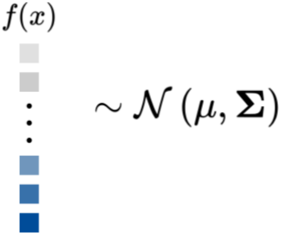
Introduction to Machine Learning

Gaussian Processes



Learning goals

- XXX
- XXX

WEIGHT-SPACE VIEW

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

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

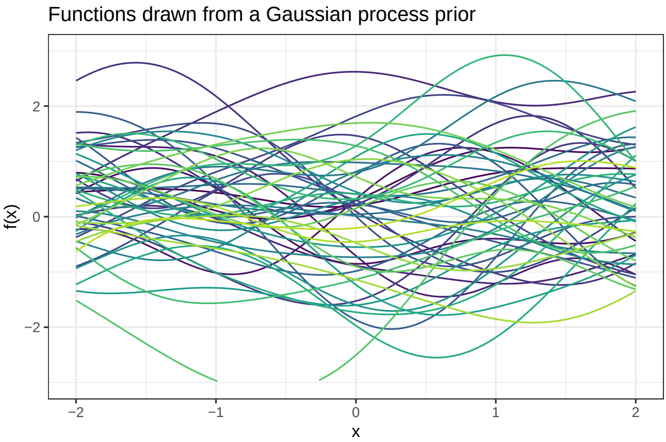
FUNCTION-SPACE VIEW

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 still use Bayesian inference, but instead specifying a prior distribution over a parameter, we specify a prior distribution **over functions** and update it according to the data points we have observed.

FUNCTION-SPACE VIEW

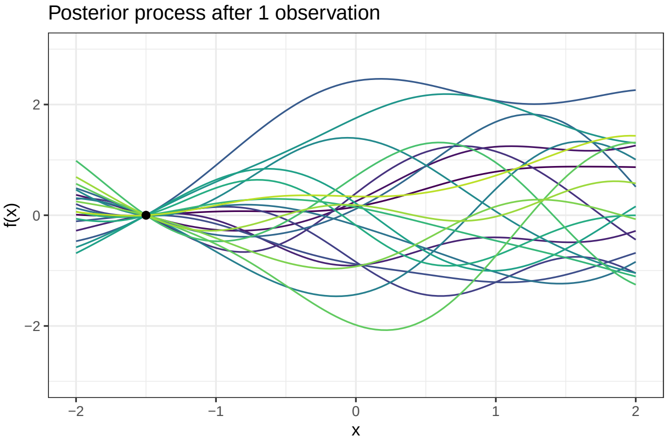
Intuitively, imagine we could draw a huge number of functions from some prior distribution over functions $^{(*)}$.



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

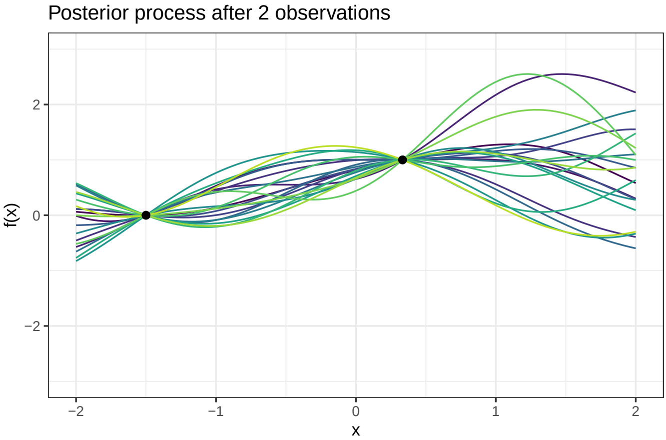
FUNCTION-SPACE VIEW

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



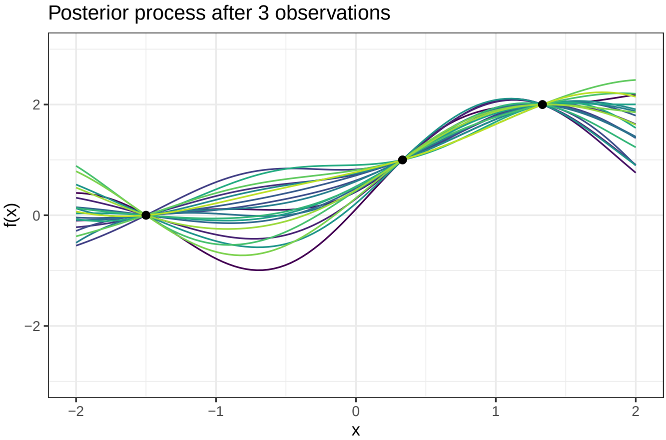
FUNCTION-SPACE VIEW

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



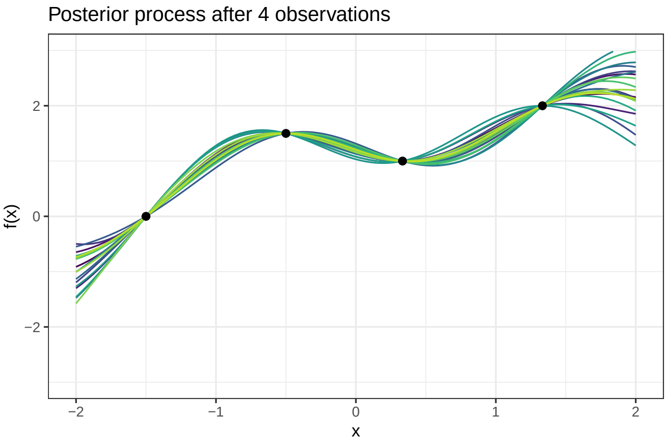
FUNCTION-SPACE VIEW

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



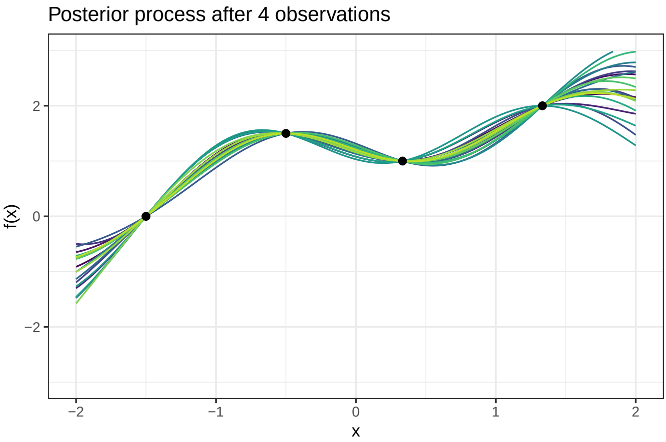
FUNCTION-SPACE VIEW

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



FUNCTION-SPACE VIEW

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



WEIGHT-SPACE 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

For simplicity, let us consider functions with finite domains first.

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 from $\mathcal{X} \to \mathbb{R}$.

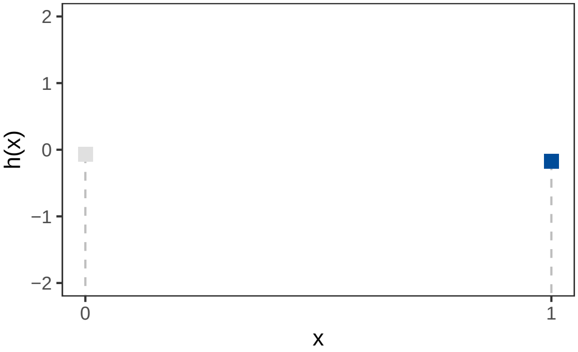
Since the domain of any $h(.) \in \mathcal{H}$ has only n elements, we can represent the function $h(.)$ compactly as a n-dimensional vector

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

DISCRETE FUNCTIONS

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

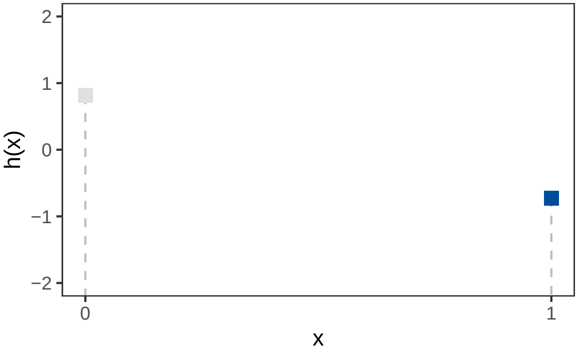
Examples for functions that live in this space:



DISCRETE FUNCTIONS

Example 1: Let us consider $h : \mathcal{X} \rightarrow \mathcal{Y}$ where the input space consists of **two** points $\mathcal{X} = \{0, 1\}$.

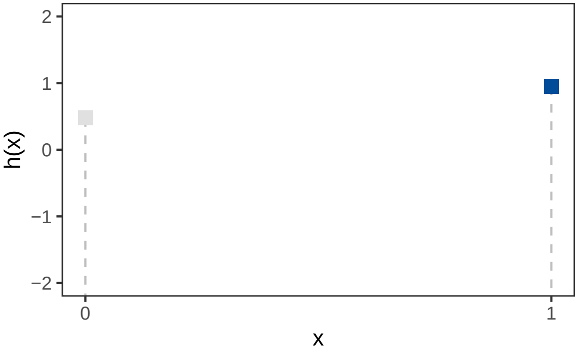
Examples for functions that live in this space:



DISCRETE FUNCTIONS

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

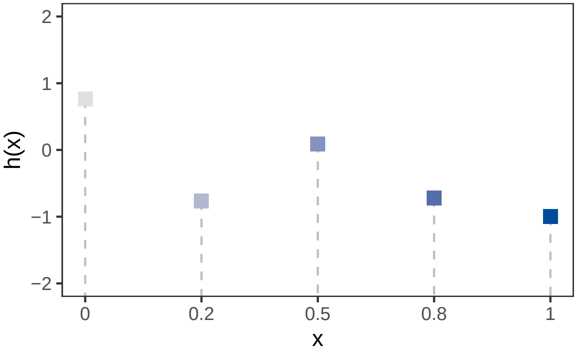
Examples for functions that live in this space:



DISCRETE FUNCTIONS

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

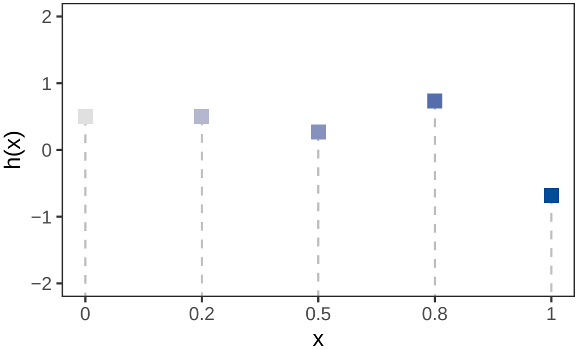
Examples for functions that live in this space:



DISCRETE FUNCTIONS

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

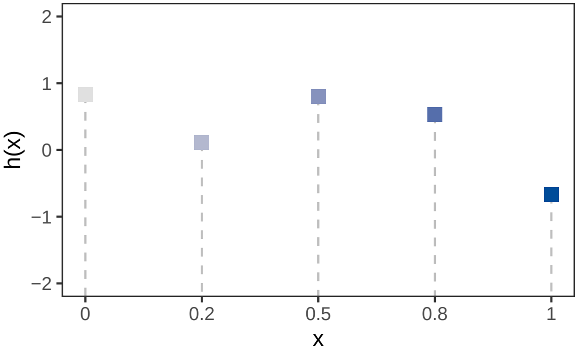
Examples for functions that live in this space:



DISCRETE FUNCTIONS

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

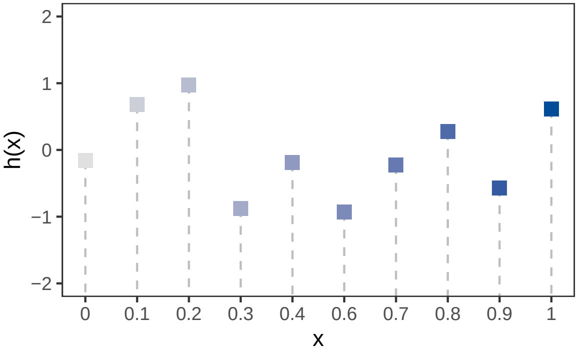
Examples for functions that live in this space:



DISCRETE FUNCTIONS

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

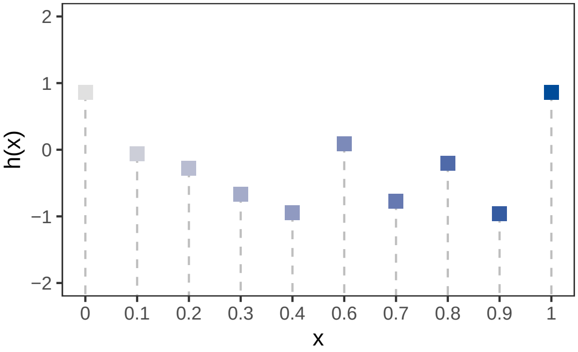
Examples for functions that live in this space:



DISCRETE FUNCTIONS

Example 3: Let us consider $h : \mathcal{X} \rightarrow \mathcal{Y}$ where the input space consists of **ten** points.

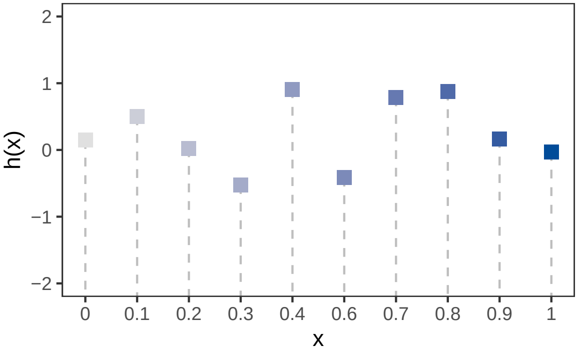
Examples for functions that live in this space:



DISCRETE FUNCTIONS

Example 3: Let us 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

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

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

of the function.

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

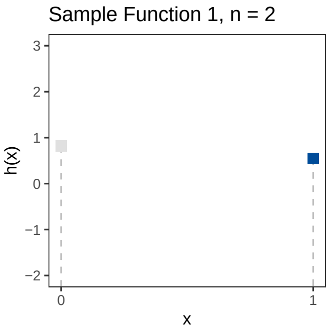
$$\boldsymbol{h} \sim \mathcal{N}\left(\boldsymbol{m}, \boldsymbol{K}\right).$$

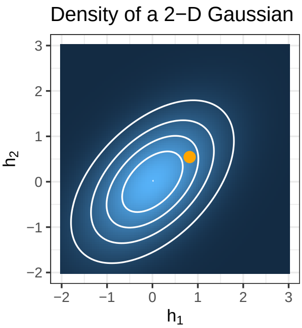
Note: For now, we set $\boldsymbol{m} = \mathbf{0}$ and take the covariance matrix $\boldsymbol{K}$ as given. We will see later how they are chosen / estimated.

DISCRETE FUNCTIONS

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

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



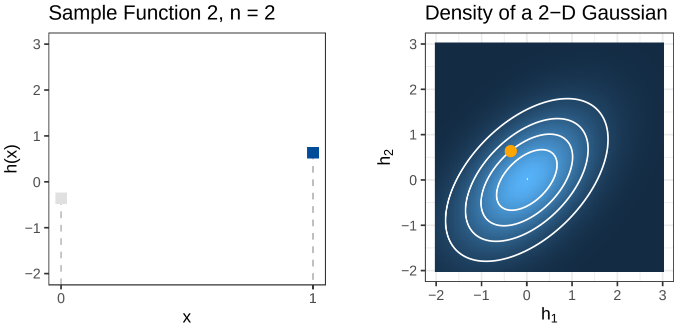


In this example, $m = (0, 0)$ and $K = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

DISCRETE FUNCTIONS

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

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

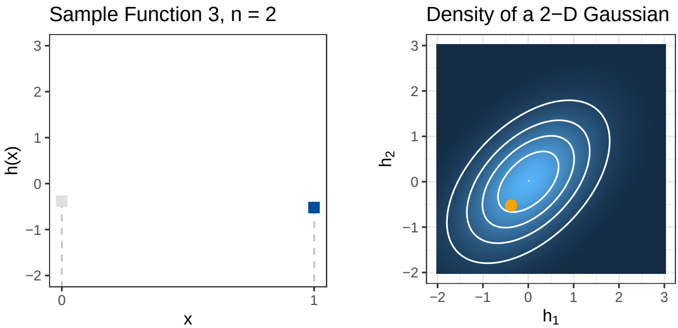


In this example, $m = (0, 0)$ and $K = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

DISCRETE FUNCTIONS

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

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

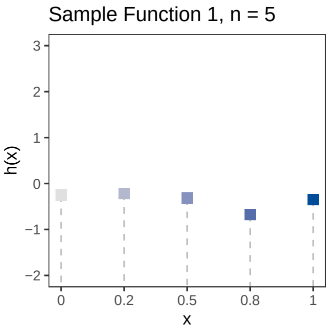


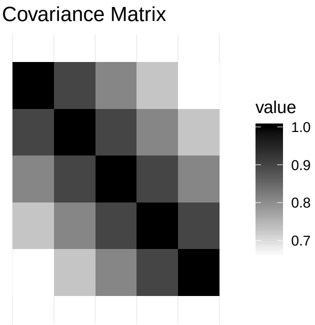
In this example, $m = (0, 0)$ and $K = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$.

DISCRETE FUNCTIONS

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

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

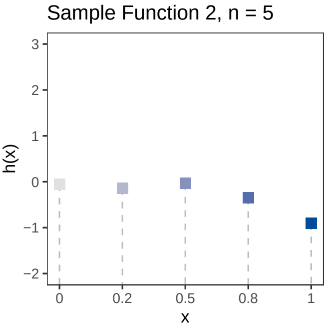


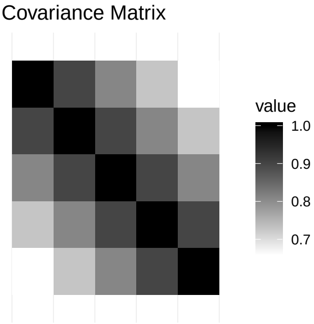


DISCRETE FUNCTIONS

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

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

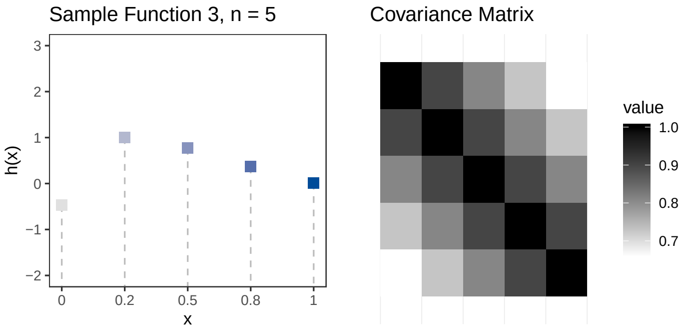




DISCRETE FUNCTIONS

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

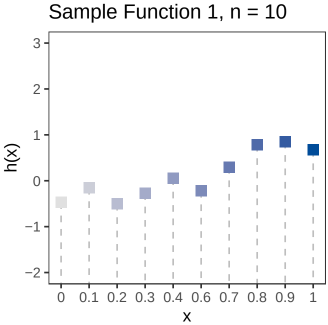
$$\boldsymbol{h} = [h(1), h(2), h(3), h(4), h(5)] \sim \mathcal{N}(\boldsymbol{m}, \boldsymbol{K})$$

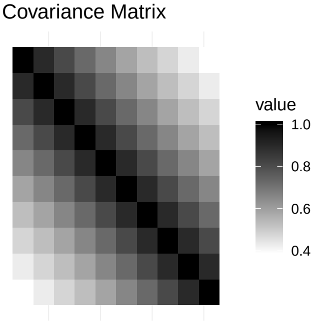


DISCRETE FUNCTIONS

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 ten-dimensional normal variable

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

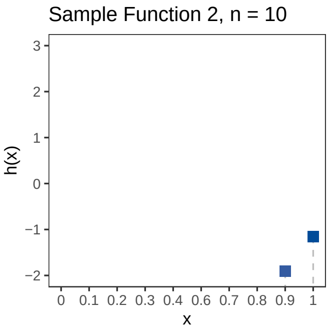


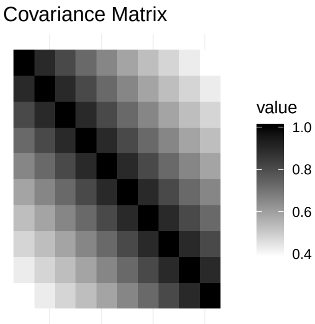


DISCRETE FUNCTIONS

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 ten-dimensional normal variable

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

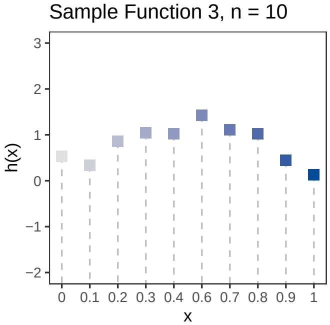


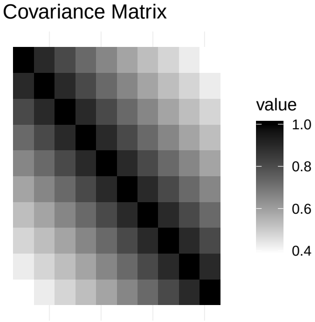


DISCRETE FUNCTIONS

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 ten-dimensional normal variable

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



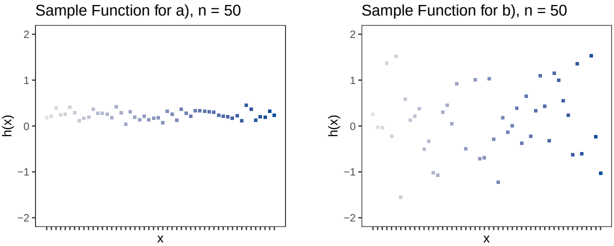


ROLE OF THE COVARIANCE FUNCTION

Note that the covariance controls the “shape” of the drawn function. 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: $\boldsymbol{K} = \boldsymbol{I}$



ROLE OF THE COVARIANCE FUNCTION

- "Meaningful" functions (on a numeric space $\mathcal{X}$) may be characterized by a spatial property:

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

  In other words: If they are close in $\mathcal{X}$-space, their functions values should be **correlated**!

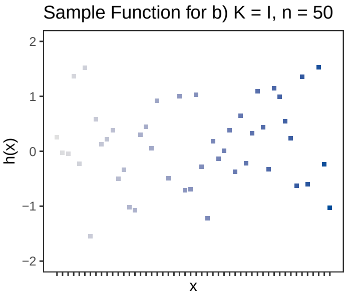
- We can enforce that by choosing a covariance function with

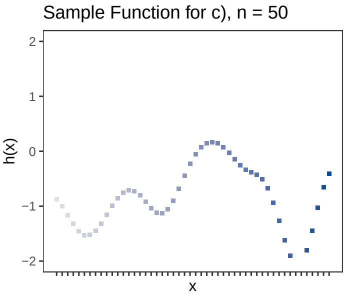
$$\boldsymbol{K}_{ij} \text{ high, if } \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \text{ close.}$$

ROLE OF THE COVARIANCE FUNCTION

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

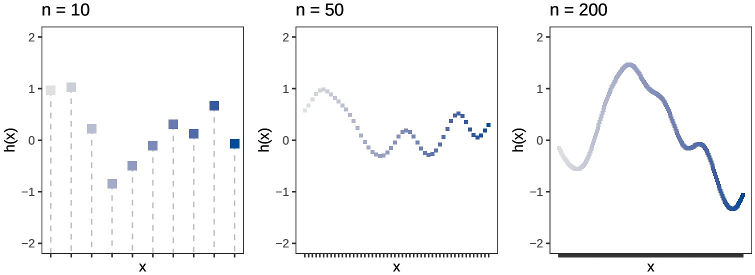
Gaussian Processes

FROM DISCRETE TO CONTINUOUS FUNCTIONS

- We defined distributions on functions with discrete domain by defining a Gaussian 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 do this for $n \to \infty$ (as “granular” as we want)



FROM DISCRETE TO CONTINUOUS FUNCTIONS

- No matter how large *n* is, we are still considering a function over a discrete domain.
- How can we extend our definition to functions with **continuous domain** $\mathcal{X} \subset \mathbb{R}$?

GAUSSIAN PROCESSES: INTUITION

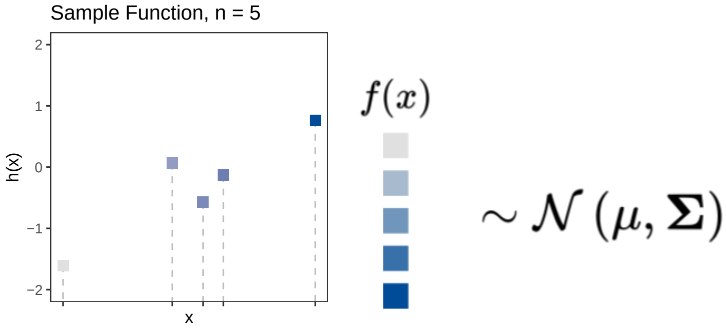
- Intuitively, a function *f* drawn from **Gaussian process** can be understood as an “infinite” long Gaussian random vector.
- It is unclear how to handle an “infinite” long Gaussian random vector!

GAUSSIAN PROCESSES: INTUITION

- Thus, it is required that for **any finite set** of inputs $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$, the vector **f** has a Gaussian distribution

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

  with $\boldsymbol{m}$ and $\boldsymbol{K}$ being calculated by a mean function $m(.)$ / covariance function $k(.,.)$.
- This property is called **Marginalization Property**.

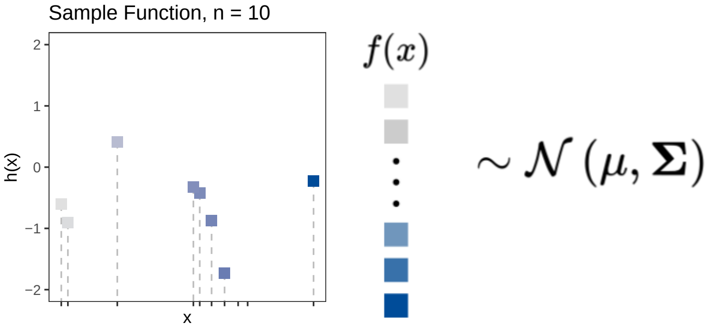


GAUSSIAN PROCESSES: INTUITION

- 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

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

  with $\boldsymbol{m}$ and $\boldsymbol{K}$ being calculated by a mean function $m(.)$ / covariance function $k(.,.)$.
- This property is called **Marginalization Property**.

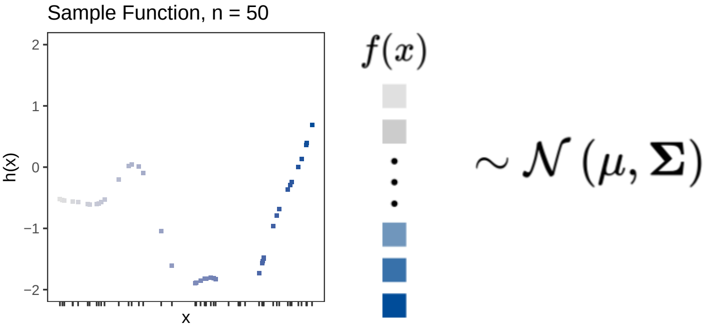


GAUSSIAN PROCESSES: INTUITION

- Thus, it is required that for **any finite set** of inputs $\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\} \subset \mathcal{X}$, the vector **f** has a Gaussian distribution

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

  with $\boldsymbol{m}$ and $\boldsymbol{K}$ being calculated by a mean function $m(.)$ / covariance function $k(.,.)$.
- This property is called **Marginalization Property**.



GAUSSIAN PROCESSES

This intuitive explanation is formally defined as follows:

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

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

with

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

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

GAUSSIAN PROCESSES

A GP is thus **completely specified** by its mean and covariance function

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

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

SAMPLING FROM A GAUSSIAN PROCESS PRIOR

We can draw functions from a Gaussian process prior. Let us consider $f(\mathbf{x}) \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$ 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 specifies the Gaussian process completely.

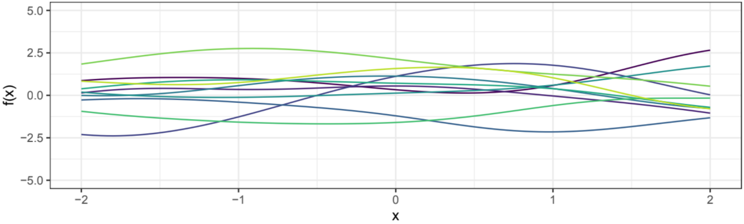
$^{(*)}$ We will talk later about different choices of covariance functions.

SAMPLING FROM A GAUSSIAN PROCESS PRIOR

To visualize a sample function, we

- choose a high number n (equidistant) points $\left\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(n)}\right\}$
- compute the corresponding covariance matrix $\mathbf{K} = \left(k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)_{i,j}$ by plugging in all pairs $\mathbf{x}^{(i)}, \mathbf{x}^{(j)}$
- 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 $m(\mathbf{x}) \equiv 0$, the drawn functions have zero mean.

Gaussian Processes as 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”) to map indices $t \in T$ to objects in $\mathcal{S}$.

Definition

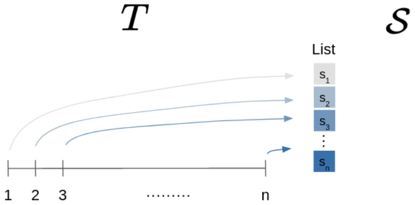
A **family of elements in $\mathcal{S}$ indexed by T** (indexed family) is a surjective function

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

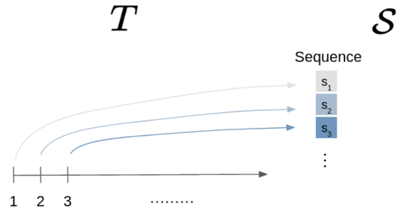
INDEXED FAMILY

Some simple examples for indexed families are:

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



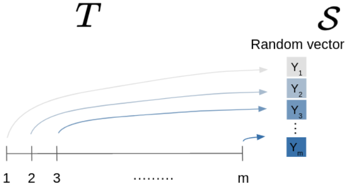
- infinite sequences:
  $T = \mathbb{N}$ and $(s_t)_{t \in T} \in \mathbb{R}$



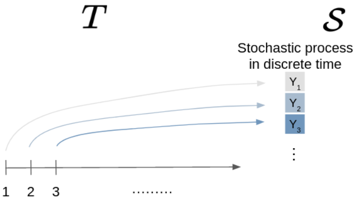
INDEXED FAMILY

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.



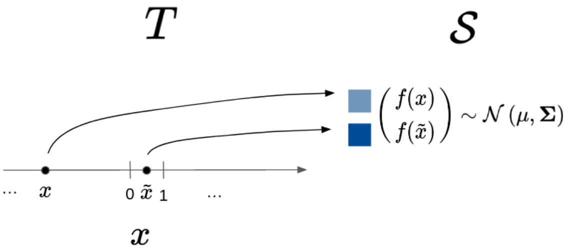
- $T = \{1, \ldots, m\}$, Y_t's are RVs: Indexed family is a stochastic process in discrete time



- $T = \mathbb{Z}^2$, Y_t's are RVs: Indexed family is a 2D-random walk.

INDEXED FAMILY

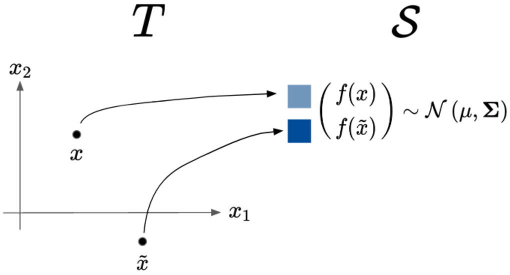
- A Gaussian process is also an indexed family, where the random variables $f(\mathbf{x})$ are indexed by the input values $\mathbf{x} \in \mathcal{X}$.
- Their special feature: Any indexed (finite) random vector has a multivariate Gaussian distribution (which comes with all the nice properties of Gaussianity!).



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

INDEXED FAMILY

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



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