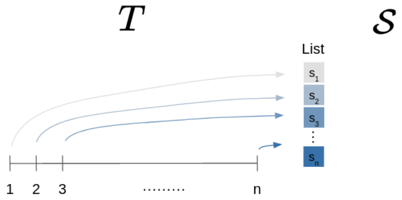
Definition: *an* **index family** *(or a family of elements in $\mathcal{S}$ indexed by T) is a surjective function that is defined as follows:*

$$
\begin{aligned}
s : T &\rightarrow \mathcal{S} \\
t &\mapsto s_t = s(t)
\end{aligned}
$$

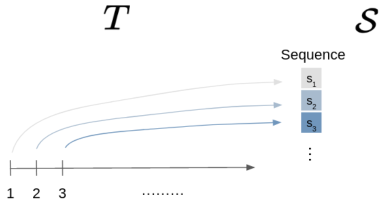
Index Family I

Some simple examples for indexed families are:

- Finite sequences (lists): $T = \{1, 2, \dots, n\}$ and $(s_t)_{t \in T} \in \mathbb{R}$



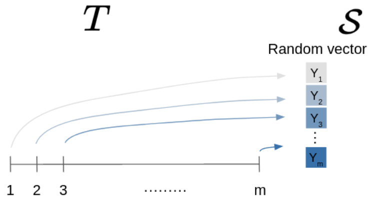
- Infinite sequences: $T = \mathbb{N}$ and $(s_t)_{t \in T} \in \mathbb{R}$

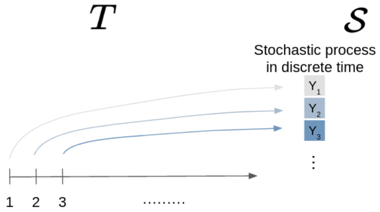


Index Family II

But the indexed set $\mathcal{S}$ can be something more complicated, for example functions or **random variables** (RV):

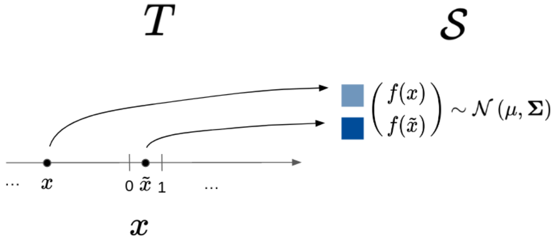
- $T = \{1, \ldots, m\}$, Y_t's are RVs: Indexed family is a random vector.
- $T = \{1, \ldots, m\}$, Y_t's are RVs: Indexed family is a stochastic process in discrete time
- $T = \mathbb{Z}^2$, Y_t's are RVs: Indexed family is a 2D-random walk.





Index Family III

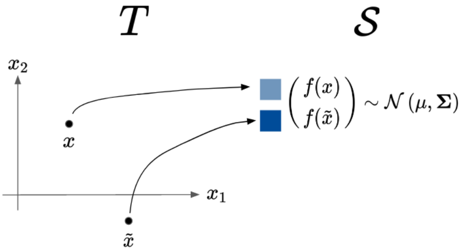
- A Gaussian process is also an indexed family, where the random variables $f(\mathbf{x})$ are indexed by the input values $\mathbf{x} \in \mathcal{X}$.
- Importantly, any indexed (finite) random vector has a multivariate Gaussian distribution (which comes with all the nice properties of Gaussianity!).



Visualization for a one-dimensional $\mathcal{X}$.

Index Family IV

- A Gaussian process is also an indexed family, where the random variables $f(\mathbf{x})$ are indexed by the input values $\mathbf{x} \in \mathcal{X}$.
- Importantly, any indexed (finite) random vector has a multivariate Gaussian distribution (which comes with all the nice properties of Gaussianity!).



Visualization for a two-dimensional $\mathcal{X}$.