# Machine Learning

## Bayesian Optimisation

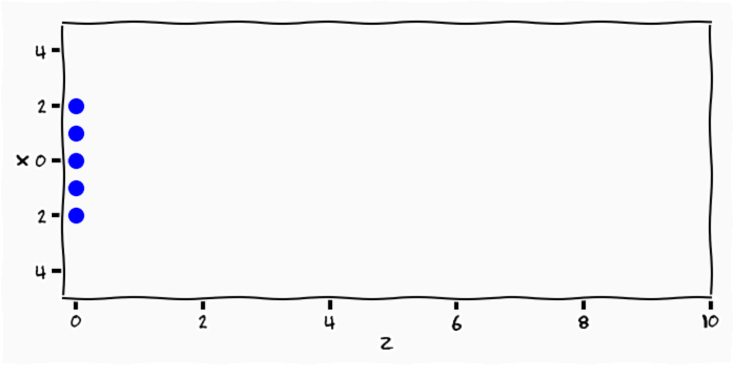
Carl Henrik Ek - carlhenrik.ek@bristol.ac.uk

October 28, 2019

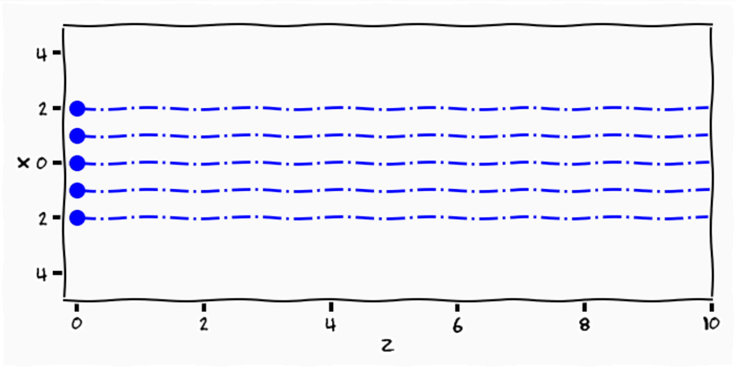
http://www.carlhenrik.com

# Introduction

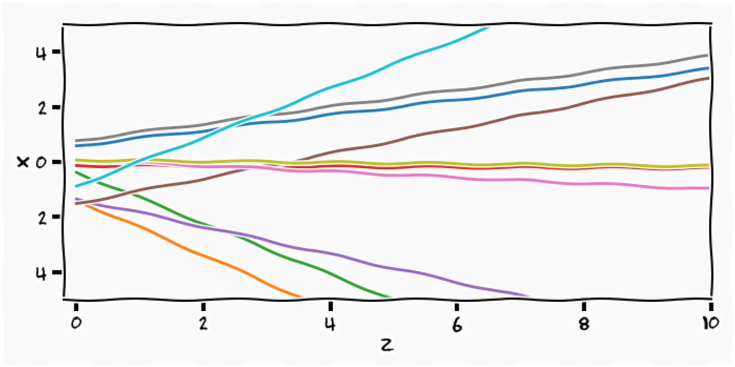
## Unsupervised Learning



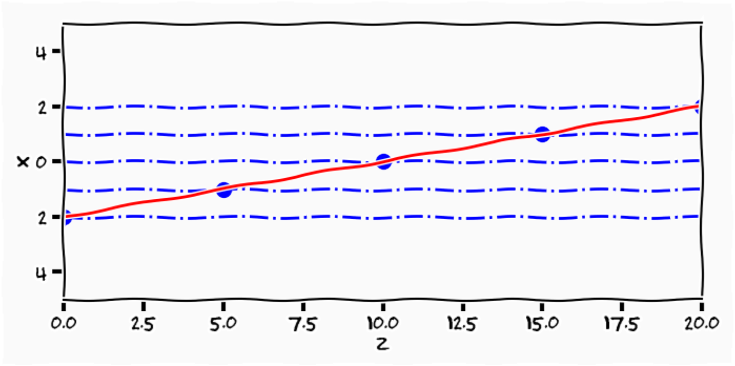
# Unsupervised Learning



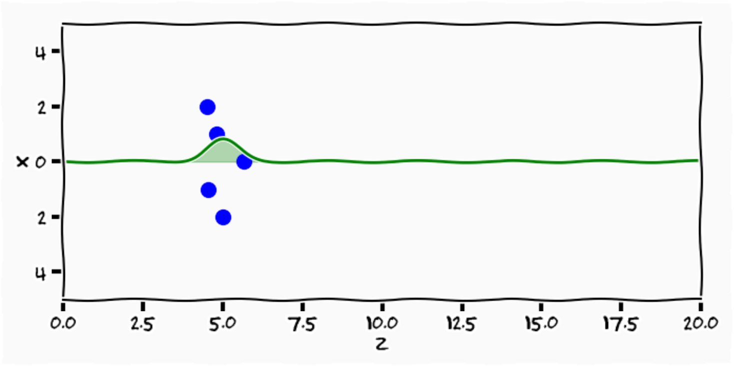
# Unsupervised Learning

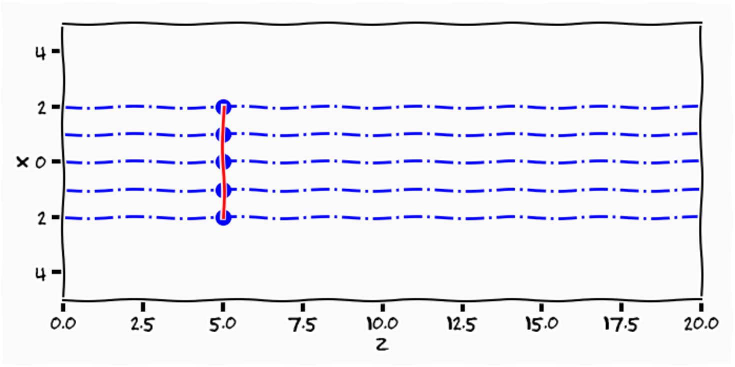


# Unsupervised Learning

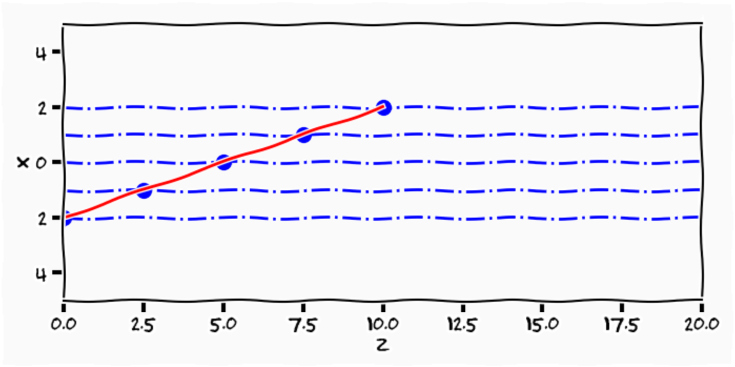


# Unsupervised Learning

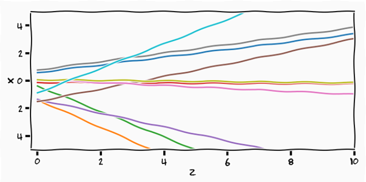


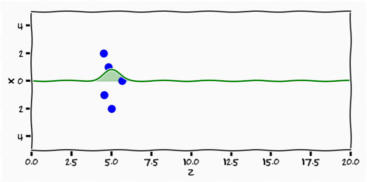

4
2
x 0
2
4
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
z

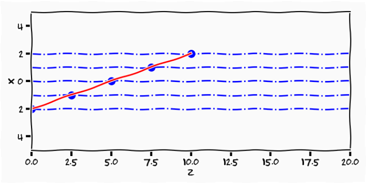
# Unsupervised Learning



# Unsupervised Learning



# Linear Latent Variable Models [1] 12.2

- Linear Regression

$$p(\mathbf{W}|\mathbf{t}, \mathbf{X}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

# Linear Latent Variable Models [1] 12.2

- Linear Regression

$$p(\mathbf{W}|\mathbf{t}, \mathbf{X}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

- Linear Unsupervised Learning

$$p(\mathbf{W}, \mathbf{X}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})p(\mathbf{X})$$

# Linear Latent Variable Models [1] 12.2

- Linear Regression

$$p(\mathbf{W}|\mathbf{t}, \mathbf{X}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

- Linear Unsupervised Learning

$$p(\mathbf{W}, \mathbf{X}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})p(\mathbf{X})$$
$$p(\mathbf{W}, \mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{W}, \mathbf{z})p(\mathbf{W})p(\mathbf{z})$$

# Linear Latent Variable Models [1] 12.2

- Linear Regression

$$p(\mathbf{W}|\mathbf{t}, \mathbf{X}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})$$

- Linear Unsupervised Learning

$$p(\mathbf{W}, \mathbf{X}|\mathbf{t}) \propto p(\mathbf{t}|\mathbf{W}, \mathbf{X})p(\mathbf{W})p(\mathbf{X})$$
$$p(\mathbf{W}, \mathbf{z}|\mathbf{x}) \propto p(\mathbf{x}|\mathbf{W}, \mathbf{z})p(\mathbf{W})p(\mathbf{z})$$

- Actually formulating the posterior over both $\mathbf{W}$ and $\mathbf{z}$ is intractable

# Type II Maximum Likelihood

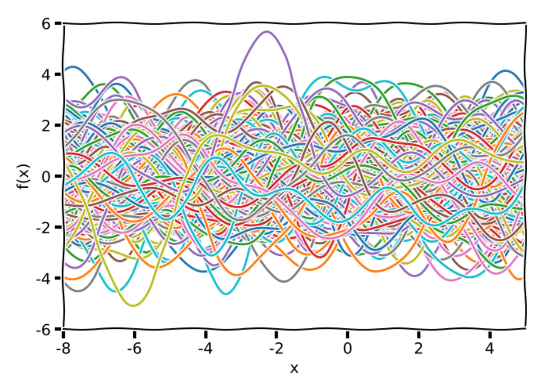
$$\hat{\mathsf{W}} = \text{argmax}_{\mathsf{W}} p(\mathsf{x}|\mathsf{W}) = \int p(\mathsf{x}|\mathsf{W}, \mathsf{z}) p(\mathsf{z}) \mathrm{d}\mathsf{z}$$

1. Intractable to reach posterior of both
2. Integrate out one variable, marginalise, expectation
3. Take point-estimate of the remaining

# Principal Component Analysis

- You might have seen this explained in a different way
  - *Retain variance*
  - *Error minimisation*
- These provides the same solution as the maximum likelihood but solved by an eigenvalue problem
- Do not provide intuition as it doesn't state assumptions

# Non-linear latent variable model




$$p(\mathsf{x}, \mathsf{z}|\theta) = \int p(\mathsf{x}|\mathbf{f})p(\mathbf{f}|\mathsf{z}, \theta)p(\mathsf{z})\mathrm{d}\mathbf{f}$$

# Unsupervised Learning

- Unsupervised learning is a misnomer, there is no such thing, you have to have beliefs in order to learn.
- Think about unsupervised learning as "less supervised" learning, you have to have stronger beliefs

# Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

2. Formulate your likelihood

$$p(\mathbf{Y}|\theta = \theta_i)$$

# Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

2. Formulate your likelihood

$$p(\mathbf{Y}|\theta = \theta_i)$$

3. Formluate your belief in the "setting" of the model

$$p(\theta)$$

# Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations
$$p(\mathcal{Y}|\theta)$$
2. Formulate your likelihood
$$p(\mathbf{Y}|\theta = \theta_i)$$
3. Formluate your belief in the "setting" of the model
$$p(\theta)$$
4. Aquire data

# Machine Learning

1. Specify your statistical model over sample space $\mathcal{Y}$, relationship between "parameters" and observations

$$p(\mathcal{Y}|\theta)$$

2. Formulate your likelihood

$$p(\mathbf{Y}|\theta = \theta_i)$$

3. Formluate your belief in the "setting" of the model

$$p(\theta)$$

4. Aquire data
5. Derive your updated belief, derive knowledge from data

$$p(\theta|\mathbf{Y}) = \frac{p(\mathbf{Y}|\theta)p(\theta)}{\int p(\mathbf{Y}|\theta)p(\theta)}$$

# Next lectures

| | |
|---|---|
| **Today** | Bayesian Optimisation |
| **Tuesday** | Evidence and Graphical Models |
| **Monday** | Dirichlet Processes |
| **Tuesday** | Topic Models |
| **Monday** | Neural Networks |
| **Tuesday** | Reinforcement Learning |

The aim is for you to connect these things to what you have learnt so far and see that what the methodology allows you to do and different tools following the same methodology. Try to see that it is all the same thing.

# Bayesian Optimisation

# Uncertainty

- Deterministic world

$$x = 4$$

- Point estimate world

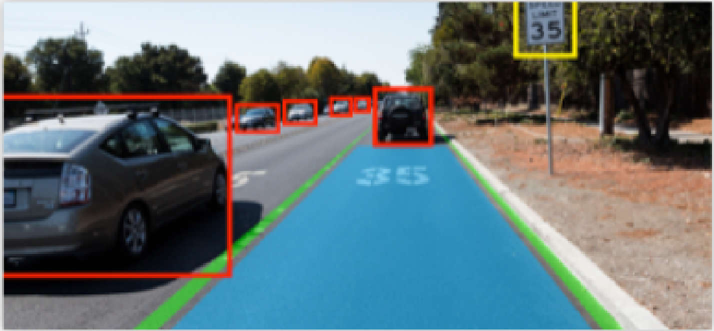
$$\text{argmax}_x p(x) = 4$$

- Stochastic world

$$p(x) = \mathcal{N}(4, 10^2)$$

# Uncertainty

# Uncertainty



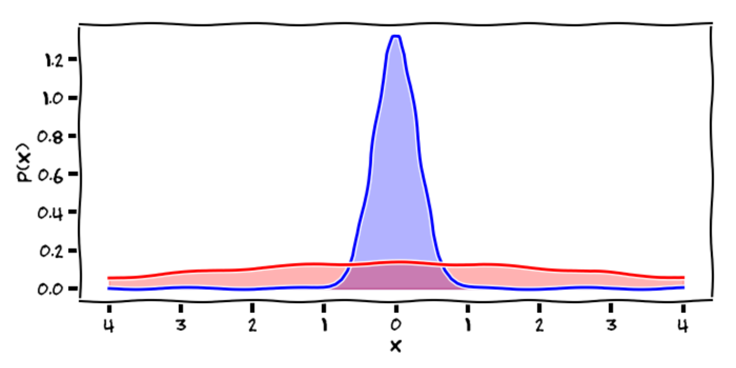
# The Uber Crash

# The Uber Crash[1]

*..the system first registered radar and LIDAR observations of the pedestrian about 6 seconds before impact, when the vehicle was traveling at 43 mph. As the vehicle and pedestrian paths converged, the self-driving system software classified the pedestrian as an unknown object, as a vehicle, and then as a bicycle with varying expectations of future travel path.*

[1] Link to report

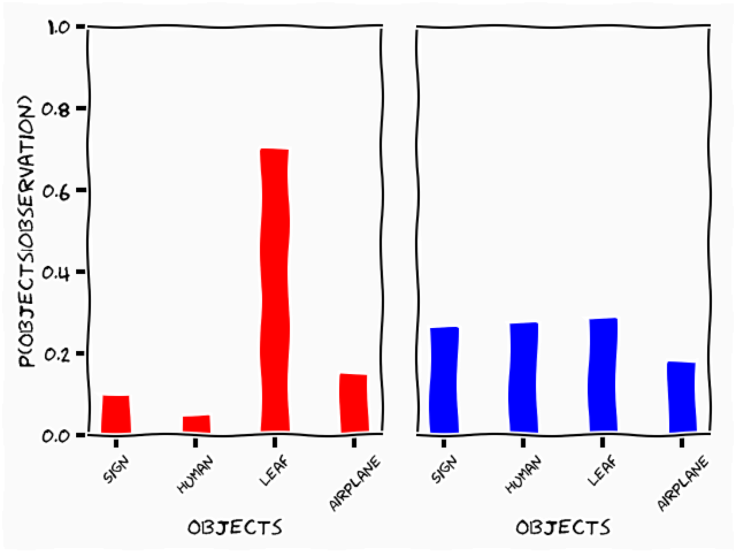
# The importance of uncertainty




$$\hat{x} = \mathrm{argmax}_x \, p(x) = \mathrm{argmax}_x \, p(x)$$

$$p(\hat{x}) \neq p(\hat{x})$$

# The importance of uncertainty



# The Uber Crash[1]

*According to Uber, emergency braking maneuvers are not enabled while the vehicle is under computer control, to reduce the potential for erratic vehicle behavior. The vehicle operator is relied on to intervene and take action. The system is not designed to alert the operator.*

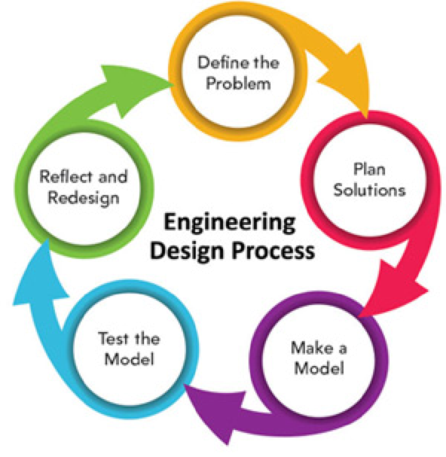
# Real Doctors

# Real Doctors

# Medicine

# Data science

# Design


Define the Problem
Plan Solutions
Make a Model
Test the Model
Reflect and Redesign
Engineering Design Process

# Optimisation

- All problems above can be seen as optimisation problems
- Classic optimisation

$$\hat{x} = \mathrm{argmin}_x f(x)$$

- Much more common

$$\hat{x} = \mathrm{argmin}_x \text{black-box}(x)$$

# Optimisation

- in most cases we have an objective function that we do not know

# Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function

# Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge

# Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge
- the cost of each test is expensive

# Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge
- the cost of each test is expensive
- the test is noisy

# Optimisation

- in most cases we have an objective function that we do not know
- we can test something and see how it works, i.e. evaluate the function
- the number of parameters, possible tests are huge
- the cost of each test is expensive
- the test is noisy
- *can we use machine learning to do this for us?*

# Global Optimisation

$$x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$$

- $\mathcal{X}$ is a bounded domain
- $f$ is explicitly unknown
- Evaluations of $f$ may be noisy
- Evaluations of $f$ is expensive

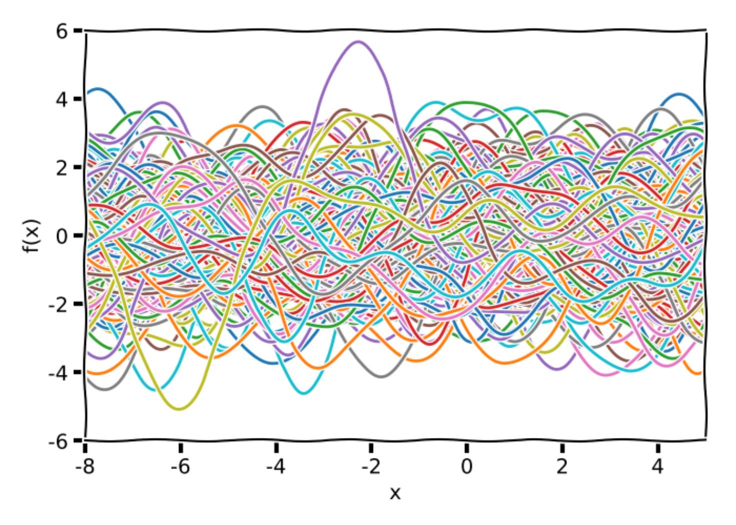
# Strategies

- Implicit knowledge
- Grid search

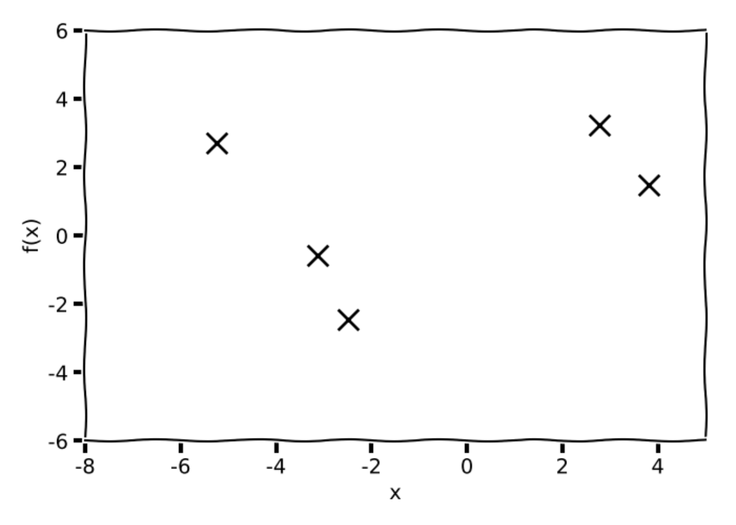
$$f(x_M) - f(x'_M) \leq \epsilon$$

  we need $(L/\epsilon)^D$ evaluations (if f is L-Lipschitz)
- Random sampling

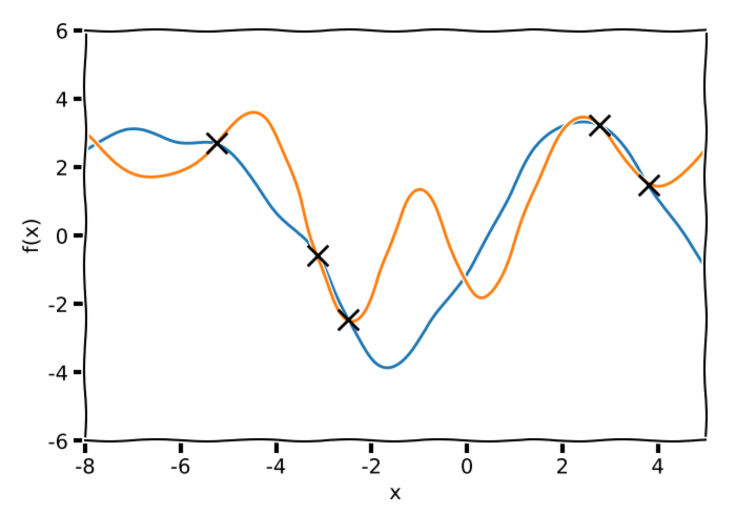
# BO in practice



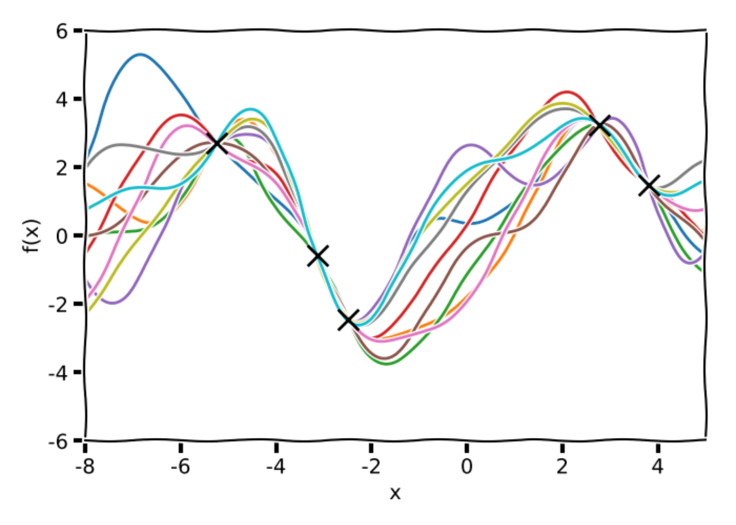
# BO in practice



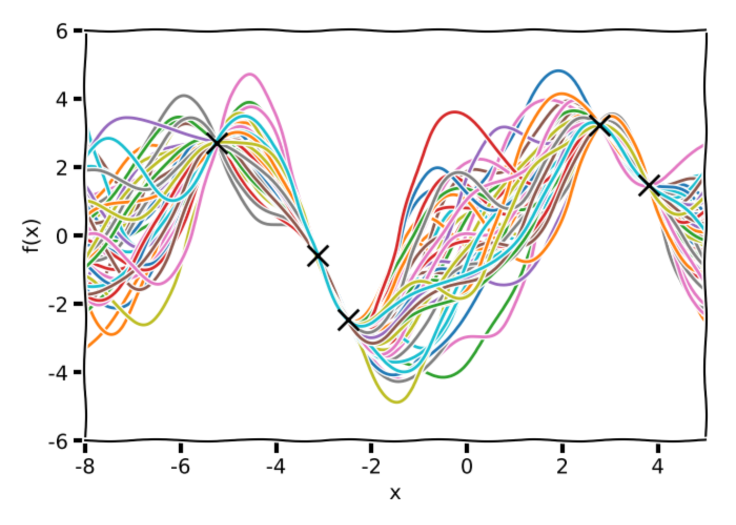
# BO in practice



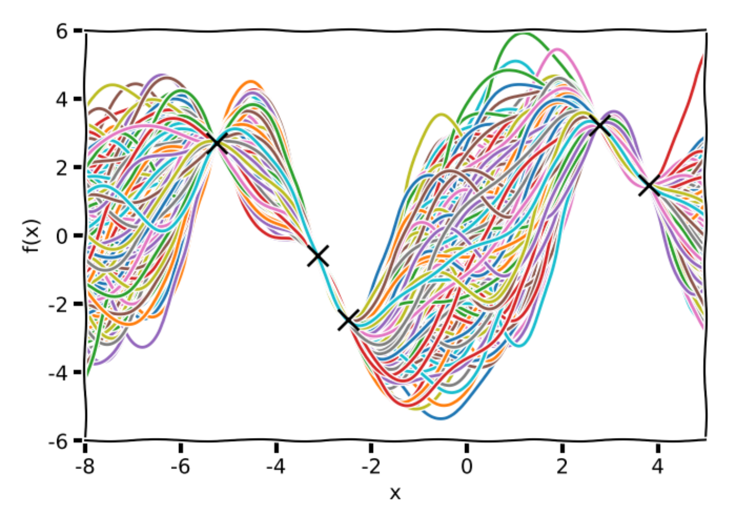
# BO in practice



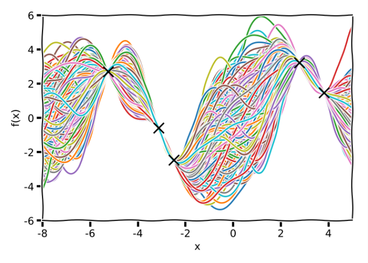
# BO in practice

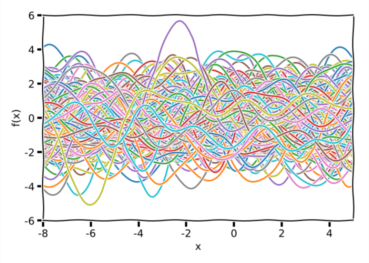


# BO in practice



# What did we actually do




$$p(f_*|f, \mathbf{X}, \mathbf{x}_*) = \frac{p(f_*, f|\mathbf{X}, \mathbf{x}_*)}{p(f|\mathbf{X})}$$

# Bayesian Optimisation

1. Choose a prior over the space of possible objective functions $f$

# Bayesian Optimisation

1. Choose a prior over the space of possible objective functions $f$
2. Combine prior and likelihood to get a posterior over the space

# Bayesian Optimisation

1. Choose a prior over the space of possible objective functions $f$
2. Combine prior and likelihood to get a posterior over the space
3. Use posterior to choose a set of evaluation according to a strategy

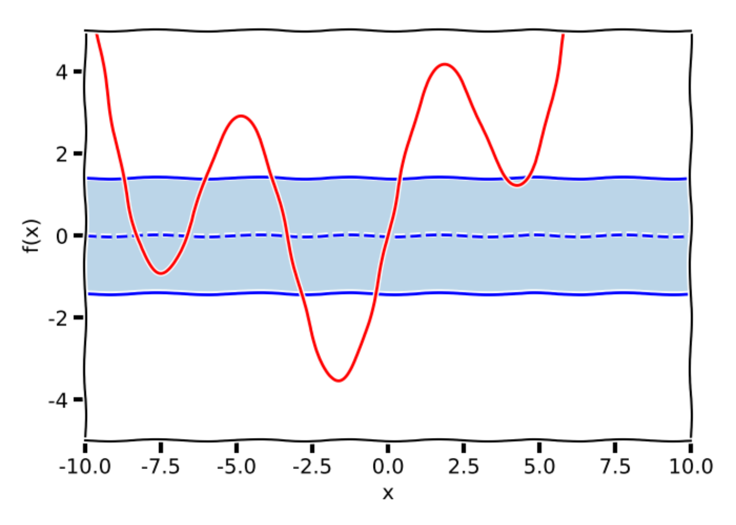
# Bayesian Optimisation

1. Choose a prior over the space of possible objective functions $f$
2. Combine prior and likelihood to get a posterior over the space
3. Use posterior to choose a set of evaluation according to a strategy
4. Add new data and update posterior

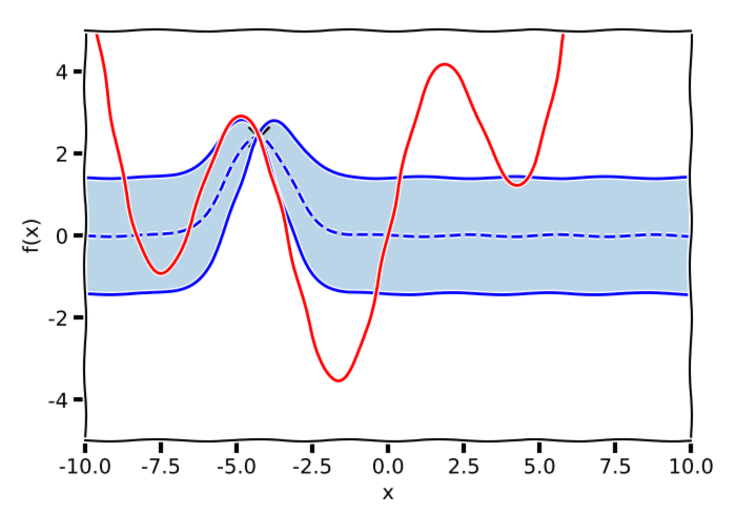
# Bayesian Optimisation

1. Choose a prior over the space of possible objective functions $f$
2. Combine prior and likelihood to get a posterior over the space
3. Use posterior to choose a set of evaluation according to a strategy
4. Add new data and update posterior
5. Repeate until budget is gone

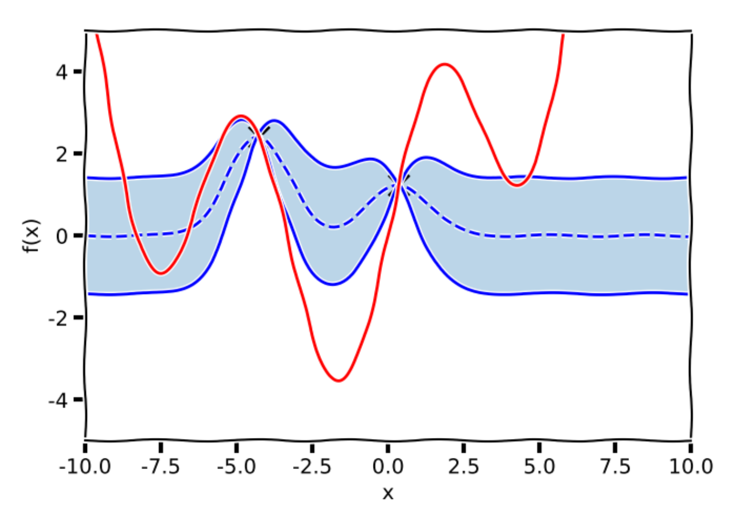
# Gaussian Processes



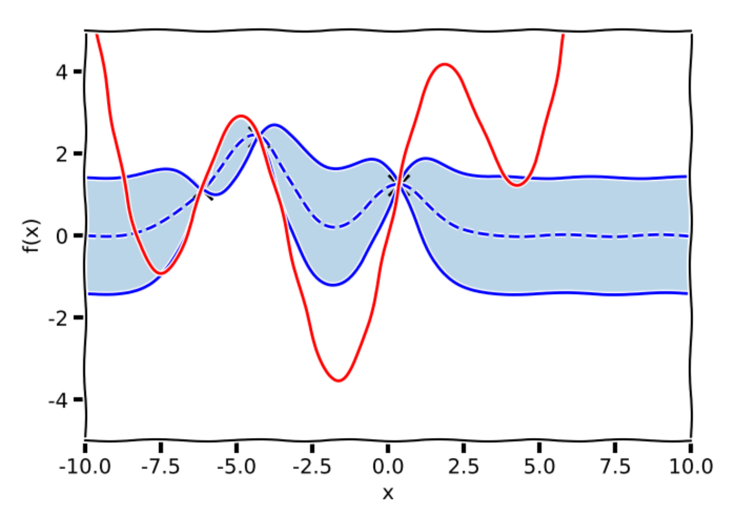
# Gaussian Processes



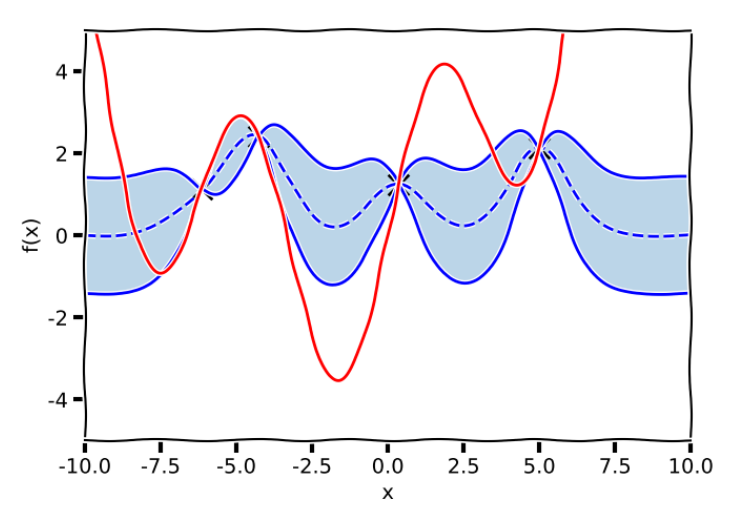
# Gaussian Processes



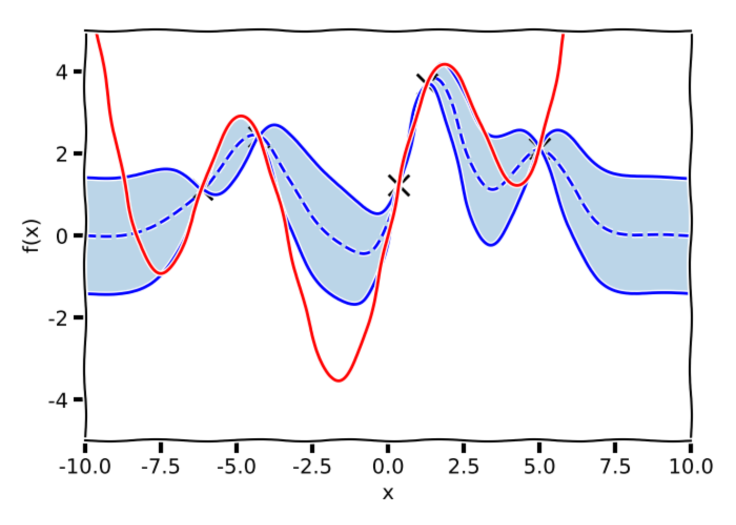
# Gaussian Processes

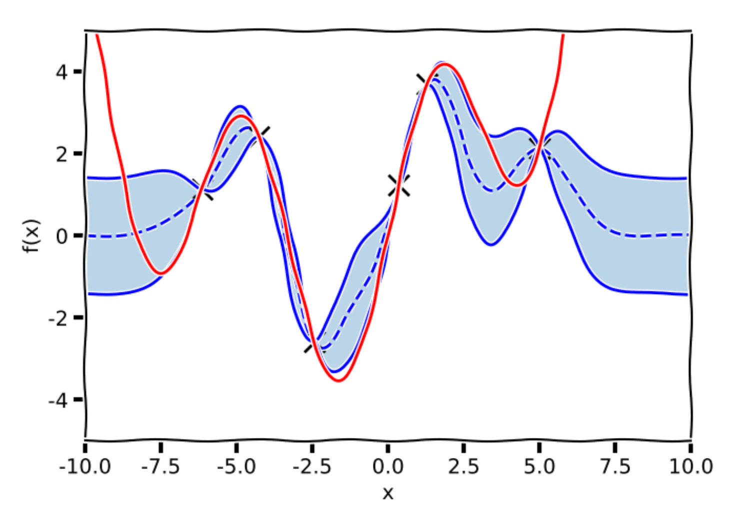


# Gaussian Processes

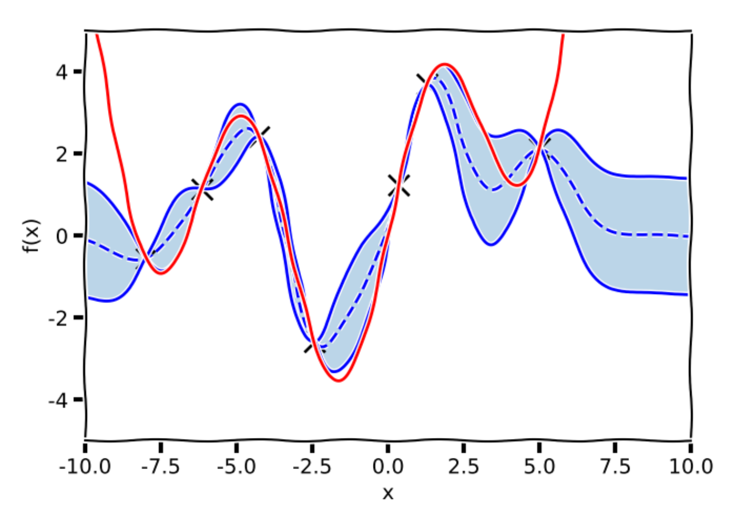


# Gaussian Processes

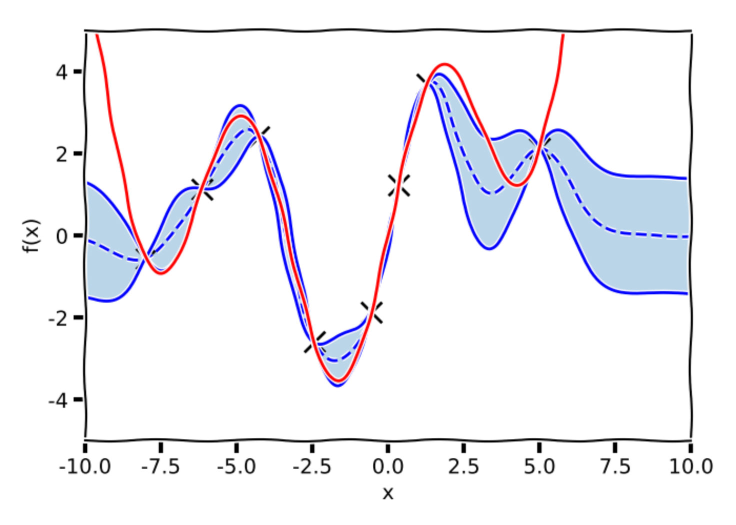



4
2
0
-2
-4
f(x)
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0
x

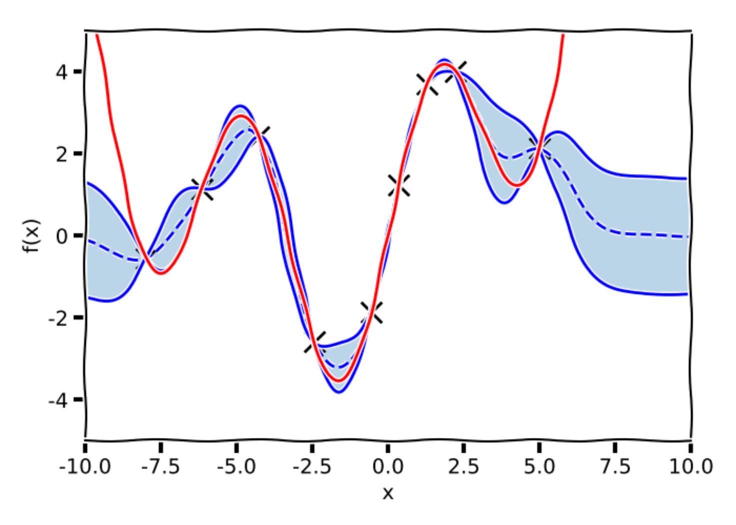
# Gaussian Processes



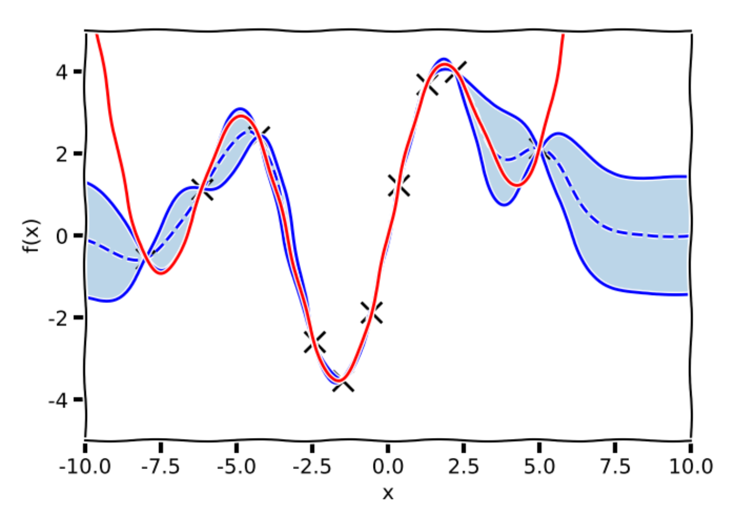
# Gaussian Processes



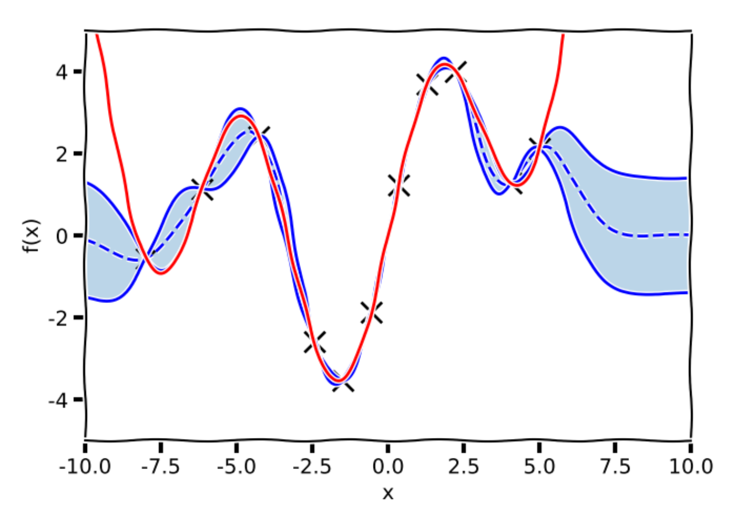
# Gaussian Processes



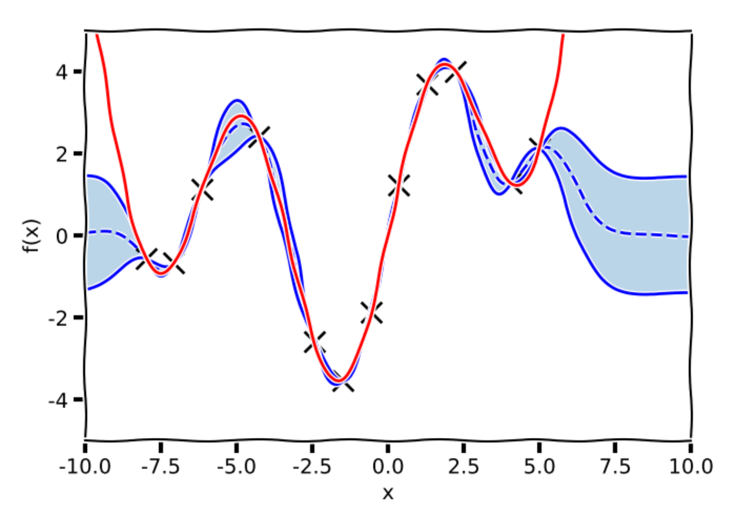
# Gaussian Processes



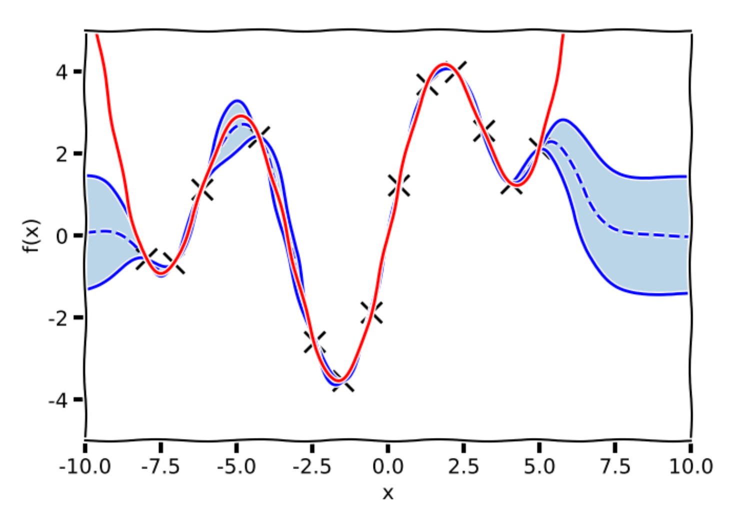
# Gaussian Processes



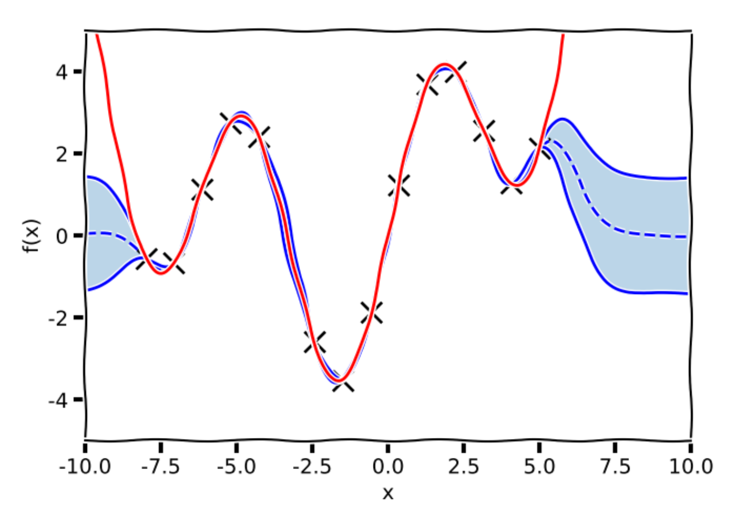
# Gaussian Processes



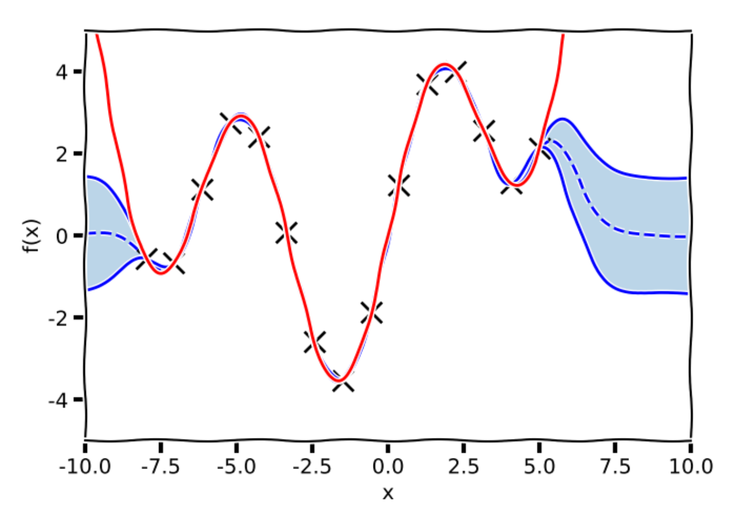
# Gaussian Processes



# Gaussian Processes



# Gaussian Processes



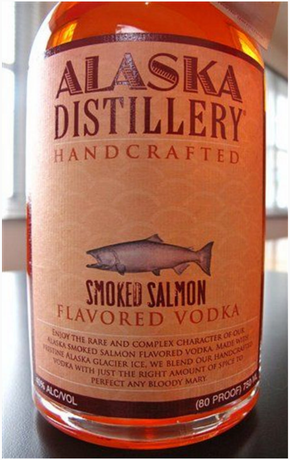
# Bayesian Optimisation

1. Choose a prior over the space of possible objective functions $f$
2. Combine prior and likelihood to get a posterior over the space
3. Use posterior to choose a set of evaluation according to a strategy
4. Add new data and update posterior
5. Repeate until budget is gone

# Exploration vs Exploitation

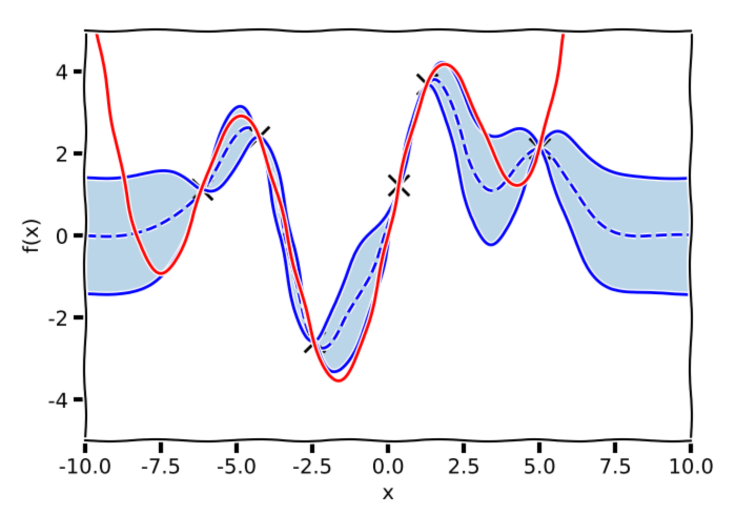
## Exploration vs Exploitation


# Balance is the key

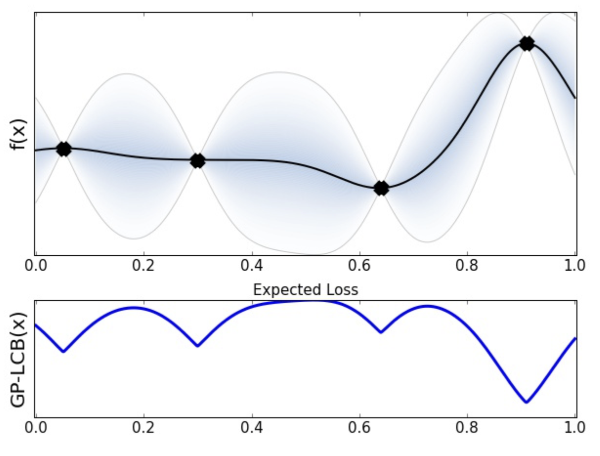


# Aquisition function

- how should we explore the space?
- humans do active search?
    - studying for exam

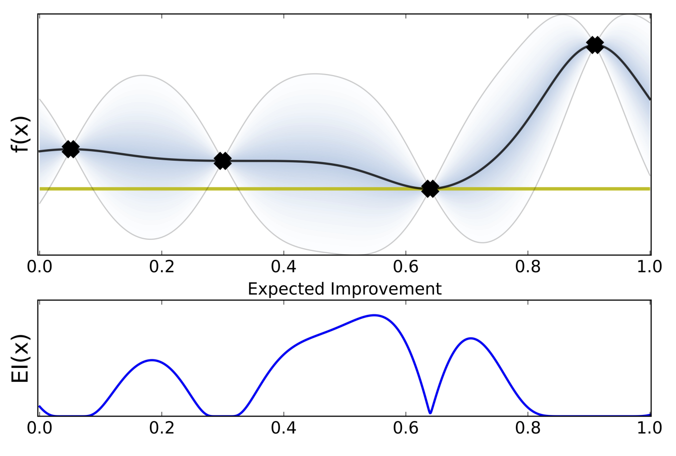

f(x)
4
2
0
-2
-4
-10.0
-7.5
-5.0
-2.5
0.0
2.5
5.0
7.5
10.0
x

# Confidence



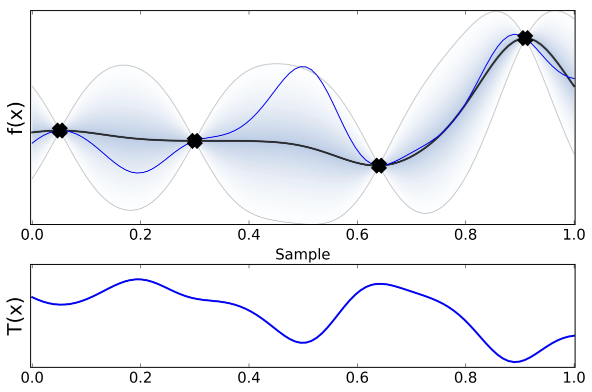

$$\alpha(\mathsf{x}; \theta, \mathcal{D}) = -\mu(\mathsf{x}; \theta, \mathcal{D}) + \beta_t \sigma(\mathsf{x}; \theta, \mathcal{D})$$

# Expected Improvement




$$\alpha(\mathbf{x}; \theta, \mathcal{D}) = \int \max(0, y_{\text{best}} - y) p(y|\mathbf{x}, \theta, \mathcal{D}) \mathrm{d}y$$

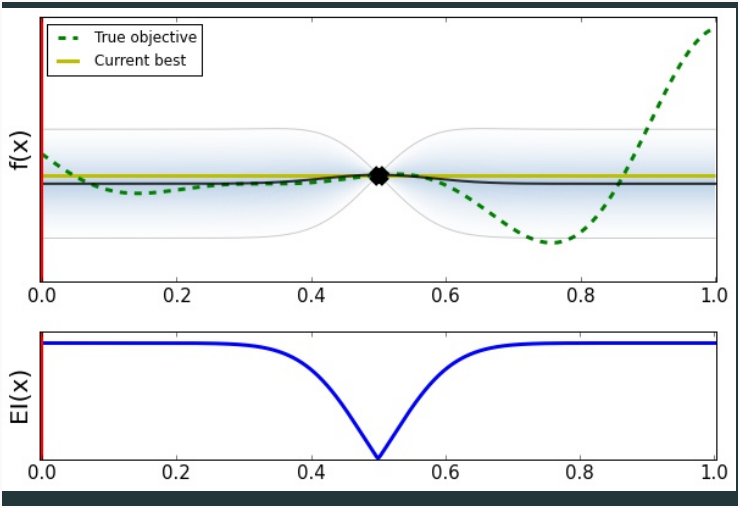
# Thomson Sampling




$$\alpha(\mathbf{x}; \theta, \mathcal{D}) = g(\mathbf{x})$$

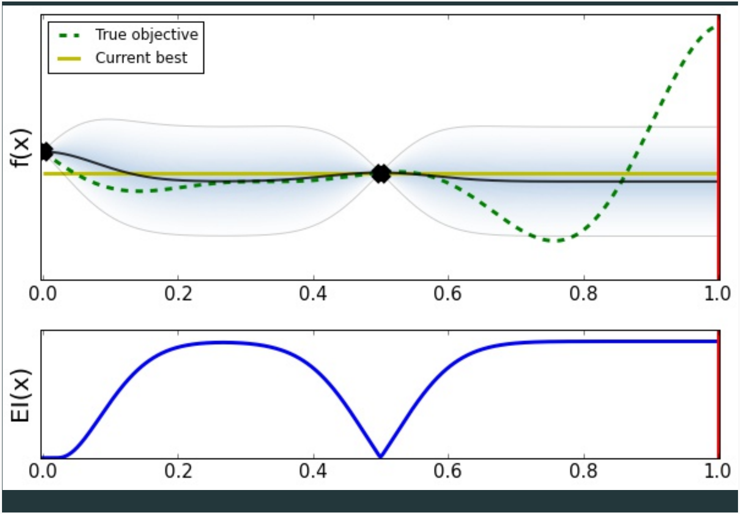
$$g(\mathbf{x}) \sim p(y|\mathbf{x}, \theta, \mathcal{D})$$

# Bayesian Optimisation [2]



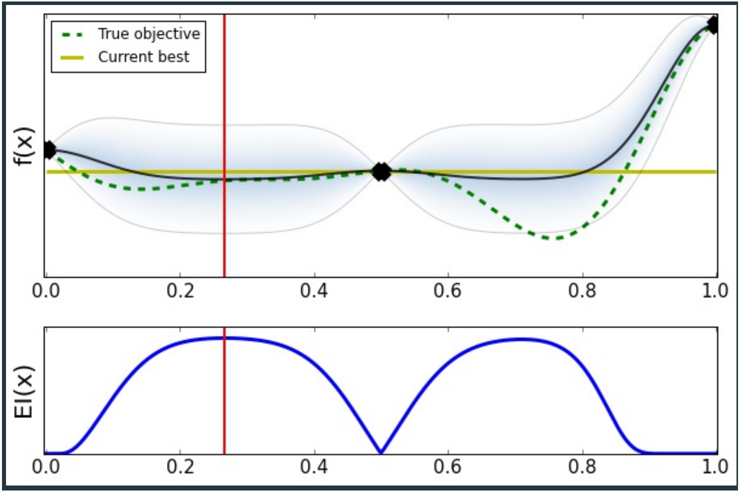

[2]Slides courtesy of Javier Gonzales

# Bayesian Optimisation [2]



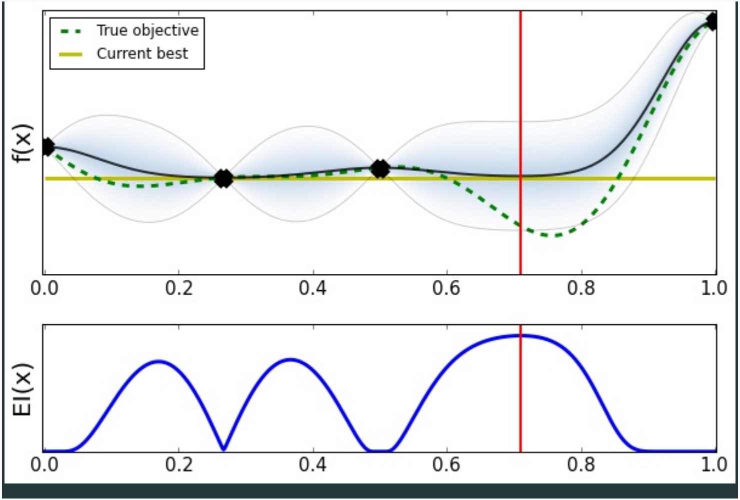

[2] Slides courtesy of Javier Gonzales

# Bayesian Optimisation [2]



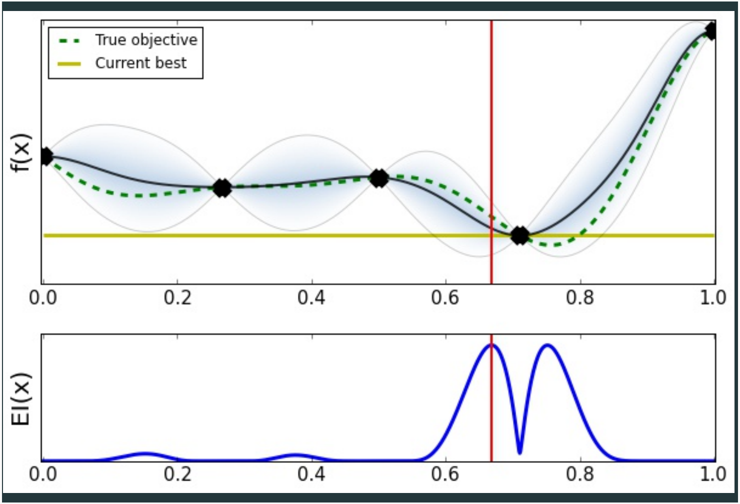

[2] Slides courtesy of Javier Gonzales

# Bayesian Optimisation [2]



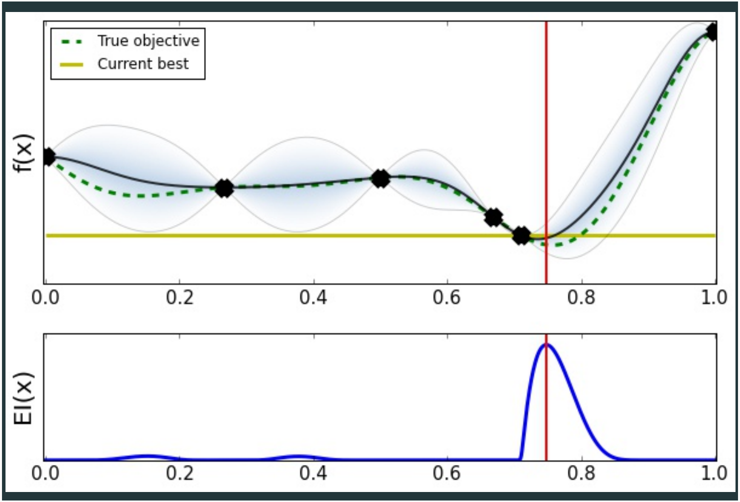

[2] Slides courtesy of Javier Gonzales

# Bayesian Optimisation [2]



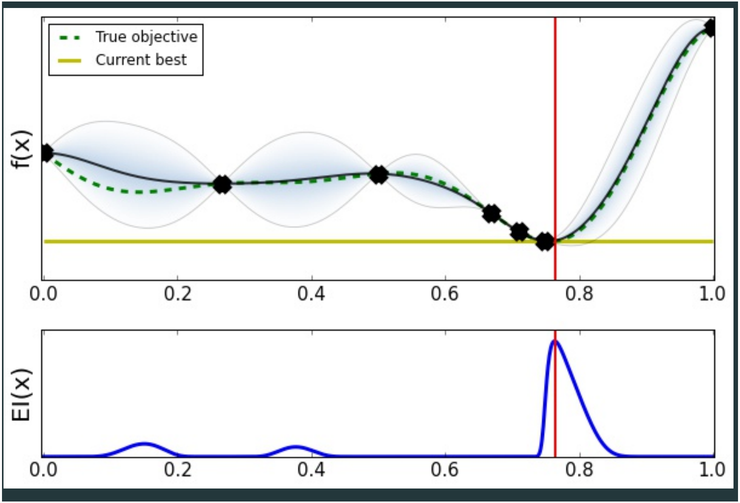

[2] Slides courtesy of Javier Gonzales

# Bayesian Optimisation [2]



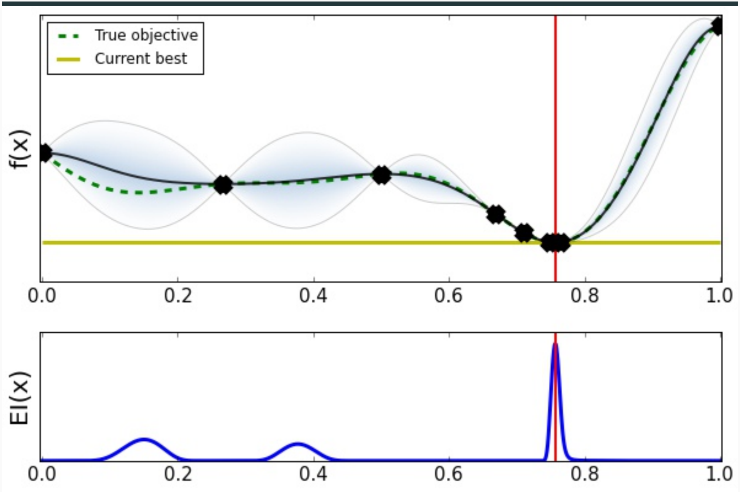

[2] Slides courtesy of Javier Gonzales

# Bayesian Optimisation [2]




[2] Slides courtesy of Javier Gonzales

# Bayesian Optimisation [2]




[2] Slides courtesy of Javier Gonzales

- We cannot solve the direct problem

$$x_M = \mathrm{argmin}_{x \in \mathcal{X}} f(x)$$

## Bayesian Optmisation

- We cannot solve the direct problem

$$x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$$

- Transform to a series of simpler problems

$$x_{n+1} = \text{argmin}_{x \in \mathcal{X}} \alpha(x; \mathcal{D}_n, \mathcal{M}_n)$$

# Bayesian Optmisation

- We cannot solve the direct problem

$$x_M = \text{argmin}_{x \in \mathcal{X}} f(x)$$

- Transform to a series of simpler problems

$$x_{n+1} = \text{argmin}_{x \in \mathcal{X}} \alpha(x; \mathcal{D}_n, \mathcal{M}_n)$$

- this will work well if

# Bayesian Optmisation

- We cannot solve the direct problem

$$x_M = \mathrm{argmin}_{x \in \mathcal{X}} f(x)$$

- Transform to a series of simpler problems

$$x_{n+1} = \mathrm{argmin}_{x \in \mathcal{X}} \alpha(x; \mathcal{D}_n, \mathcal{M}_n)$$

- this will work well if
    - $\alpha(x)$ is cheap to compute

# Bayesian Optmisation

- What assumptions can we make about the function that we are optimising?
  - kernel
  - kernel parameters
  - etc.
- What search strategy should we have
  - exploration vs. exploitation
  - cost of search
- How should we optimise the aquisition function

# Why [3]

| Consumption | $CO_2$ (lbs) |
|---|---:|
| Air travel NY-SF | 1 984 |
| Human life (1 year) | 11 023 |
| American life (1 year) | 36 156 |
| Car life-time | 126 000 |
| Transformers NN | 626 155 |

[3] Strubell, E., Ganesh, A., & McCallum, A., Energy and policy considerations for deep learning in nlp, CoRR, (), (2019).

# Open Questions

- Parallell BO

# Open Questions

- Parallell BO
- Traditional BO is myopic but most tasks are not

# Open Questions

- Parallell BO
- Traditional BO is myopic but most tasks are not
- Non-stationary functions

# Open Questions

- Parallell BO
- Traditional BO is myopic but most tasks are not
- Non-stationary functions
- Multiple objectives

# Open Questions

- Parallell BO
- Traditional BO is myopic but most tasks are not
- Non-stationary functions
- Multiple objectives
- Constraints

# Summary

# Summary

- BO is a rapidly evolving field

## Summary

- BO is a rapidly evolving field
- it is happening now, its the thing that you want on your CV ;-)

## Summary

- BO is a rapidly evolving field
- it is happening now, its the thing that you want on your CV ;-)
- Two key aspects,

# Summary

- BO is a rapidly evolving field
- it is happening now, its the thing that you want on your CV ;-)
- Two key aspects,
    - Gaussian processes to model function

## Summary

- BO is a rapidly evolving field
- it is happening now, its the thing that you want on your CV ;-)
- Two key aspects,
    - Gaussian processes to model function
    - Aquisition function to model exploration

## Summary

- BO is a rapidly evolving field
- it is happening now, its the thing that you want on your CV ;-)
- Two key aspects,
    - Gaussian processes to model function
    - Aquisition function to model exploration
- Take home message: *uncertainty matters*

eof

# References

Christopher M. Bishop.
***Pattern Recognition and Machine Learning (Information Science and Statistics)*.**
Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.