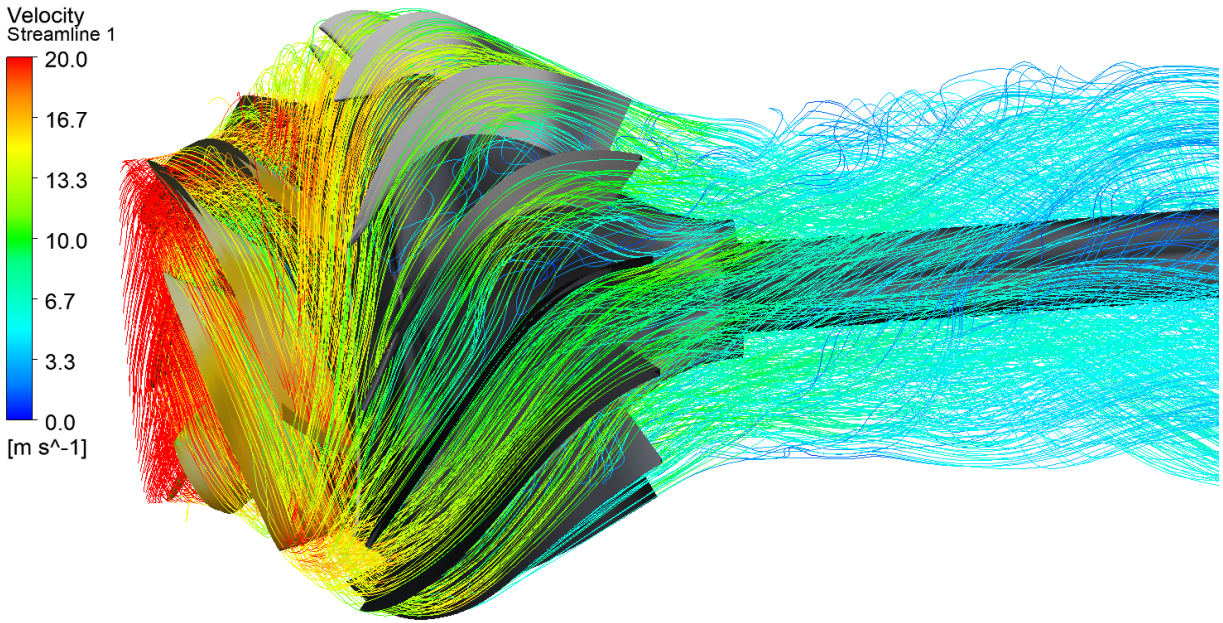
Data Driven Approaches in (Multi-objective) Bayesian Optimisation - An Introductory Course

Tinkle Chugh

Department of Computer Science
University of Exeter, UK

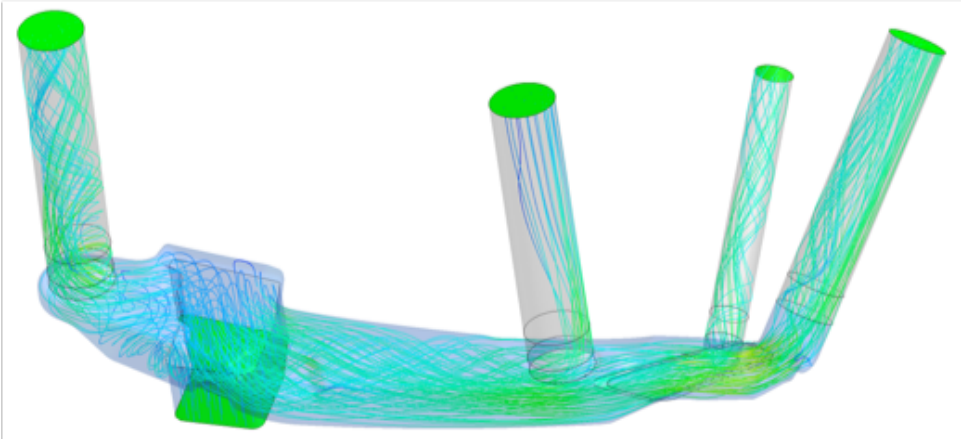
Expensive Optimisation



Pump Optimisation

1. One evaluation: 24 to 48 hours
2. Utilises CFD simulations

Expensive Optimisation



Air Intake Ventilation System

1. One evaluation: 30 minutes
2. Utilises CFD simulations

Bayesian Optimisation

Problem

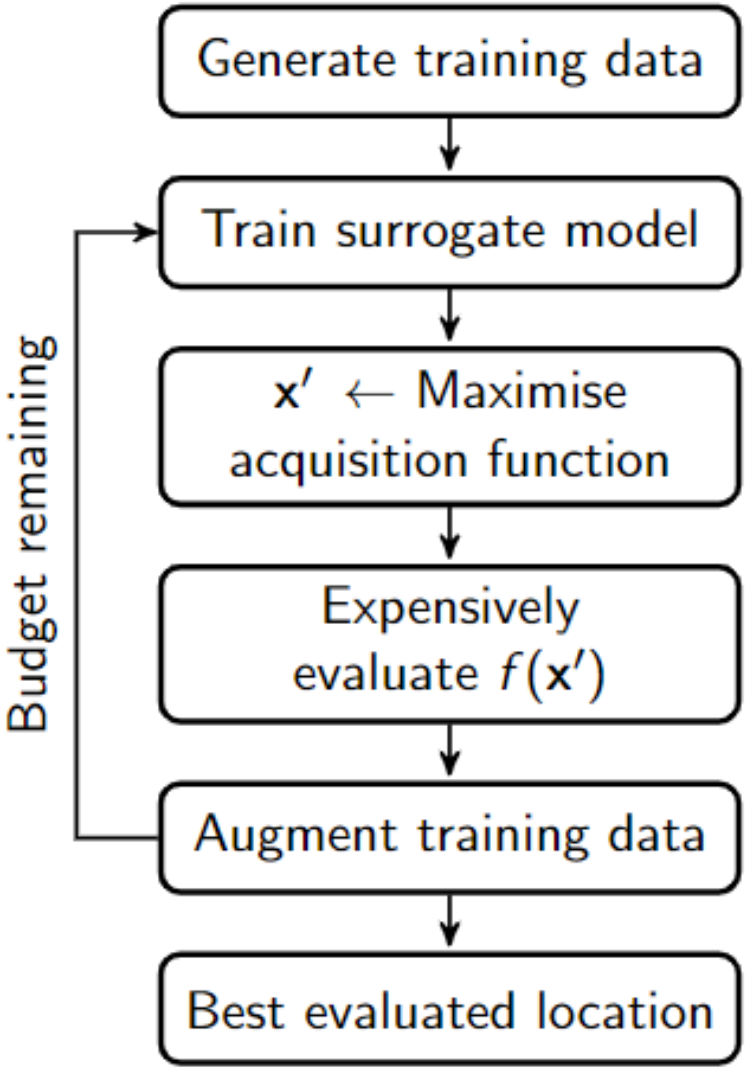
$$\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \qquad \mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$$

Bayesian optimisation

- Build a probabilistic surrogate model of $f(\mathbf{x})$
- Maximise an *acquisition* function $\alpha(\mathbf{x})$

$$\mathbf{x}' \leftarrow \operatorname*{argmax}_{\mathbf{x}\in\mathcal{X}} \alpha(\mathbf{x})$$

  Balances exploitation with exploration
- Expensively evaluate $f(\mathbf{x}')$



Bayesian Optimisation

Two important elements

1. Bayesian model: $p(f|D)$
2. Acquisition function : $\alpha(x, p(f|D))$

- Focus of this lecture: Bayesian Model

Gaussian Processes

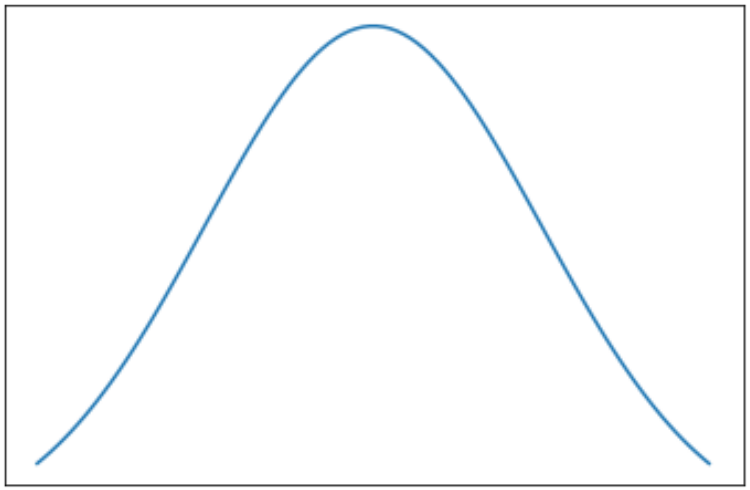
- Time series: Wiener, Kolmogorov 1940's
- Geostatistics: kriging 1970's
- Spatial statistics in general: see Cressie [1993] for overview
- General regression: O'Hagan [1978]
- Computer experiments (noise free): Sacks et al. [1989] and a survey on Bayesian optimization: Shahriari et al. [2016]
- Machine learning: Williams and Rasmussen [1996], Neal [1996]

Reading material: Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning, MIT Press, 2006
`http://www.gaussianprocess.org/gpml/chapters/`

Gaussian Processes

- Given a data set $D = \{(\mathbf{x}_n, y_n)\}_{n=1}^{N}$, we want to learn a function f, which not only provides the mean prediction (or point prediction) but also the uncertainty around them
- The prediction can be for regression or classification

Bayesian modelling



Bayesian Modelling

1. Univariate normal (or Gaussian) distribution has two parameters: mean (μ) and standard deviation (σ)

Bayesian Modelling

1. Univariate normal (or Gaussian) distribution has two parameters: mean (μ) and standard deviation (σ)
2. Let us denote the parameters with $\boldsymbol{\theta} = (\mu, \sigma)$
3. Given a data $\mathbf{y} = (y_1, \ldots, y_n)$, the distribution is denoted as:

$$p(y_i|\boldsymbol{\theta}) \sim \mathcal{N}(\mu, \sigma^2) \text{ Or}$$

$$p(y_i|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-0.5 \frac{(y_i - \mu)^2}{\sigma^2} \right)$$

A statistical model can specify a probability distribution for data $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ as $p(\mathbf{y}|\boldsymbol{\theta})$

Likelihood

- $p(\mathbf{y}|\boldsymbol{\theta})$ with varying values of model parameters $\boldsymbol{\theta}$

Likelihood

- $p(\mathbf{y}|\boldsymbol{\theta})$ with varying values of model parameters $\boldsymbol{\theta}$
- In other words, how likely is it to get $\mathbf{y}$ for different values of $\boldsymbol{\theta}$

Likelihood

- $p(\mathbf{y}|\boldsymbol{\theta})$ with varying values of model parameters $\boldsymbol{\theta}$
- In other words, how likely is it to get $\mathbf{y}$ for different values of $\boldsymbol{\theta}$
- For n independent and identically distributed (IID) samples, $(y_1, y_2, \ldots, y_n)$, the likelihood is:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

Likelihood

- $p(\mathbf{y}|\boldsymbol{\theta})$ with varying values of model parameters $\boldsymbol{\theta}$
- In other words, how likely is it to get $\mathbf{y}$ for different values of $\boldsymbol{\theta}$
- For n independent and identically distributed (IID) samples, $(y_1, y_2, \ldots, y_n)$, the likelihood is:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

$$p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

- Likelihood depends on the model - two different models for the same data $\mathbf{y}$ will typically have two different likelihoods

Maximising likelihood

- Find the model parameters $\boldsymbol{\theta}$ by looking at the data $\mathbf{y}$ Or
- Maximise the likelihood to find the model parameters i.e.

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\text{argmax}} \; L(\boldsymbol{\theta}) \;\; \text{Or}$$

$$\hat{\boldsymbol{\theta}} = \underset{\theta}{\text{argmax}} \; \prod_{i=1}^{n} p(y_i|\boldsymbol{\theta})$$

Maximising Likelihood

- argmax of $L(\theta)$ is the same as the argmax of $\log L(\boldsymbol{\theta})$ - makes the maths simpler:

$$\log L(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} p(y_i|\theta) = \sum_{i=1}^{n} \log p(y_i|\boldsymbol{\theta})$$

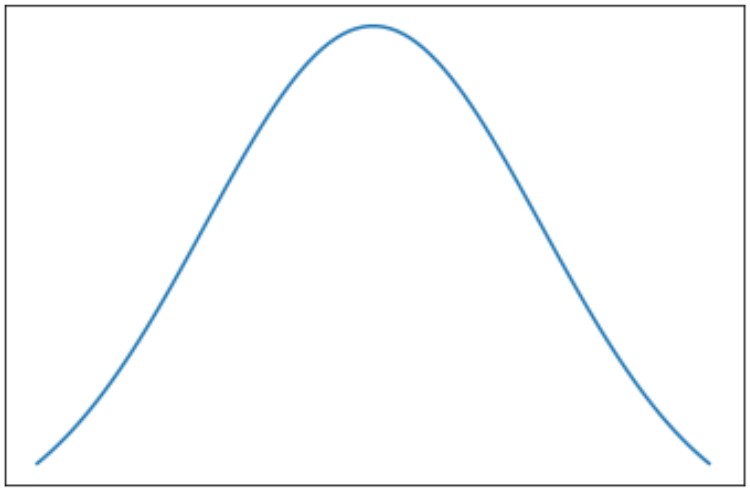
$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\text{argmax}} \ \log L(\boldsymbol{\theta}) = \underset{\boldsymbol{\theta}}{\text{argmax}} \sum_{i=1}^{n} \log p(y_i|\boldsymbol{\theta})$$

Maximising Likelihood

Let us assume the data is normally distributed

$$p(y_i|\boldsymbol{\theta}) \sim \mathcal{N}(\mu, \sigma^2) \text{ for } i = 1, \ldots, n$$

Maximise the likelihood to get the parameters



Maximising likelihood

$$p(y_i|\theta) \sim \mathcal{N}(\mu, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \frac{(y_i - \mu)^2}{\sigma^2}\right)$$

$$L(\theta) = p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta) \quad \text{for } i = 1, \ldots n$$

Maximising likelihood

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right)$$

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}}\ L(\theta) \qquad \theta = (\mu, \sigma)$$

$$\text{Log } L(\theta) = -\frac{n}{2}\log(2\pi\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}$$

$$\frac{\partial}{\partial \mu}\text{Log } L(\theta) = 0 \qquad \frac{\partial}{\partial \sigma}\text{Log } L(\theta) = 0$$

Maximising likelihood

$$\frac{\partial}{\partial \mu} \log L(\theta) = \sum_{i=1}^{n} \frac{y_i - \mu}{\sigma^2} = 0$$

$$\sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \mu = 0$$

$$\mu = \frac{1}{n} \sum_{i=1}^{n} y_i$$

$$\frac{\partial}{\partial \sigma} \log L(\theta) = -\frac{n}{2} \frac{\partial}{\partial \sigma} \log \sigma^2 - \frac{1}{2} \frac{\partial}{\partial \sigma} \left(\frac{1}{\sigma^2} \sum (y_i - \mu)^2 \right)$$

Maximising likelihood

$$= -\frac{n}{2}\frac{1}{\sigma^2}2\sigma - \frac{1}{2}\left(-\frac{2}{\sigma^3}\right)\sum_{i=1}^{n}(y_i-\mu)^2$$

$$= -\frac{n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i-\mu)^2 = 0$$

$$\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{\sigma^3} = \frac{n}{\sigma}$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i-\mu)^2}$$

Bayesian Modelling

- The parameters $\boldsymbol{\theta}$ can also be uncertain

Bayesian Modelling

- The parameters $\boldsymbol{\theta}$ can also be uncertain
  - The probability of raining tomorrow is 0.9
  - The probability of getting heads is 0.5
  - There is a one-in-a-thousand chance that Austria will win the next World Cup

Bayesian Modelling

- The parameters $\boldsymbol{\theta}$ can also be uncertain
  - The probability of raining tomorrow is 0.9
  - The probability of getting heads is 0.5
  - There is a one-in-a-thousand chance that Austria will win the next World Cup
- Uncertainty quantification on $\boldsymbol{\theta}$ using a prior distribution: $p(\boldsymbol{\theta})$

Prior

Let us assume the data $\mathbf{y} = (y_1, \ldots, y_n)$ is normally distributed (let as assume σ is given and $\boldsymbol{\theta} = \mu$) i.e.

$$p(y_i|\mu) \sim \mathcal{N}(\mu, \sigma^2) \text{ for } i = 1, \ldots, n$$

Prior

Let us assume the data $\mathbf{y} = (y_1, \ldots, y_n)$ is normally distributed (let as assume σ is given and $\boldsymbol{\theta} = \mu$) i.e.

$$p(y_i|\mu) \sim \mathcal{N}(\mu, \sigma^2) \text{ for } i = 1, \ldots, n$$

We also assume that the parameter μ also follows a normal distribution - Prior (or belief before looking at the data) i.e.

$$p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

Prior

Let us assume the data $\mathbf{y} = (y_1, \ldots, y_n)$ is normally distributed (let as assume σ is given and $\boldsymbol{\theta} = \mu$) i.e.

$$p(y_i|\mu) \sim \mathcal{N}(\mu, \sigma^2) \text{ for } i = 1, \ldots, n$$

We also assume that the parameter μ also follows a normal distribution - Prior (or belief before looking at the data) i.e.

$$p(\mu) \sim \mathcal{N}(\mu_0, \sigma_0^2)$$

How to find the $p(\mu|\mathbf{y})$?

Bayesian Inference

- Likelihood: $p(\mathbf{y}|\boldsymbol{\theta})$: probability that any $y = (y_1, \ldots, y_n)$ will occur for given $\boldsymbol{\theta}$

Bayesian Inference

- Likelihood: $p(\mathbf{y}|\boldsymbol{\theta})$: probability that any $y = (y_1, \ldots, y_n)$ will occur for given $\boldsymbol{\theta}$
- Prior: $p(\boldsymbol{\theta})$: randomness or uncertainty on the parameter $\boldsymbol{\theta}$ before observing any data

Bayesian Inference

- Likelihood: $p(\mathbf{y}|\boldsymbol{\theta})$: probability that any $y = (y_1, \ldots, y_n)$ will occur for given $\boldsymbol{\theta}$
- Prior: $p(\boldsymbol{\theta})$: randomness or uncertainty on the parameter $\boldsymbol{\theta}$ before observing any data
- Posterior: $p(\boldsymbol{\theta}|\mathbf{y})$: complete probability distribution which can quantify the uncertainty in the parameters $\boldsymbol{\theta}$

Bayesian Inference

Bayes' rule:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

Bayesian Inference

Bayes' rule:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

The marginal distribution of $\mathbf{y}$ is for the normalisation and can also be written as:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Bayesian Inference

Bayes' rule:

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})}$$

The marginal distribution of $\mathbf{y}$ is for the normalisation and can also be written as:

$$p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$$

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Bayesian Inference

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

Bayesian Inference

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

$$\log p(\boldsymbol{\theta}|\mathbf{y}) \propto \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Bayesian Inference

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

$$\log p(\boldsymbol{\theta}|\mathbf{y}) \propto \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

Instead of Maximum likelihood estimation (MLE), we now do the Maximum A Posterior Estimation (MAP) to get the most likely parameter values

Bayesian Inference

$$p(\mu|y) \propto p(y|\mu)\, p(\mu)$$

$$p(y|\mu) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right)$$

$$p(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{1}{2}\frac{(\mu-\mu_0)^2}{\sigma_0^2}\right)$$

$$p(\mu|y) = (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \cdot (2\pi\sigma_0^2)^{-1/2} \exp\left(-\frac{1}{2\sigma_0^2}(\mu-\mu_0)^2\right)$$

Bayesian Inference

$$\propto \exp\left(-\frac{\sum_{i=1}^{n} (y_i - \mu)^2}{2\sigma^2} - \frac{(\mu - \mu_0)^2}{2\sigma_0^2} \right)$$

$$\propto \mathcal{N}(\mu_1, \sigma_1^2)$$

$$\sigma_1^2 = \frac{1}{n\sigma^{-2} + \sigma_0^{-2}} \qquad \mu_1 = \sigma_1^2 \left(\mu_0 \sigma_0^{-2} + \sum_{i=1}^{n} y_i \sigma^{-2} \right)$$

Prior

Prior distribution -

- is important to reveal any information on the model parameters before observing any data
- can rule out values which are implausible

How to choose a prior?

- To rule out impossible values or distributions

How to choose a prior?

- To rule out impossible values or distributions
- If there is little or no knowledge, uninformative/weakly informative prior distributions can be used i.e. assigning the same probability to all parameter values

How to choose a prior?

- To rule out impossible values or distributions
- If there is little or no knowledge, uninformative/weakly informative prior distributions can be used i.e. assigning the same probability to all parameter values
- Non-informative - Normal distribution with a constant mean and very large variance - $p(\theta)$ is approximately the same for all values of θ

How to choose a prior?

- To rule out impossible values or distributions
- If there is little or no knowledge, uninformative/weakly informative prior distributions can be used i.e. assigning the same probability to all parameter values
- Non-informative - Normal distribution with a constant mean and very large variance - $p(\theta)$ is approximately the same for all values of θ
- Weakly informative - Normal distribution with a constant mean and a large variance (e.g. 50-100) - $p(\boldsymbol{\theta})$ same for $\boldsymbol{\theta}$ which are plausible

How to choose a prior?

- To rule out impossible values or distributions
- If there is little or no knowledge, uninformative/weakly informative prior distributions can be used i.e. assigning the same probability to all parameter values
- Non-informative - Normal distribution with a constant mean and very large variance - $p(\theta)$ is approximately the same for all values of θ
- Weakly informative - Normal distribution with a constant mean and a large variance (e.g. 50-100) - $p(\boldsymbol{\theta})$ same for $\boldsymbol{\theta}$ which are plausible
- Informative - Normal distribution with a constant mean and very small variance

Bayesian Inference

Conjugate priors

Parameter	Prior Distribution
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Normal (fixed variance)	Normal for the mean
Normal (fixed mean)	Gamma prior for the inverse variance
Exponential	Gamma
Multinomial	Dirichlet

Conjugacy can help in providing a closed form of the integral, and you do not rely on numerical integration

Markov Chain Monte Carlo

- There may not be a closed form for the posterior

$$\log p(\boldsymbol{\theta}|\mathbf{y}) \propto \log p(\mathbf{y}|\boldsymbol{\theta}) + \log p(\boldsymbol{\theta})$$

- Markov Chain Monte Carlo (MCMC) can be used to estimate the posterior
- There are several MCMC methods - No U turn sampler, Metropolis-Hastings

Summary

- Three key terms to remember: Prior, Likelihood and Posterior
- Prior - belief before observing or looking at the data - randomness or uncertainty in parameters $p(\boldsymbol{\theta})$
- Likelihood - probability that any $\mathbf{y} = (y_1, \ldots, y_n)$ will occur for given $\boldsymbol{\theta}$
- Posterior - obtained using Bayes rule - complete probability distribution on $\boldsymbol{\theta}$ i.e. $p(\boldsymbol{\theta}|\mathbf{y})$ - analytically or using MCMC

Gaussian Processes

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

Gaussian Processes

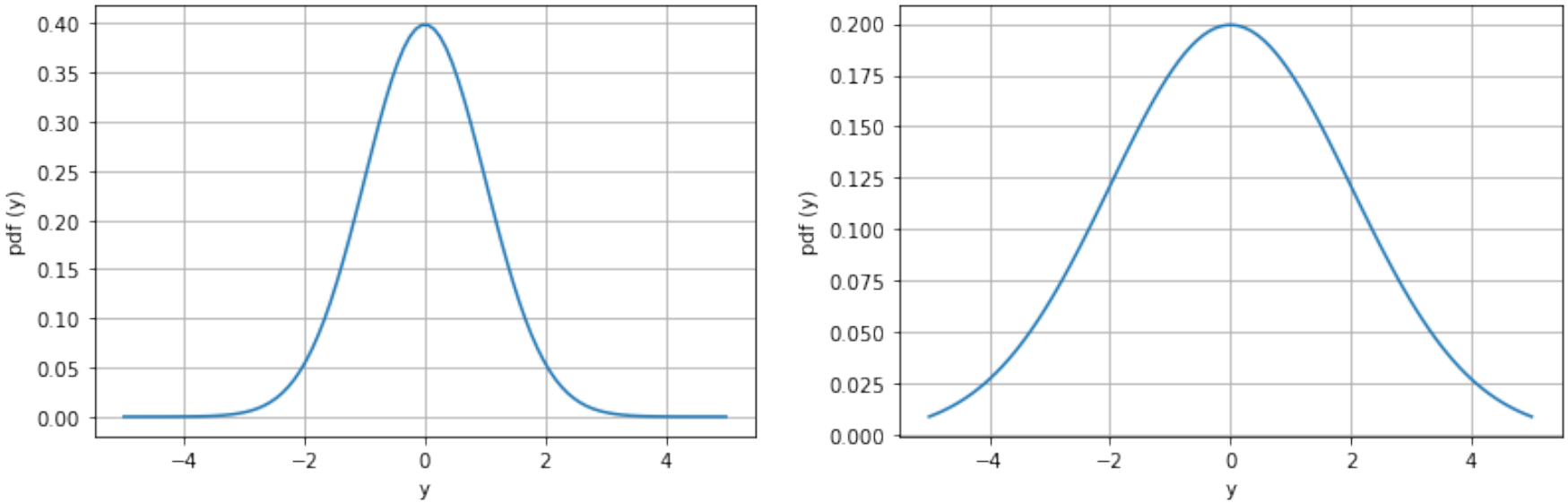


Figure: Two univariate normal distributions with means (0,0) and standard deviation (1,2)

Gaussian Processes

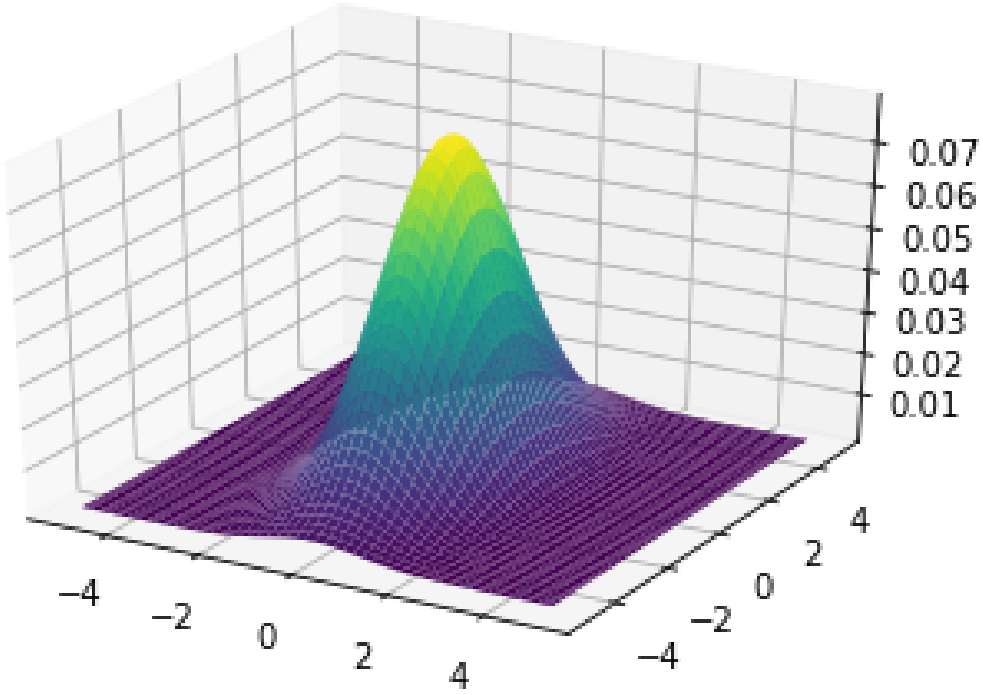


Figure: Joint distribution of two univariate normal distributions is also Normal with means (0,0) and covariance matrix [1,0; 0,2]

Gaussian Processes

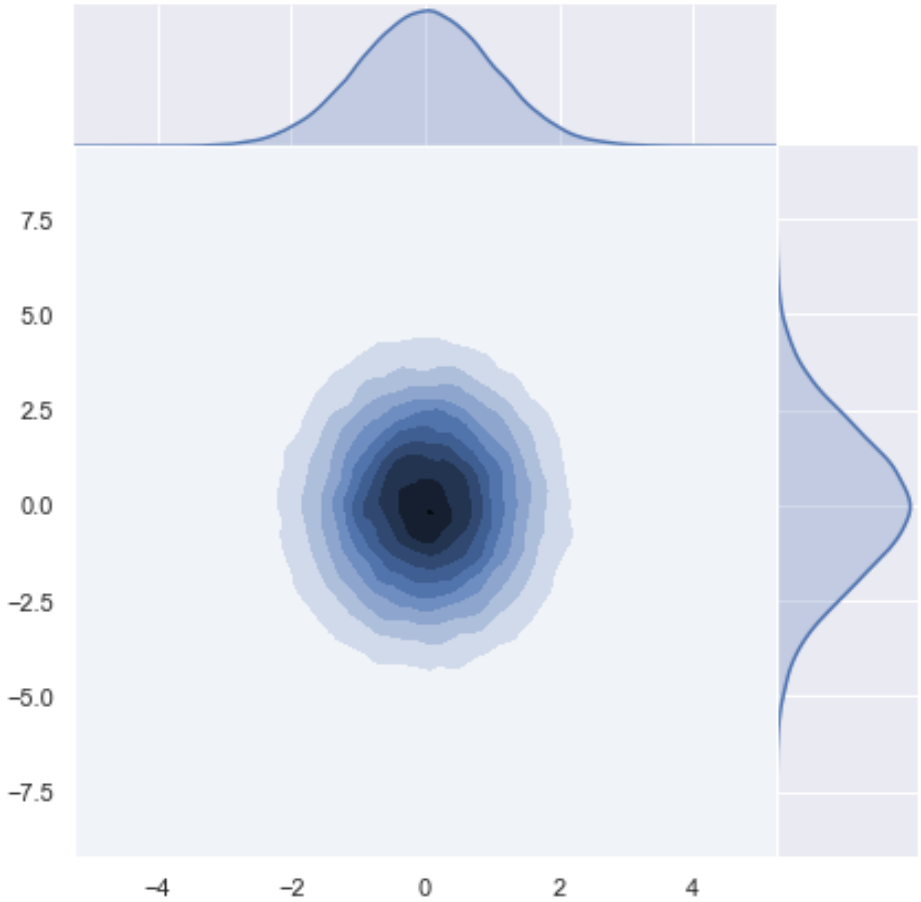


Figure: Joint distribution of two univariate normal distributions is also Normal with means (0,0) and covariance matrix [1,0; 0,2]

Gaussian Processes

In regression we need a function $y \sim f(\mathbf{x})$, which can be defined as:

$$\begin{aligned} f(\mathbf{x}) &\sim \mathcal{N}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')) \\ &\sim \mathcal{N}(\mathbf{0}, K) \end{aligned}$$

where $m(\mathbf{x})$ is the mean function and $k(\mathbf{x}, \mathbf{x}')$ is the covariance function (or kernel)

For simplicity in mathematical calculations, we will be using $m(\mathbf{x}) = \mathbf{0}$

Gaussian Processes

Let us use squared exponential (SQ) kernel (also known as RBF or Gaussian kernel):

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \exp\left(-0.5|x - x'|^2\right) \\ &= cov(f(\mathbf{x}), f(\mathbf{x}')) \end{aligned}$$

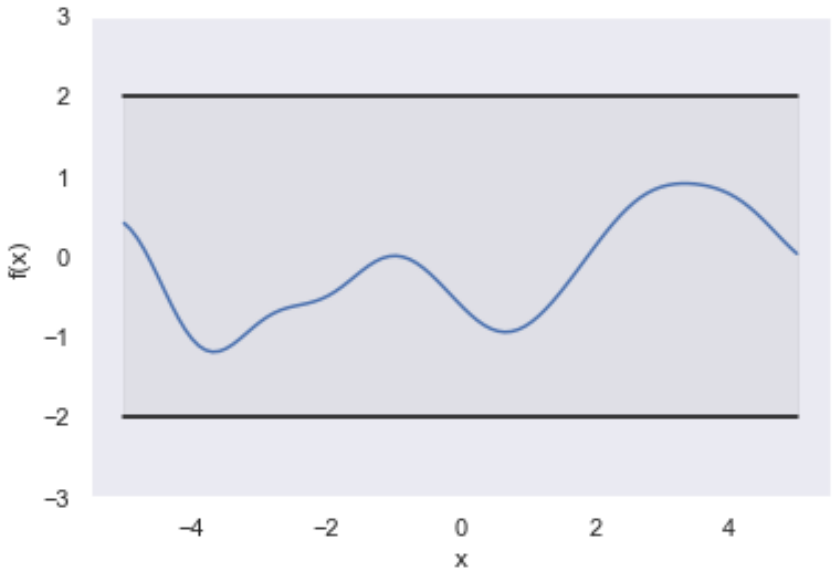
- The squared exponential kernel is infinitely differentiable
- Covariance function between the outputs is used as a function of inputs.
- Specification of covariance function implies a distribution over functions

Gaussian Processes

- Let us draw $f(\mathbf{x})$ with a multivariate Normal distribution using the squared exponential kernel

Gaussian Processes

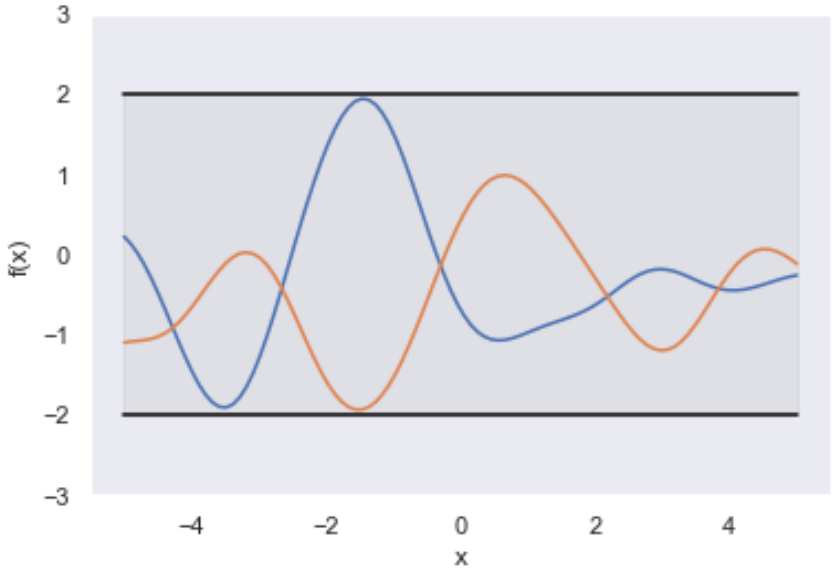
In 1-D: sampling a function randomly using prior distribution over the functions



The prior represent our belief before we observe any data

Gaussian Processes

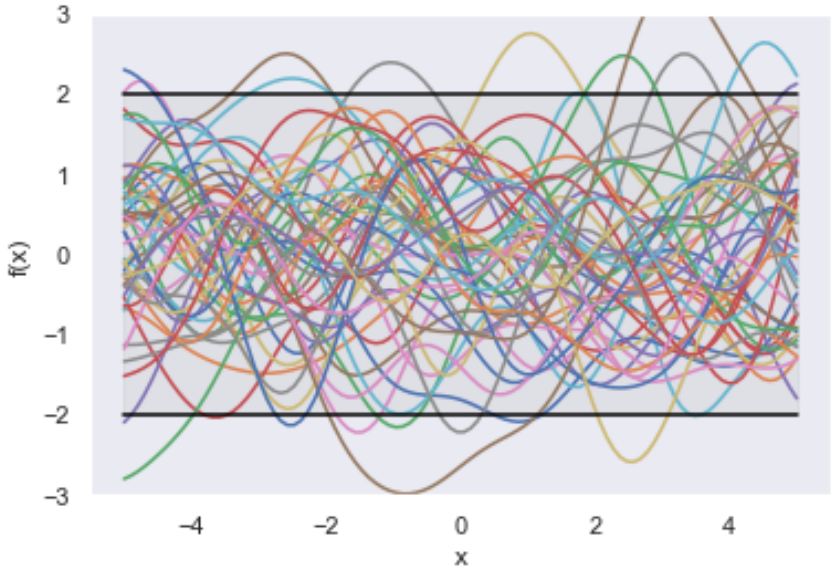
We can have two such functions:



The prior represent our belief before we observe any data

Gaussian Processes

Or we can have infinitely many such functions



The prior represent our belief before we observe any data

Gaussian Processes

Some observations from the prior functions:

1. The functions have a *length scale* - can be interpreted as distance you have to move in x space before $f(\mathbf{x})$ can change significantly

Gaussian Processes

Some observations from the prior functions:

1. The functions have a *length scale* - can be interpreted as distance you have to move in x space before $f(\mathbf{x})$ can change significantly
2. Overall variance of process (or joint of all functions) is around one

Gaussian Processes

- To control these two parameters: we can change the formulation of SQ kernel as:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-0.5 \frac{|\mathbf{x} - \mathbf{x}'|^2}{l^2}\right)$$

where l is the length scale and σ_f is the amplitude (signal variance) to control the variance

Gaussian Processes

- To control these two parameters: we can change the formulation of SQ kernel as:

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-0.5 \frac{|\mathbf{x} - \mathbf{x}'|^2}{l^2}\right)$$

  where l is the length scale and σ_f is the amplitude (signal variance) to control the variance
- How to set these two parameters - we will come back here
- As for now, assume they are equal to one.

Gaussian Processes

- We are not primarily interested in prior functions
- We are interested in training and prediction

Gaussian Processes

- Given a data set: $(X, \mathbf{y})$,, the posterior predictive distribution of new points X^* is also Gaussian:

$$p(\mathbf{y}^*|X^*, X, \mathbf{y}) \sim \mathcal{N}(K(X^*X)K(X,X)^{-1}\mathbf{y}, \\ K(X^*, X^*) - K(X^*, X)K(X,X)^{-1}K(X, X^*))$$

K is covariance matrix in which $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$

Gaussian Processes

1. Let us imagine we got some data set for training
2. What is the prediction on new samples using Gaussian processes - posterior predictive distribution?

Gaussian Processes

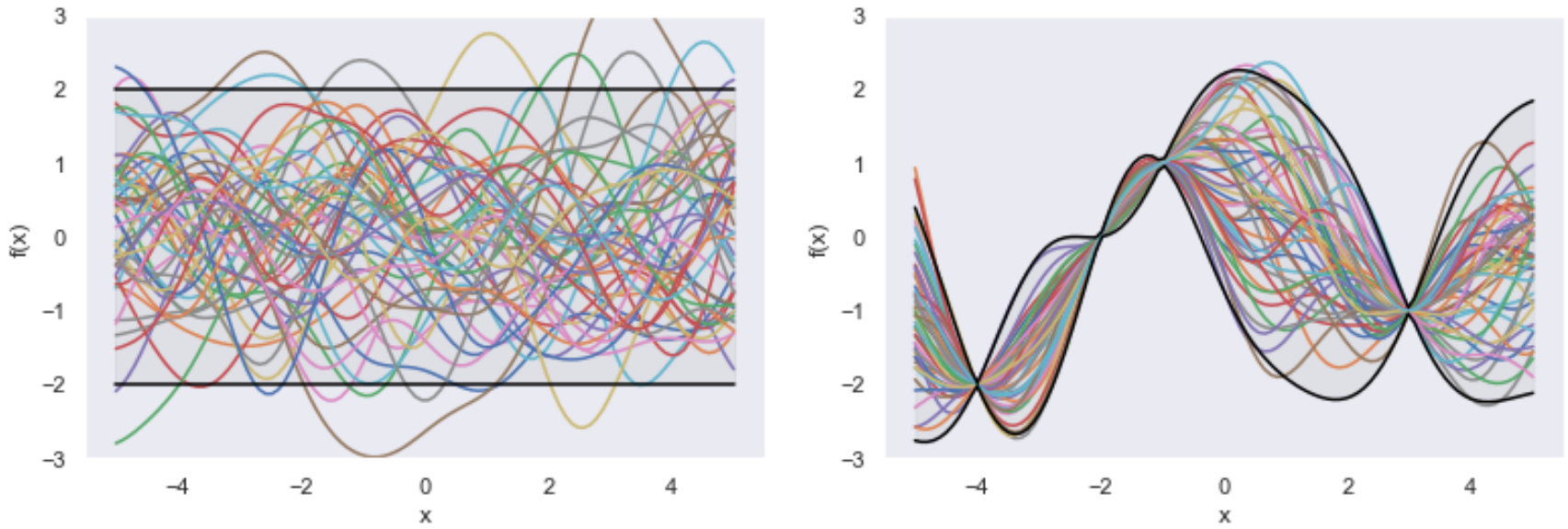


Figure: Prior functions (left) and posterior functions (right) drawn from Gaussian Processes. The black lines represent the mean plus and minus two standard deviations

Gaussian Processes

- For noisy observations,

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-0.5 \frac{|\mathbf{x} - \mathbf{x}'|^2}{l^2}\right) + \sigma_n^2 \delta_{pq} \text{ OR}$$

$$K = K + \sigma_n^2 I,$$

where δ_{pq} is a Kronecker delta which is one iff p = q and zero otherwise and σ_n^2 is the variance of noise which is assumed to be same for same data points (homoscedastic).

Gaussian Processes

Summary so far:

- A stochastic process is a generalisation of a probability distribution to functions
- These functions are Gaussian
- Combination of prior over functions and the data leads to posterior over functions
- For prior, we need to find the covariance function

Gaussian Processes

Summary so far:

- We do not need to worry about basis functions - as they are inbuilt in the covariance function
- Learning in Gaussian process is finding the covariance function
- Covariance function - length scale, amplitude (noise free data) and noise variance (noisy data)

Gaussian Processes

Learning in Gaussian Processes - Finding the parameters of the covariance function

$$k(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-0.5 \frac{|\mathbf{x} - \mathbf{x}'|^2}{l^2}\right) + \sigma_n^2 \delta_{pq},$$

where δ_{pq} is a Kronecker delta which is one iff p = q and zero otherwise.

- We need to maximise the marginal likelihood to get the covariance function parameters

Gaussian Processes

Before we maximise marginal likelihood, let us find the importance of different parameters in the covariance function

Gaussian Processes

Effect of length scale (l):

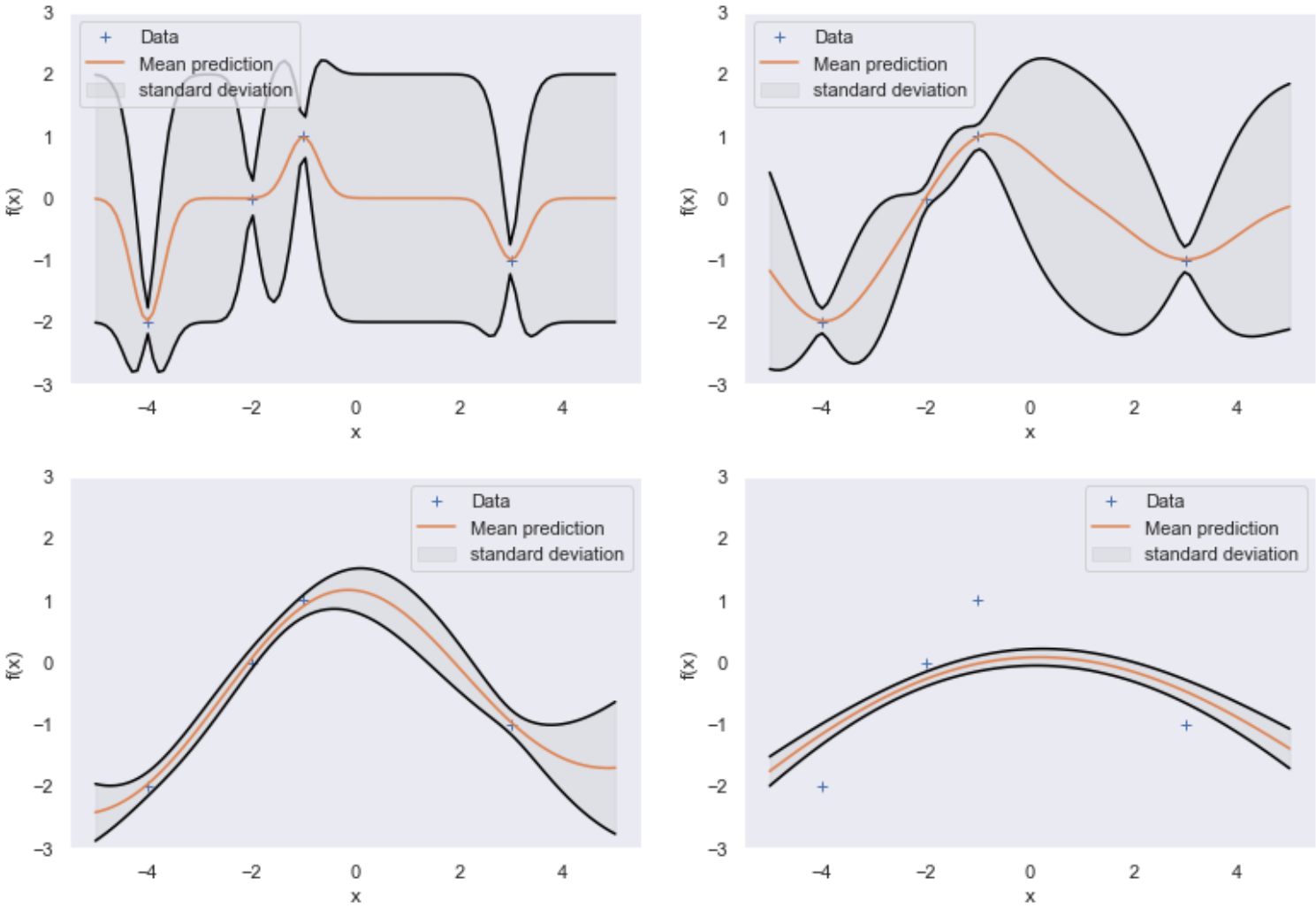


Figure: Effect of length scale, l = 0.3, 1, 3 and 10. $\sigma_f = 1$ and $\sigma_n = 0.1$

Gaussian Processes

Effect of amplitude (σ_f)

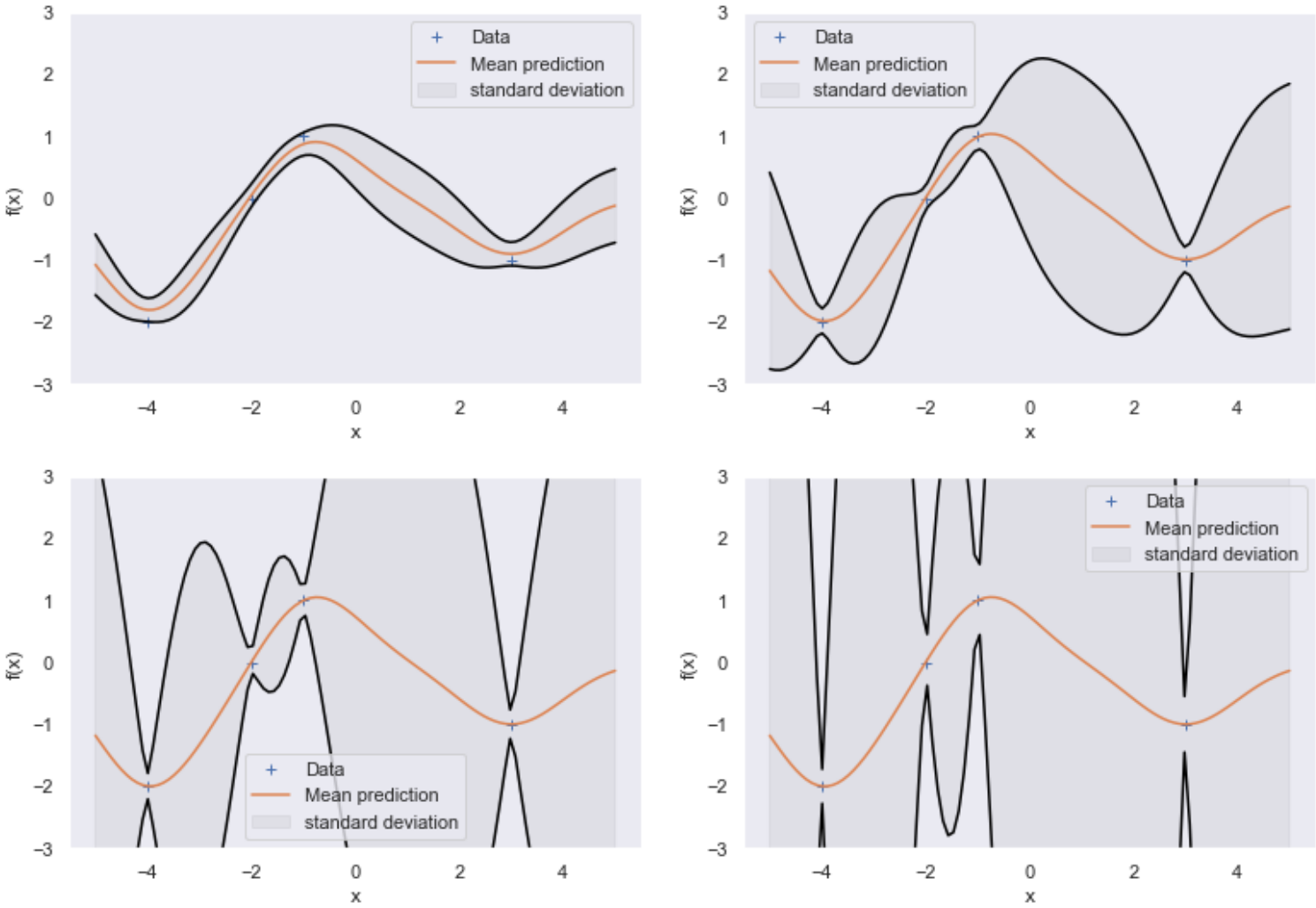


Figure: Effect of amplitude, $\sigma_f = 0.3$, 1, 3 and 10. $l = 1$ and $\sigma_n = 0.1$

Gaussian Processes

Effect of noise (σ_n)

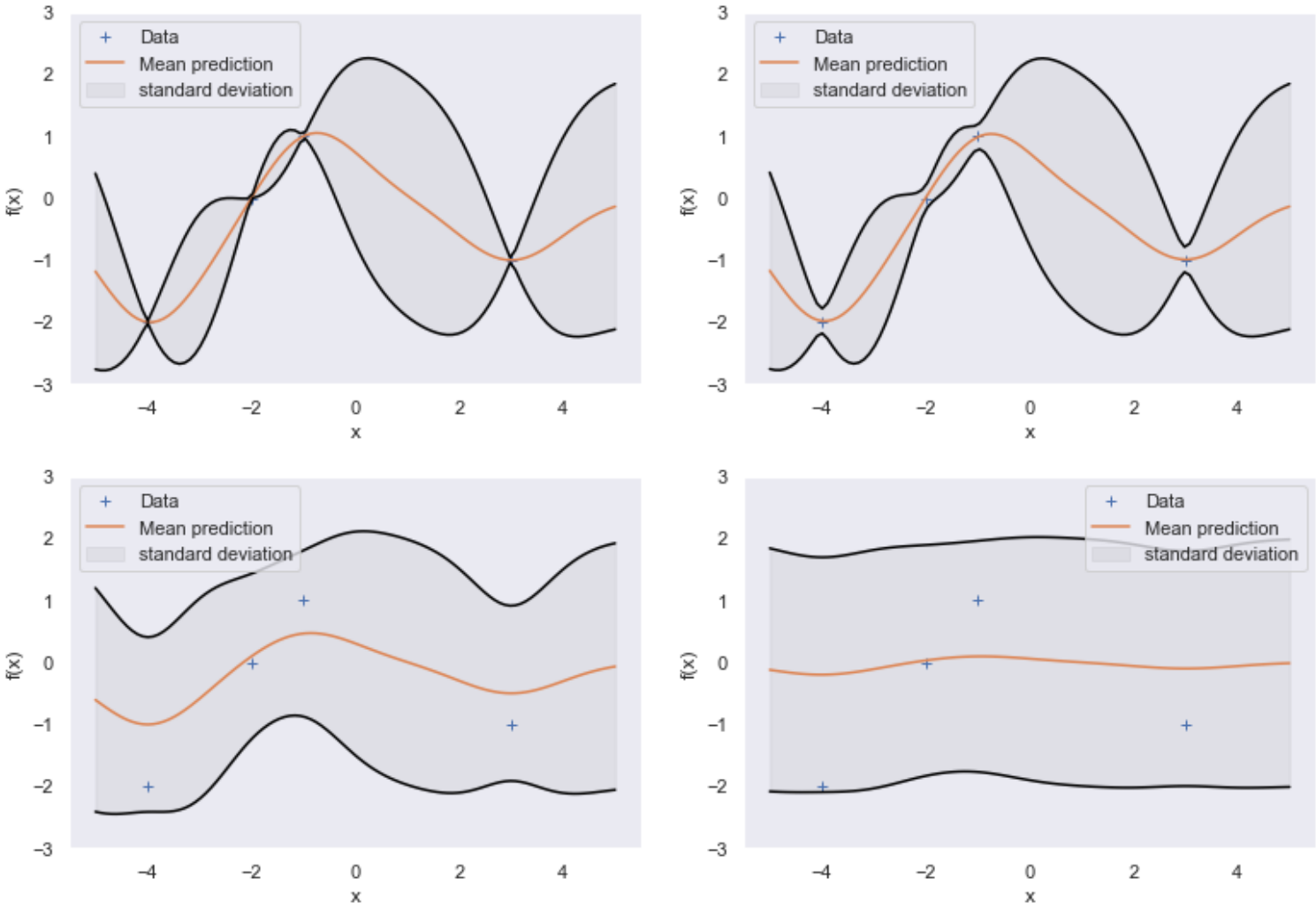


Figure: Effect of amplitude, $\sigma_n = 0$, 0.1, 1 and 3. $l = 1$ and $\sigma_f = 1$

Gaussian Processes

Estimating the covariance function parameters by maximising the marginal likelihood:

$$p(\mathbf{y}|X) = \int p(\mathbf{y}|\mathbf{f}, X)p(\mathbf{f}|X)d\mathbf{f}$$

It is the marginalisation over functions $\mathbf{y} = f(X)$. As the functions are Gaussian:

$$\log p(\mathbf{y}|X) = -0.5\mathbf{y}^T K^{-1}\mathbf{y} - 0.5|K| - \frac{n}{2}\log 2\pi$$

Gaussian Processes

$$l, \sigma_f, \sigma_n = \underset{l,\sigma_f,\sigma_n}{\text{argmax}} \ \text{Marginal Log-Likelihood}$$

Gaussian Processes

After training the model, we are interested in prediction:

$$\mathbf{f}^*|X^*, X, \mathbf{f} \sim \mathcal{N}\Big(K(X^*, X)K(X, X)^{-1}\mathbf{f}, \\ K(X^*, X^*) - K(X^*, X)K(X, X)^{-1}K(X, X^*)\Big)$$

Summary

- **Bayesian Optimisation** - needs uncertainty in prediction in addition to point prediction
- **Gaussian processes** - prior over functions, provide uncertainty in predictions
- **Bayesian models** such as Bayesian NN can also be used in BO