University of BRISTOL

Machine Learning

Bayesian Optimisation

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

October 17, 2017

http://www.carlhenrik.com

Introduction

Help Sessions

- Thursday 19/10 18-20 MVB 1.11
- Friday 20/10 17-19 MVB 1.11
- Potentially Thursday 26/10 17-19 (will confirm next week)

GitHub

`git pull`

Recap

Marginal Likelihood

$$p(\mathbf{Y}|\mathbf{X}, \theta) = \int p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta) \mathrm{d}f$$

- We are not interested in f directly
- Marginalise out f
- Gaussian likelihood and Gaussian prior $\rightarrow$ Gaussian marginal

Marginalisation

- Deterministic world

$$\mathbb{E}[y] = \int y p(y) \mathrm{d}y$$

Marginalisation

- Deterministic world

$$\mathbb{E}[y] = \int y p(y) \mathrm{d}y$$

- Stochastic world

$$\mathbb{E}[p(y)] = \int p(y|x) p(x) \mathrm{d}x$$

Learning

$$\hat{\theta} = \text{argmax}_\theta p(\mathbf{Y}|\mathbf{X}, \theta) = \text{argmax}_\theta \int p(\mathbf{Y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \theta) \mathrm{d}\mathbf{f}$$

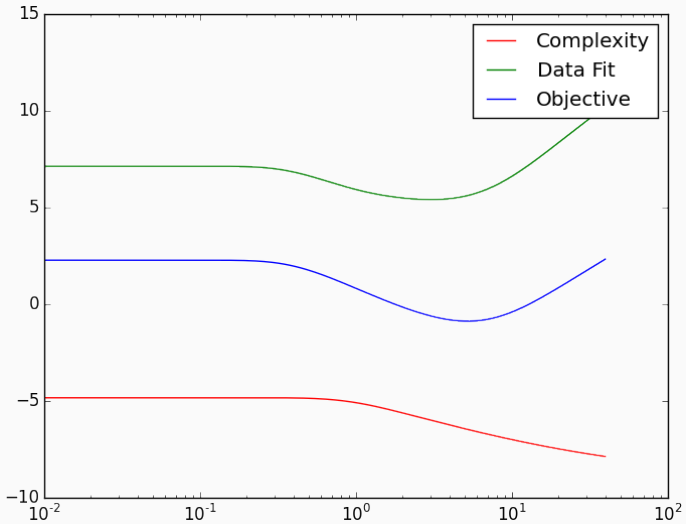
- Type-II Maximum likelihood [1] 3.5.0
- minimise logarithm of marginal likelihood

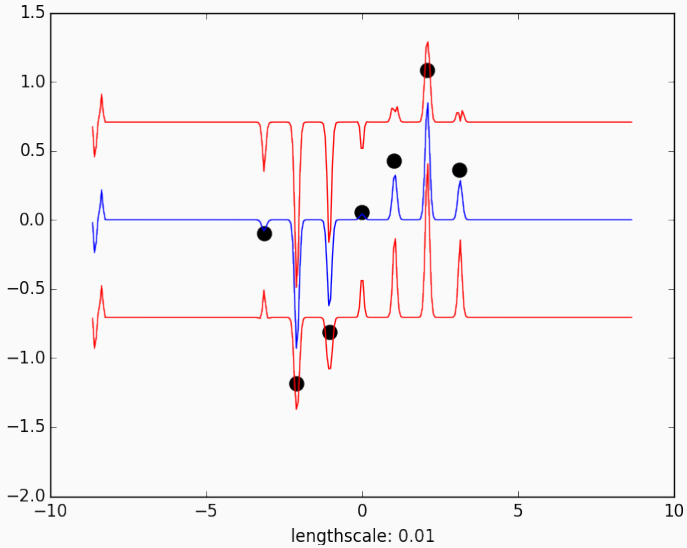
Learning

$$\text{argmax}_{\boldsymbol{\theta}} p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta}) = \text{argmin}_{\boldsymbol{\theta}} - \log\left(p(\mathbf{Y}|\mathbf{X}, \boldsymbol{\theta})\right) = \text{argmin}_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$

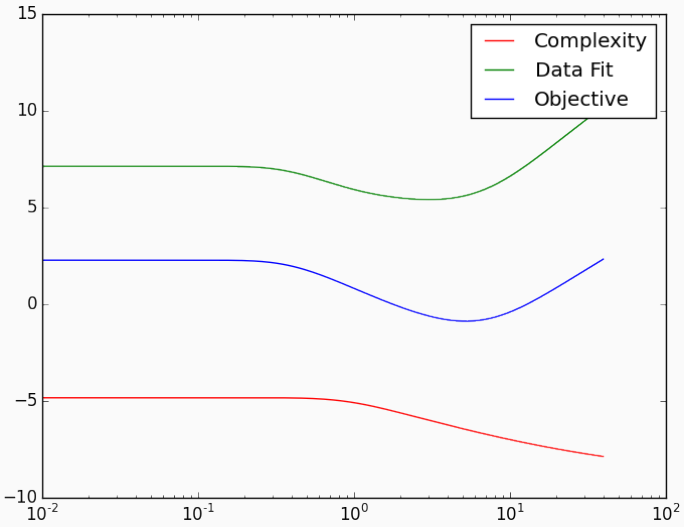
$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{y} + \frac{1}{2}\log|\mathbf{K}| + \frac{N}{2}\log(2\pi)$$

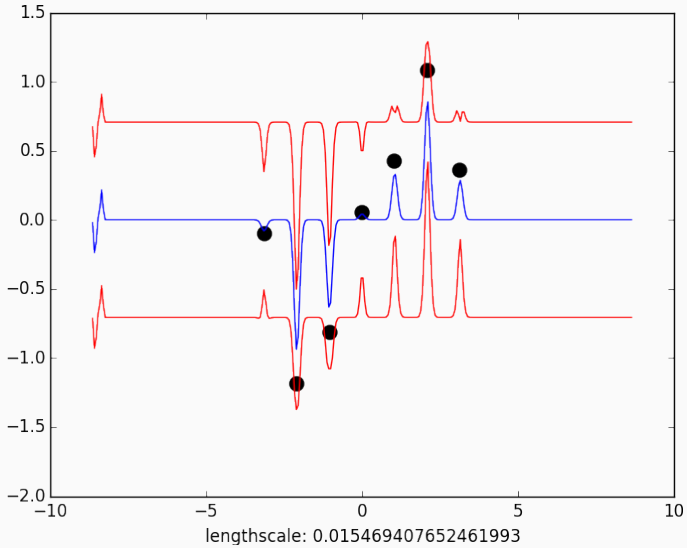
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

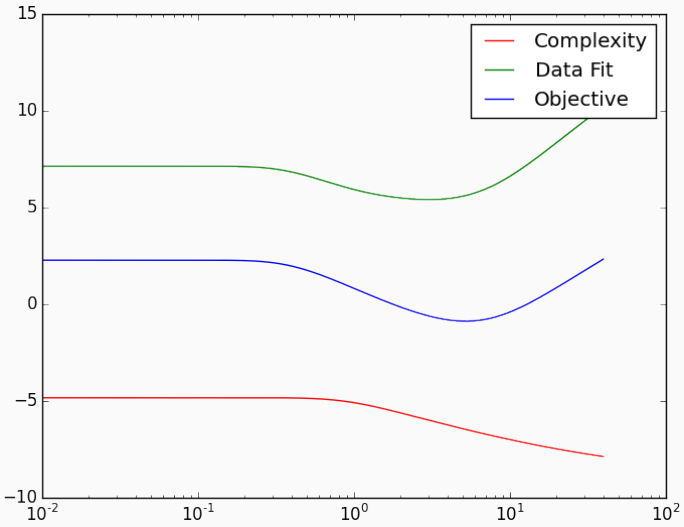

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.01

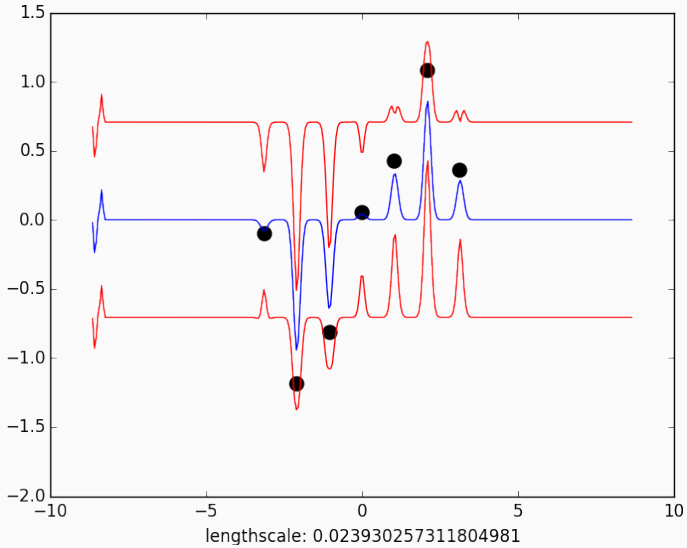
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

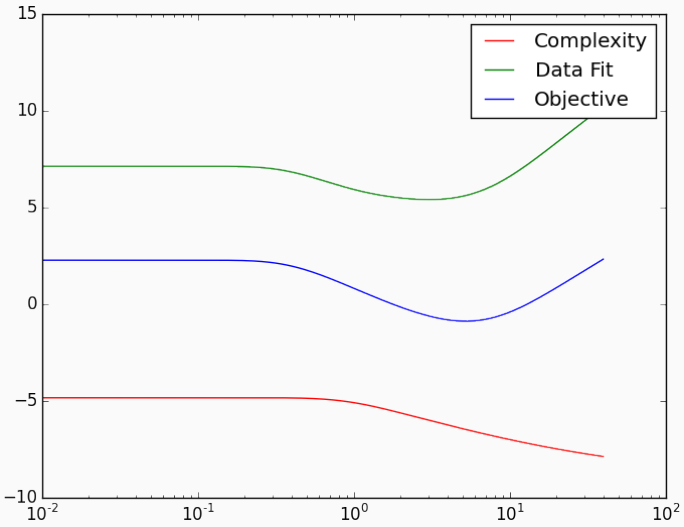

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.015469407652461993

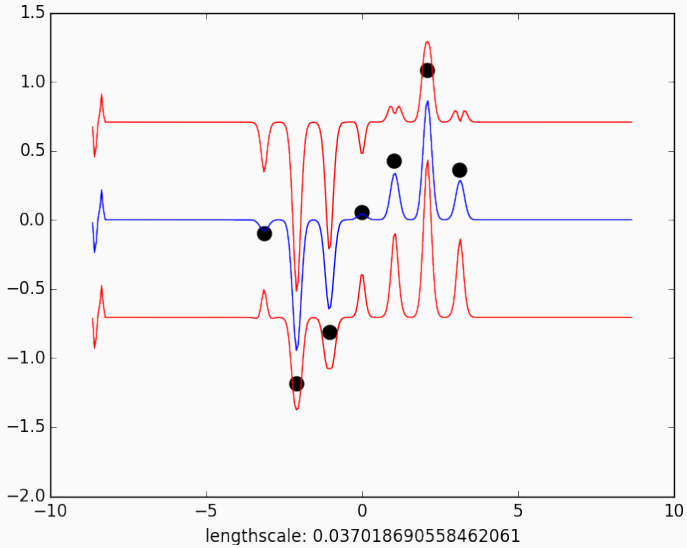
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

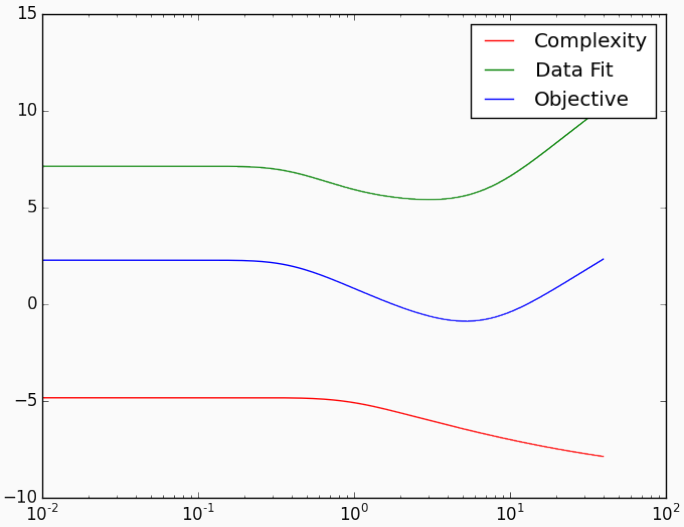

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.0239302573118049 81

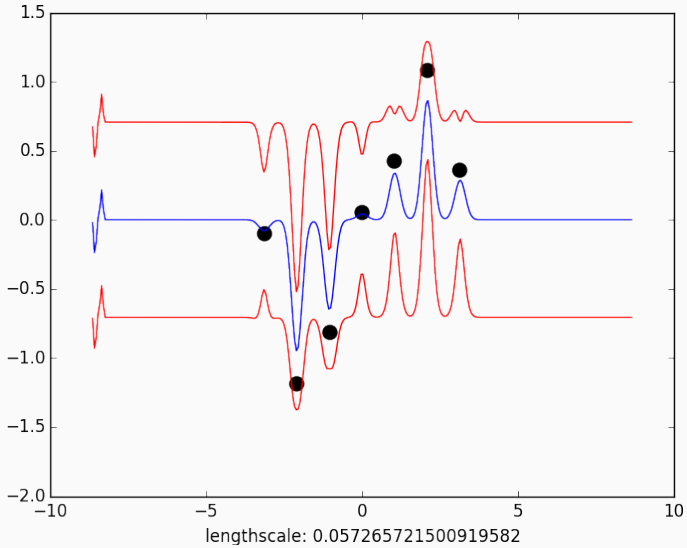
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

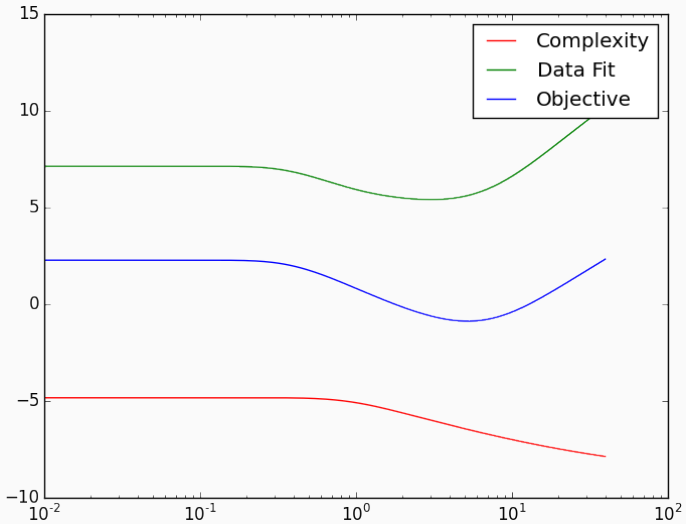

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.037018690558462061

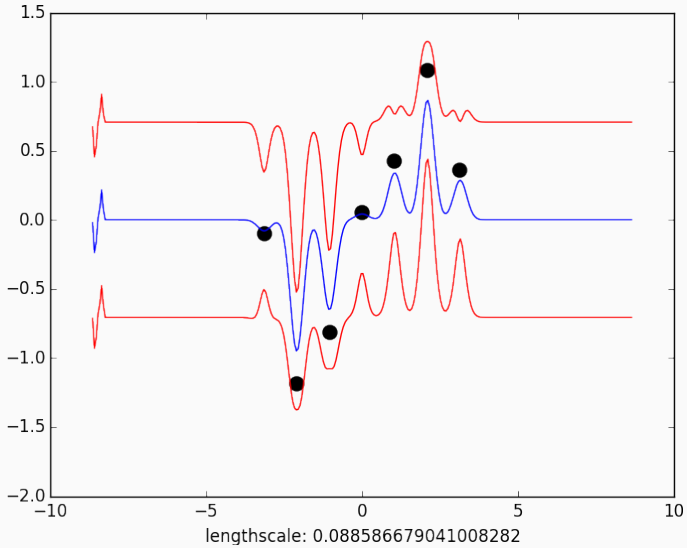
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

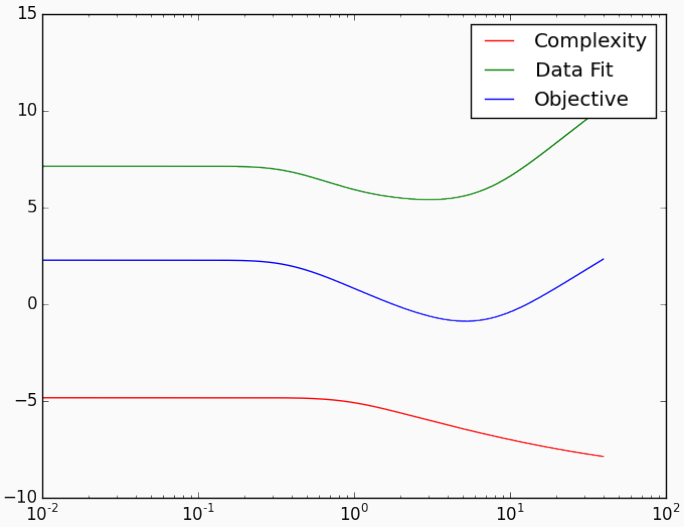

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.057265721500919582

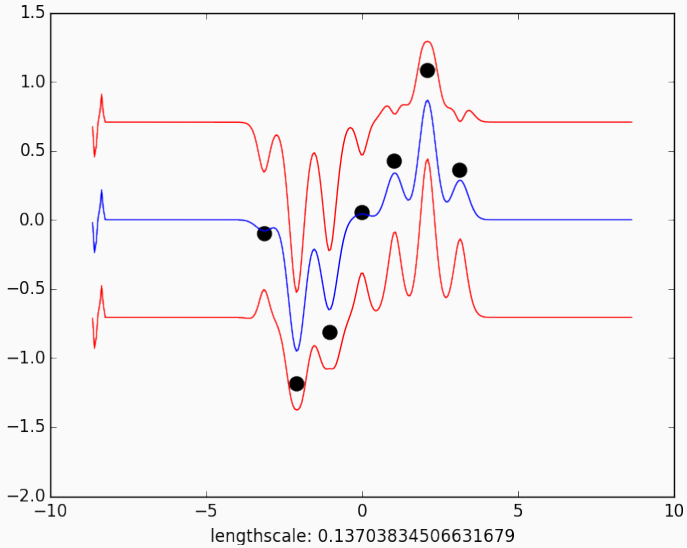
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

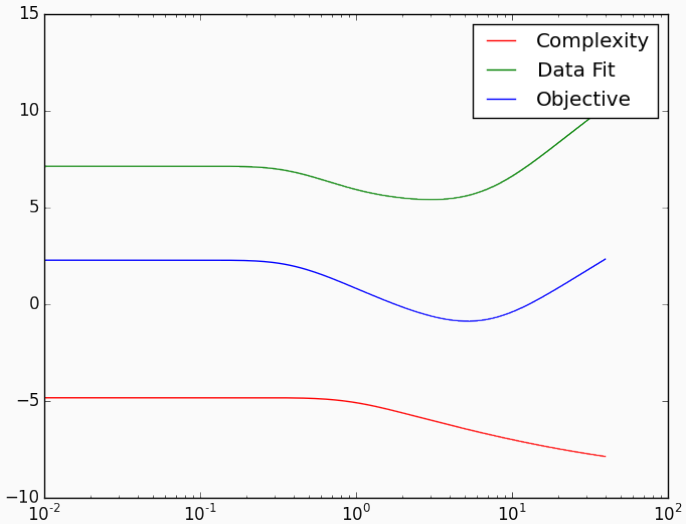

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.088586679041008282

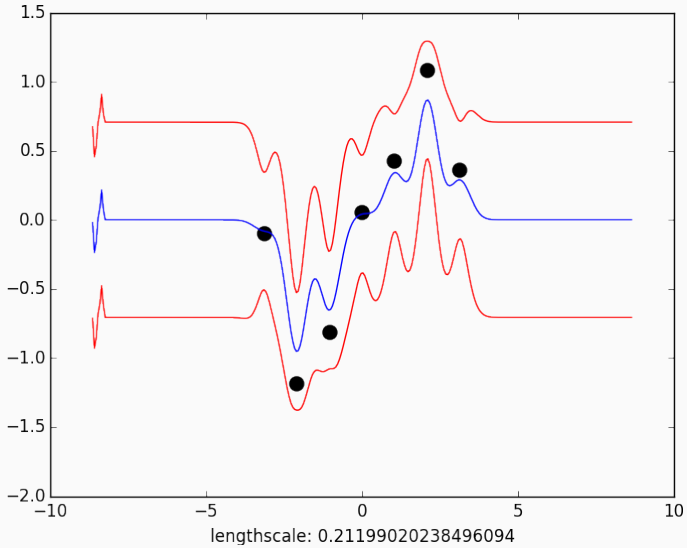
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

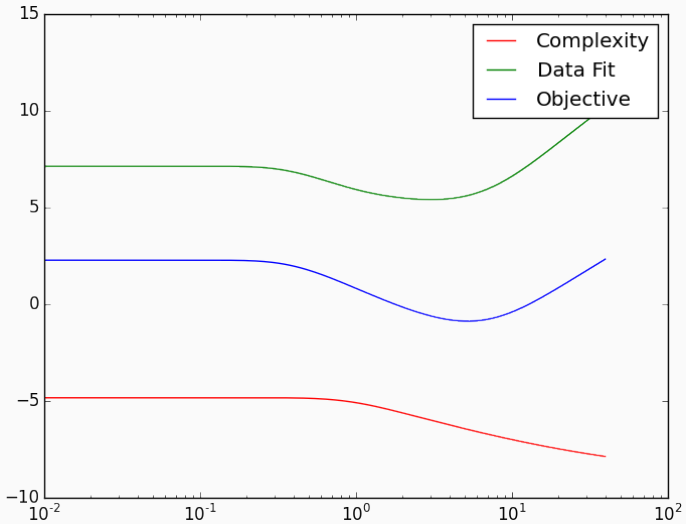

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.13703834506631679

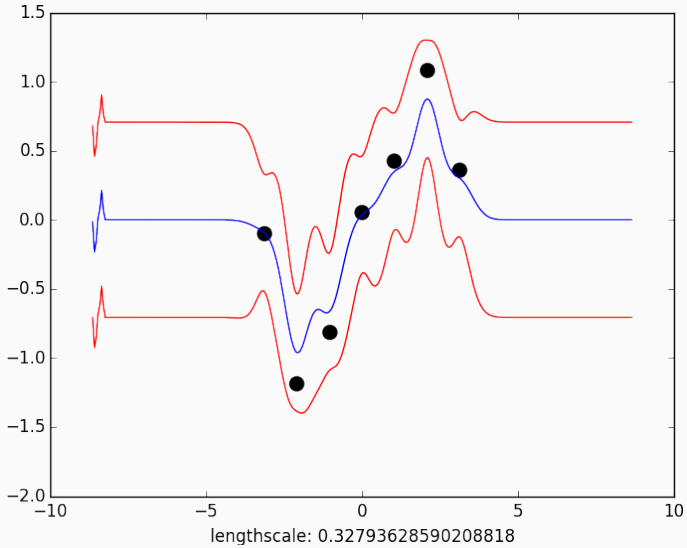
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

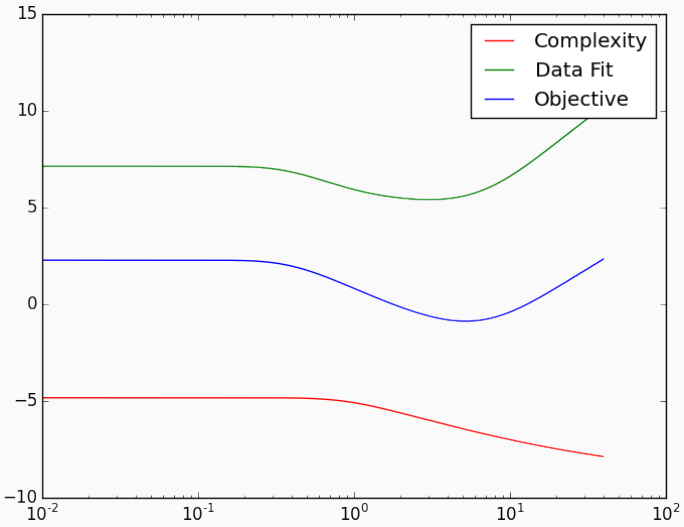

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.21199020238496094

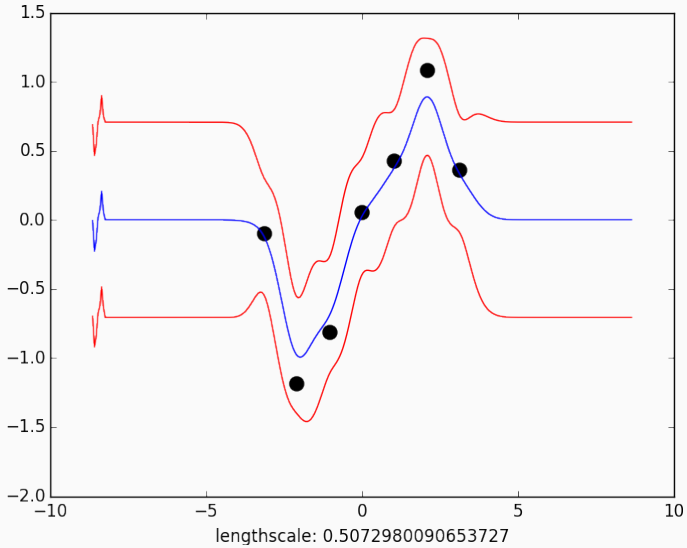
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

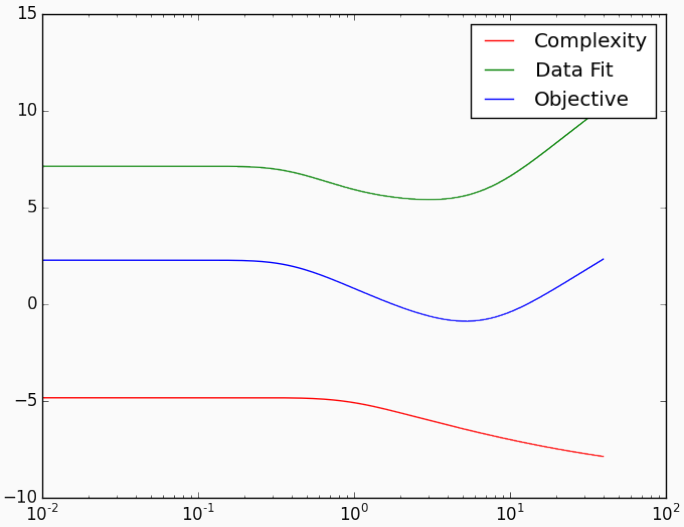

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.32793628590208818

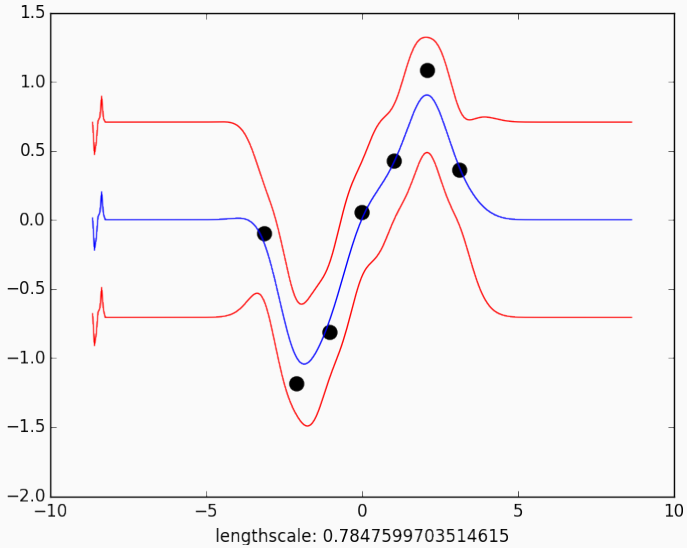
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

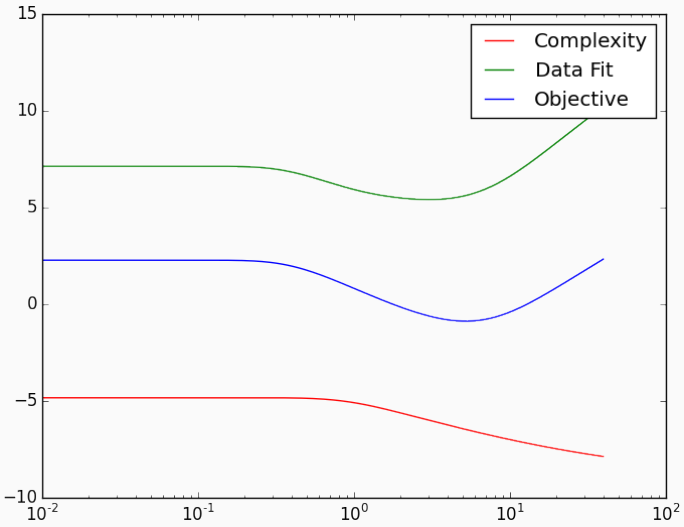

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.5072980090653727

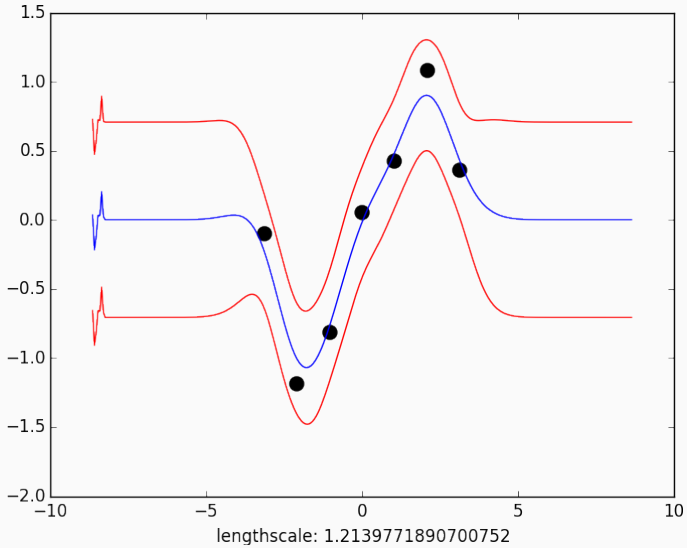
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

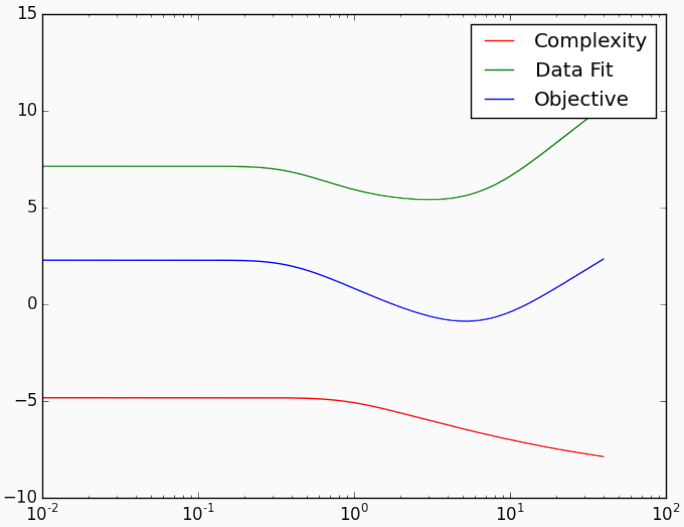

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 0.7847599703514615

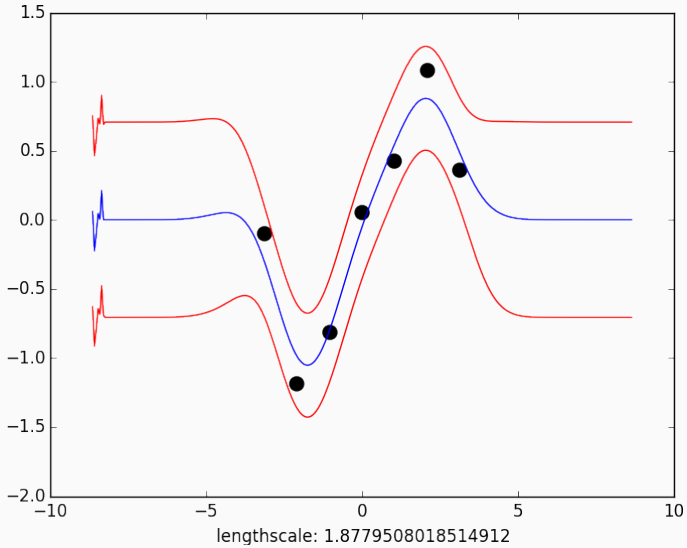
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

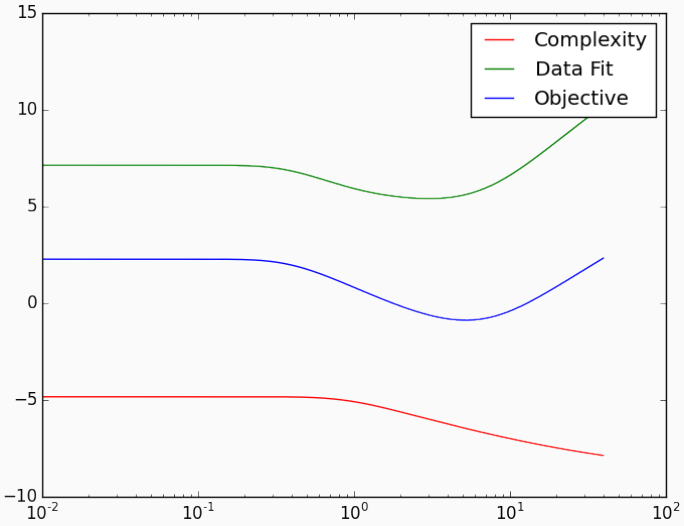

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 1.2139771890700752

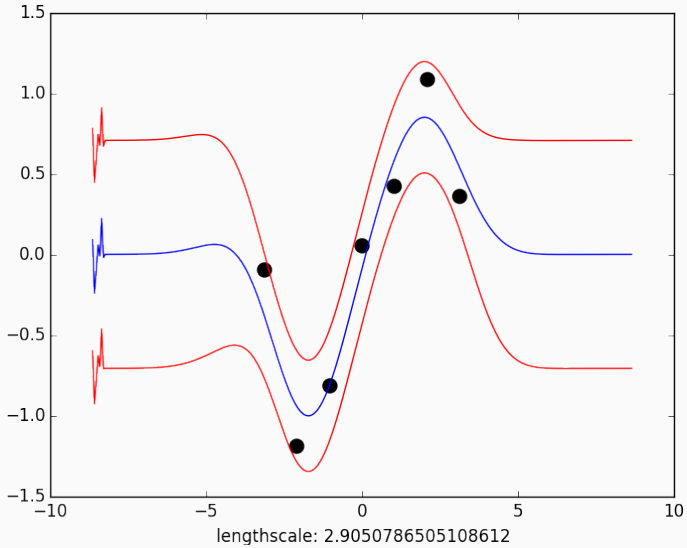
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

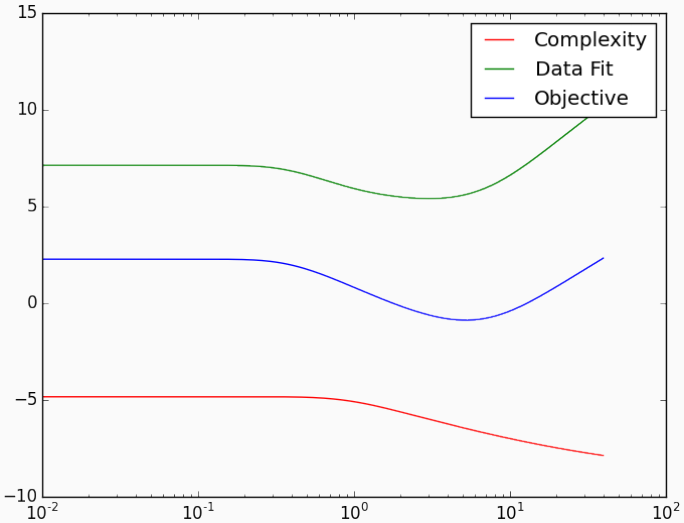

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−2.0
−10
−5
0
5
10
lengthscale: 1.8779508018514912

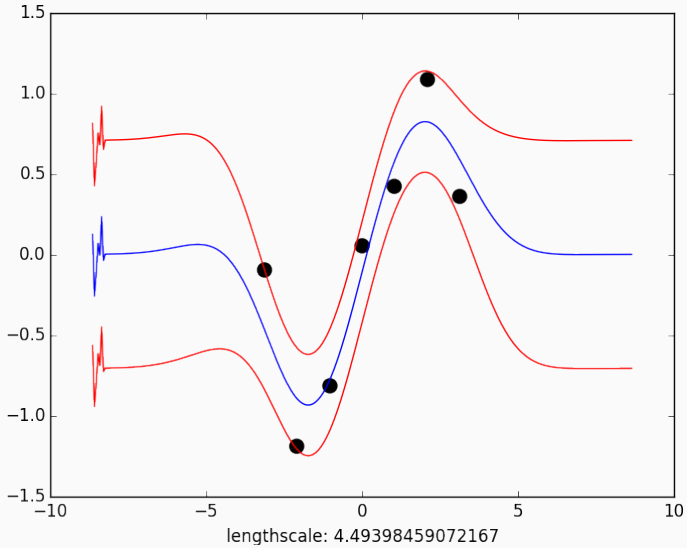
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

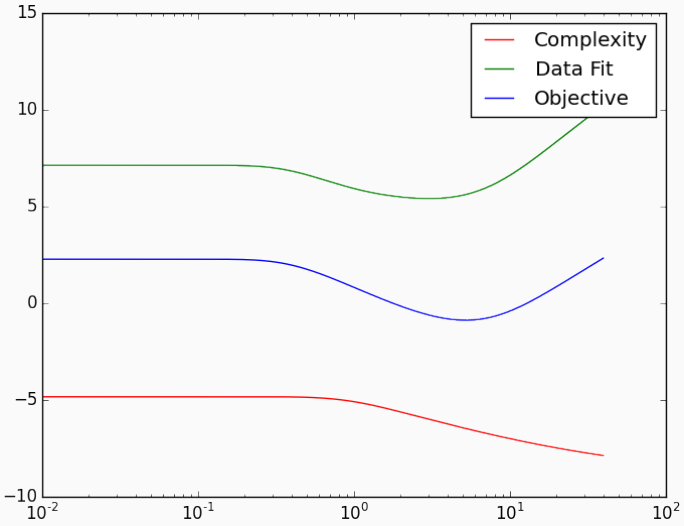

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 2.9050786505108612

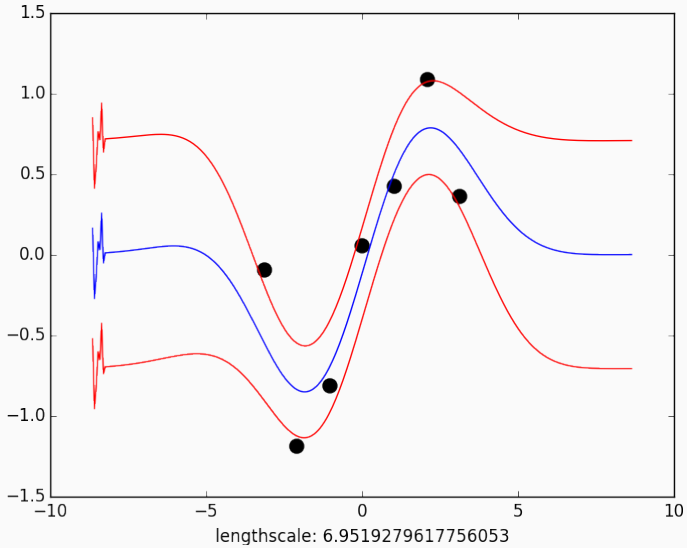
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

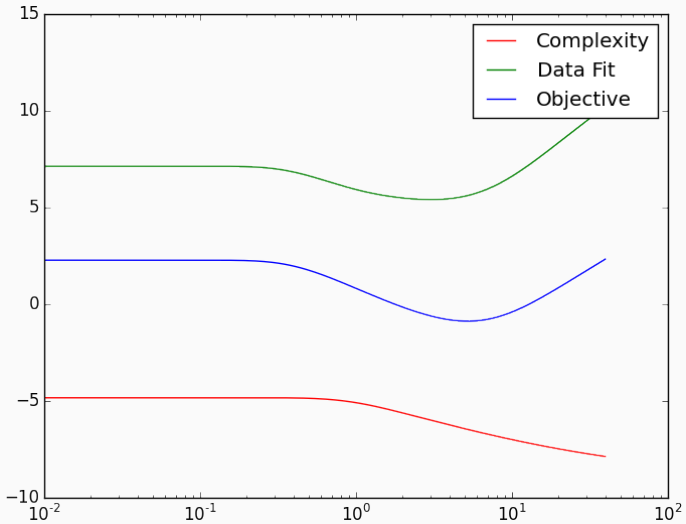

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 4.49398459072167

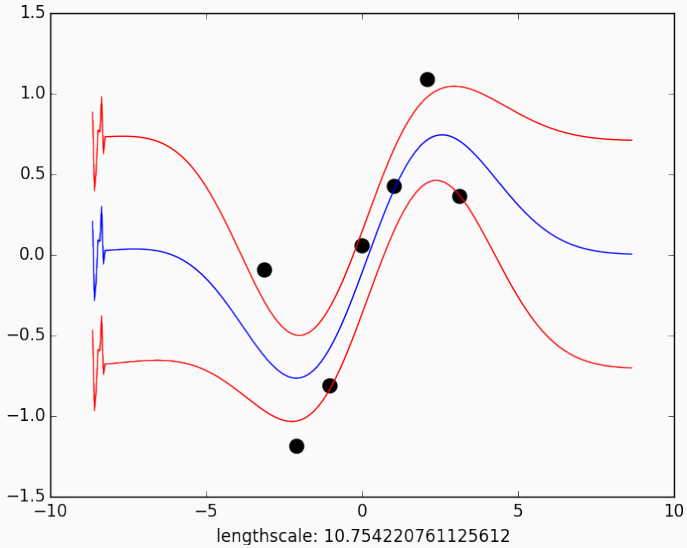
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

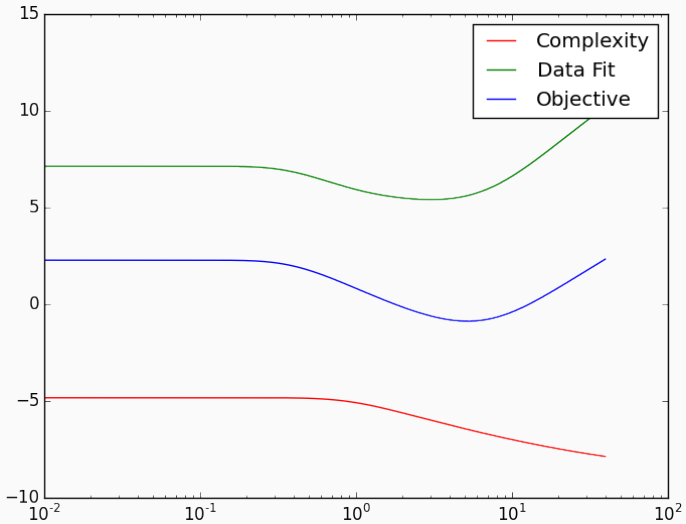

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 6.9519279617756053

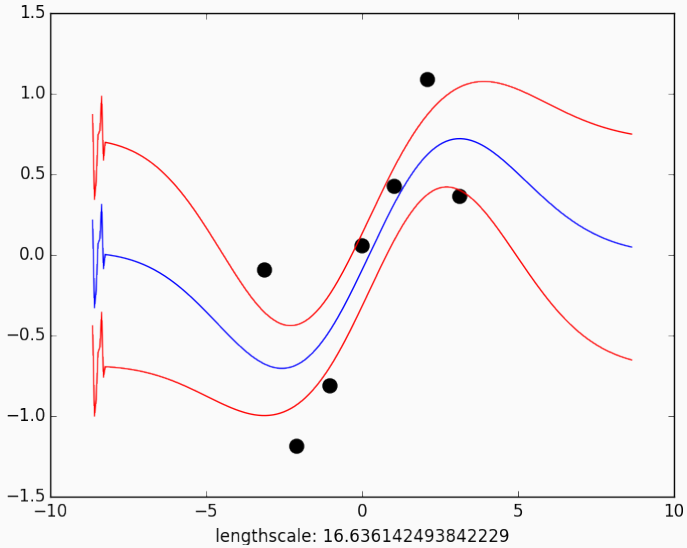
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

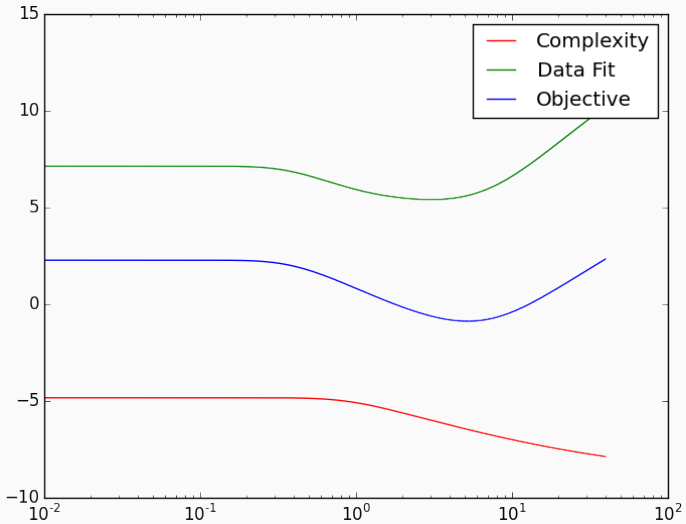

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 10.754220761125612

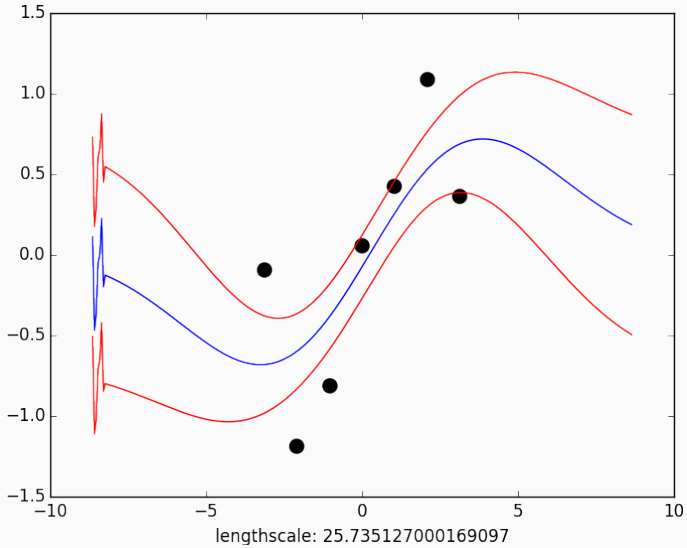
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

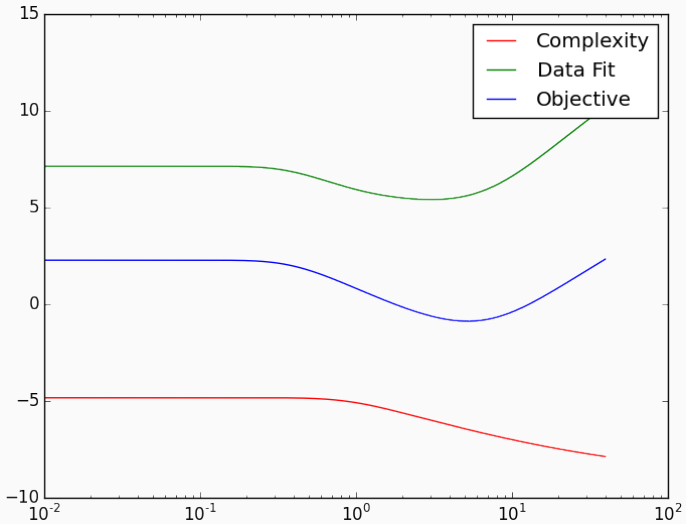

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 16.636142493842229

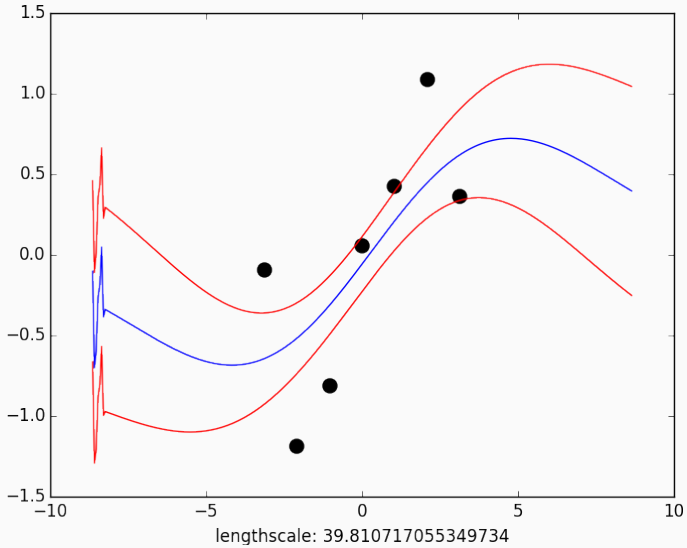
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}

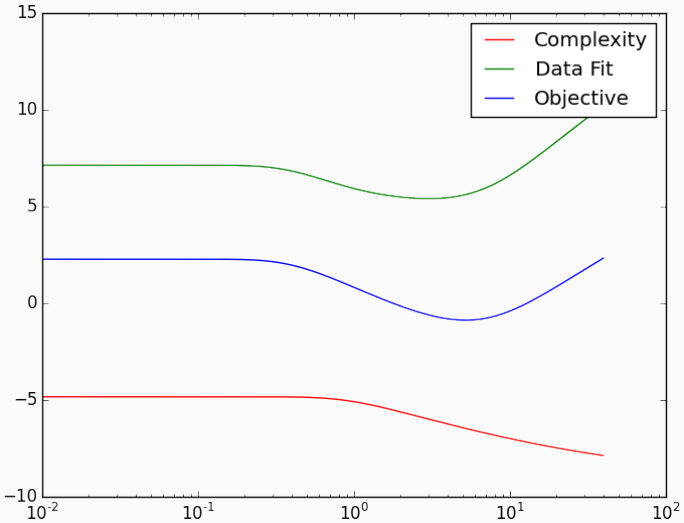

1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 25.735127000169097

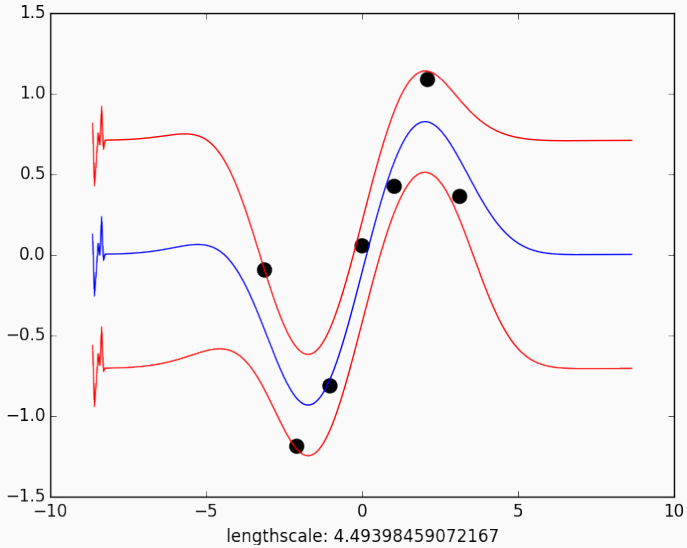
Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 39.810717055349734

Learning


Complexity
Data Fit
Objective
15
10
5
0
−5
−10
10^{-2}
10^{-1}
10^{0}
10^{1}
10^{2}


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−5
0
5
10
lengthscale: 4.49398459072167

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

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

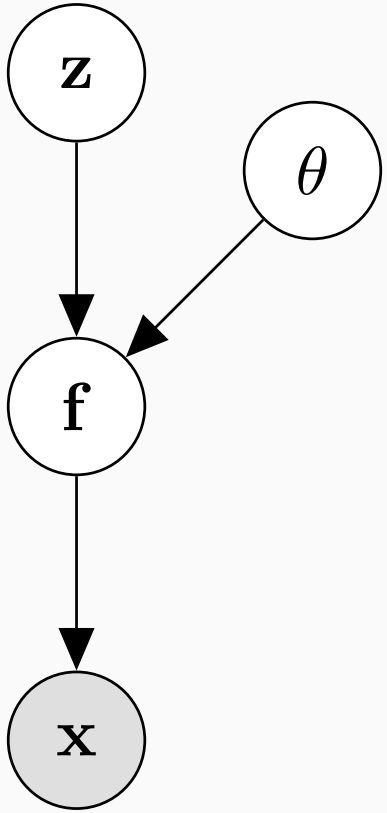
- Actually formulating the posterior over both $\mathbf{W}$ and $\mathbf{z}$ is intractable

Type II Maximum Likelihood

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

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

Non-linear Latent variable model



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

Demo

Font Demo

Bayesian Optimisation

Next lectures

Today Bayesian Optimisation

Monday Dirichlet Processes

Tuesday Topic Models

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.

Uncertainty

- Deterministic world

$$x = 4$$

- Point estimate world

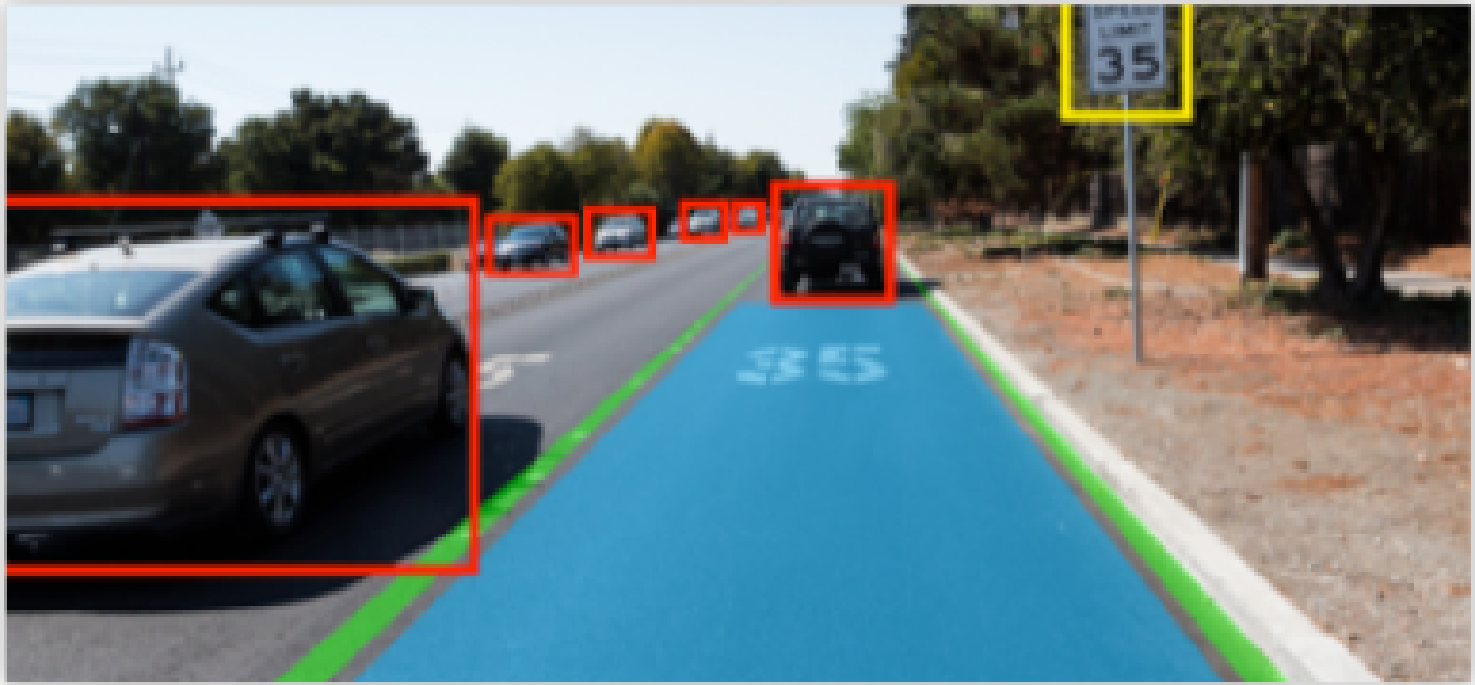
$$\text{argmax}_x p(x) = 4$$

- Stochastic world

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

Uncertainty

Uncertainty

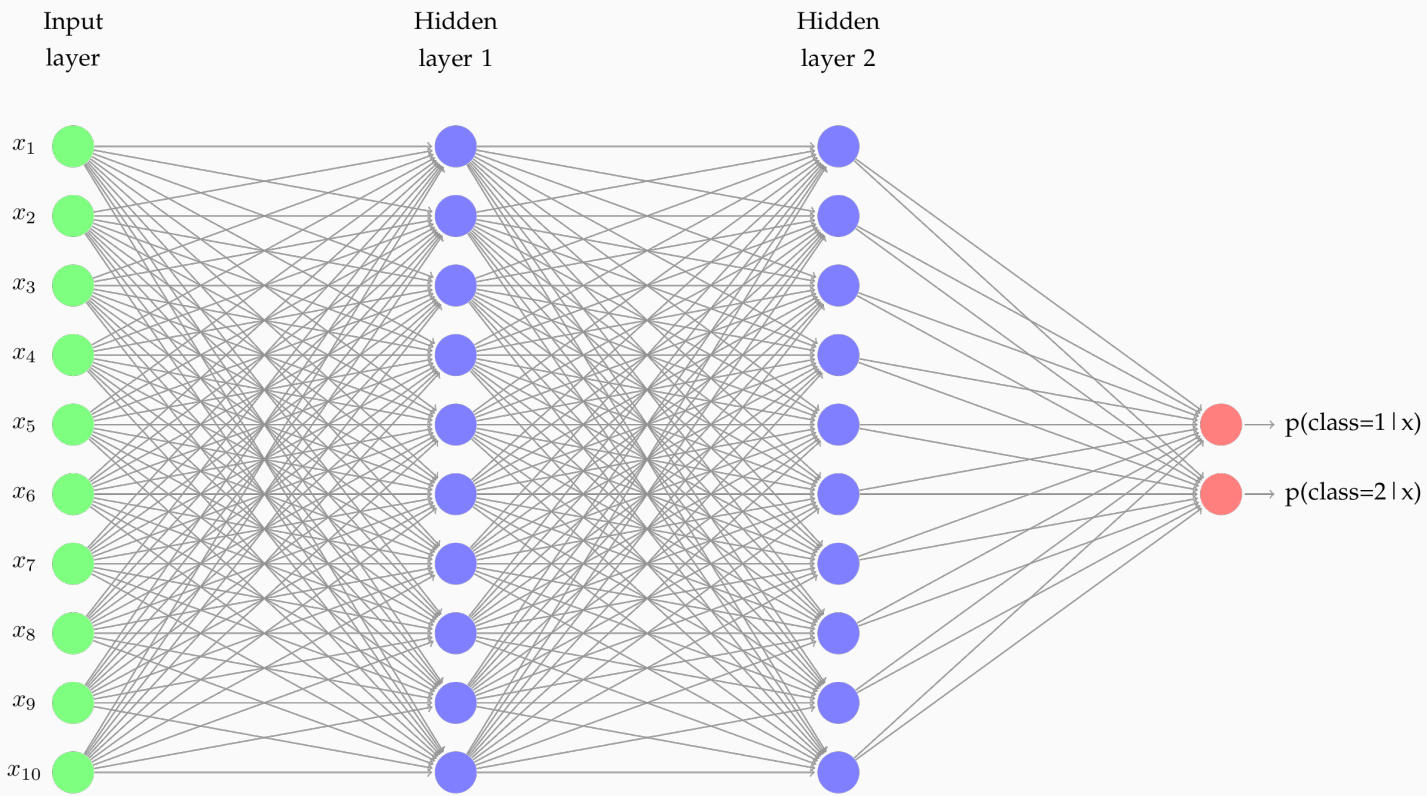


Real Doctors

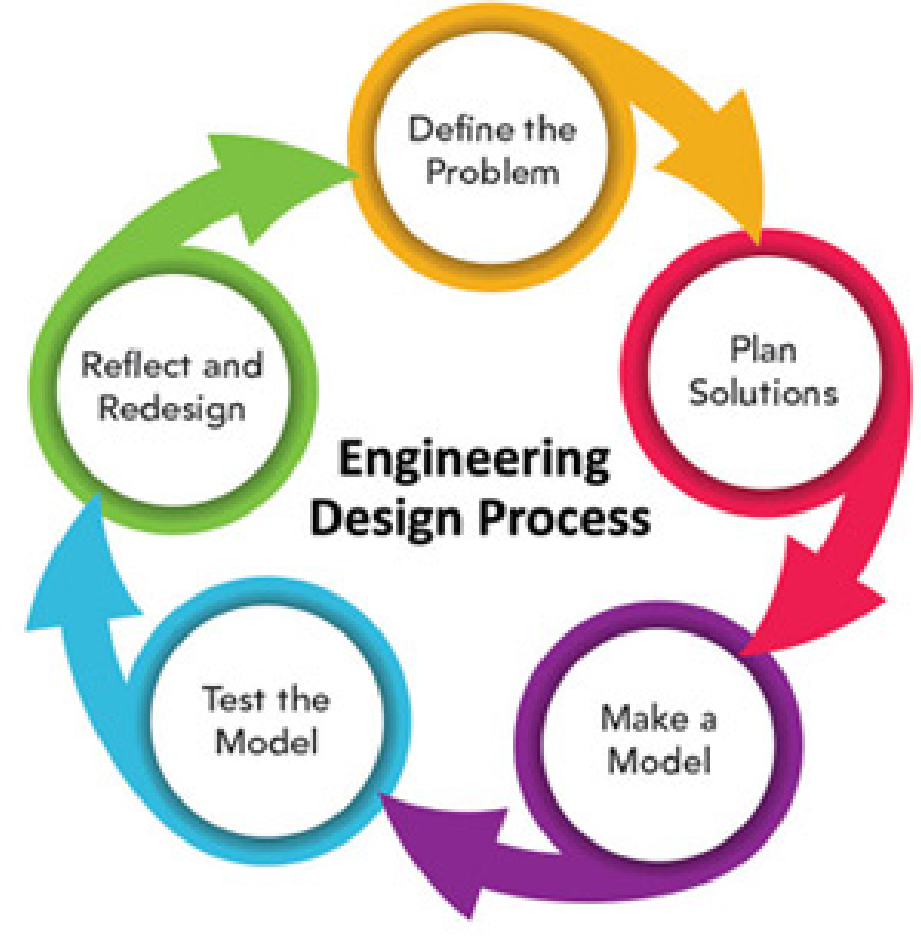
Real Doctors

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

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

- Much more common

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

Implicit understanding

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

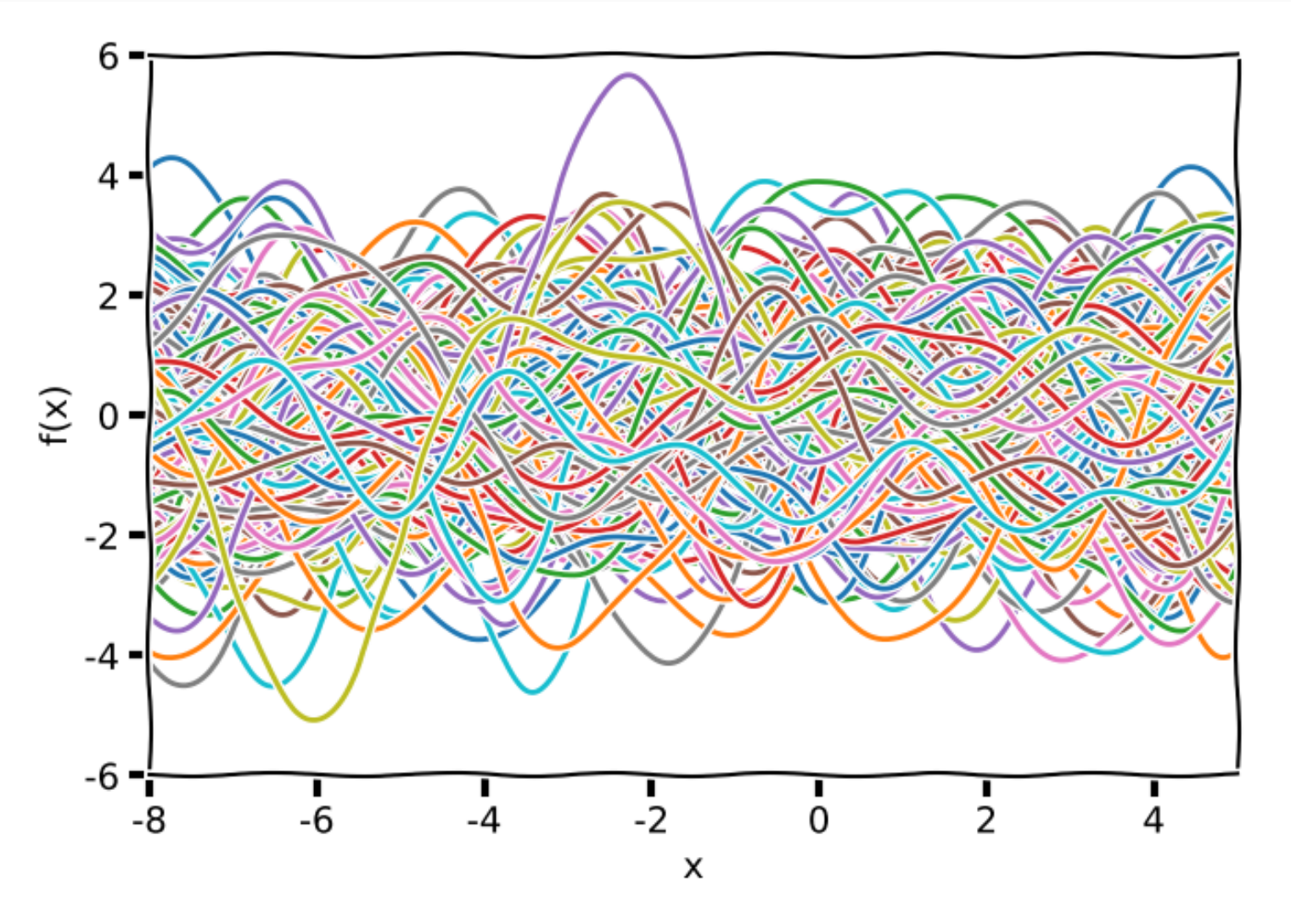
Strategies

- Implicit knowledge
- Grid search

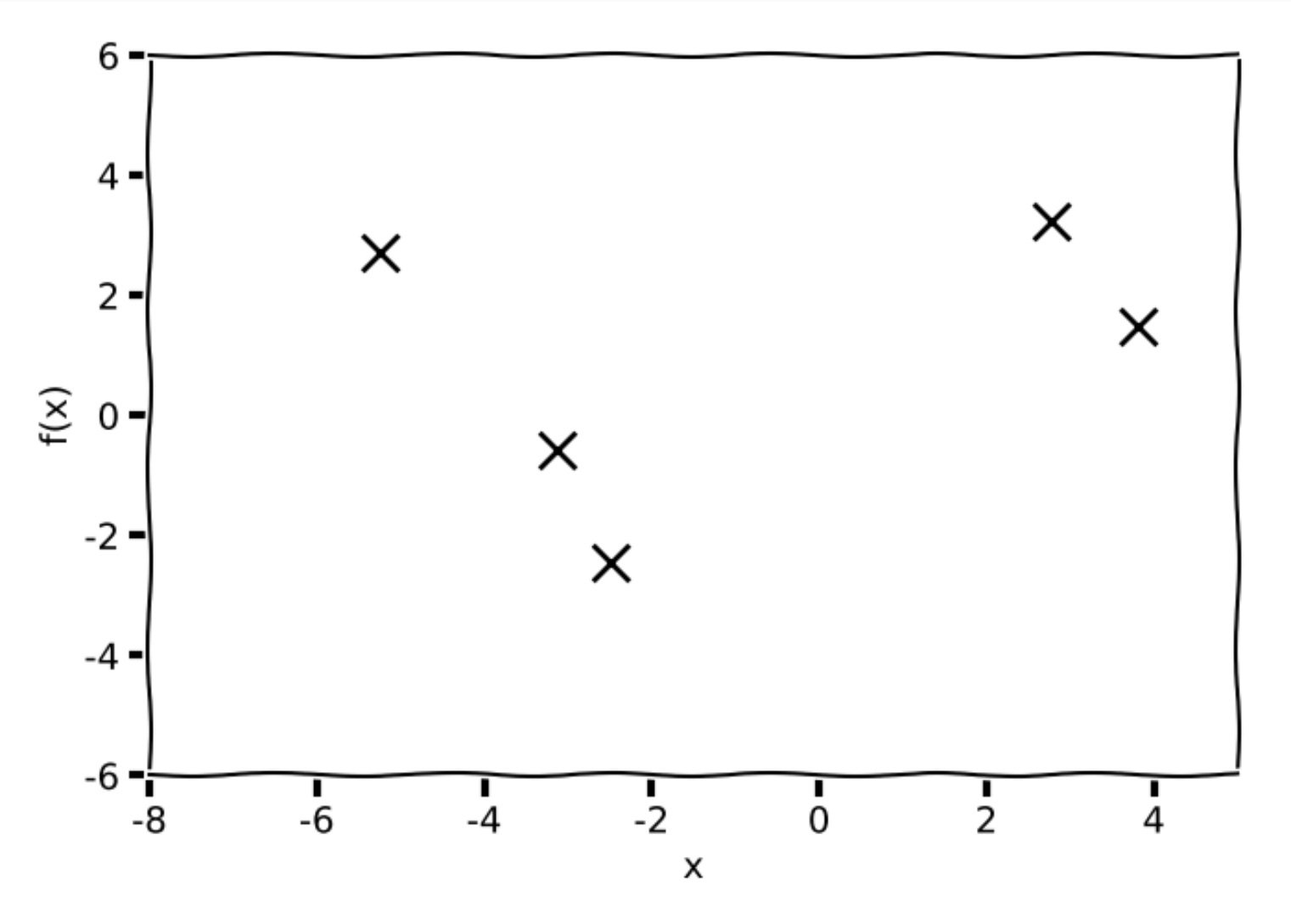
  $$f(x_M) - f(x'_M) \leq \epsilon$$

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

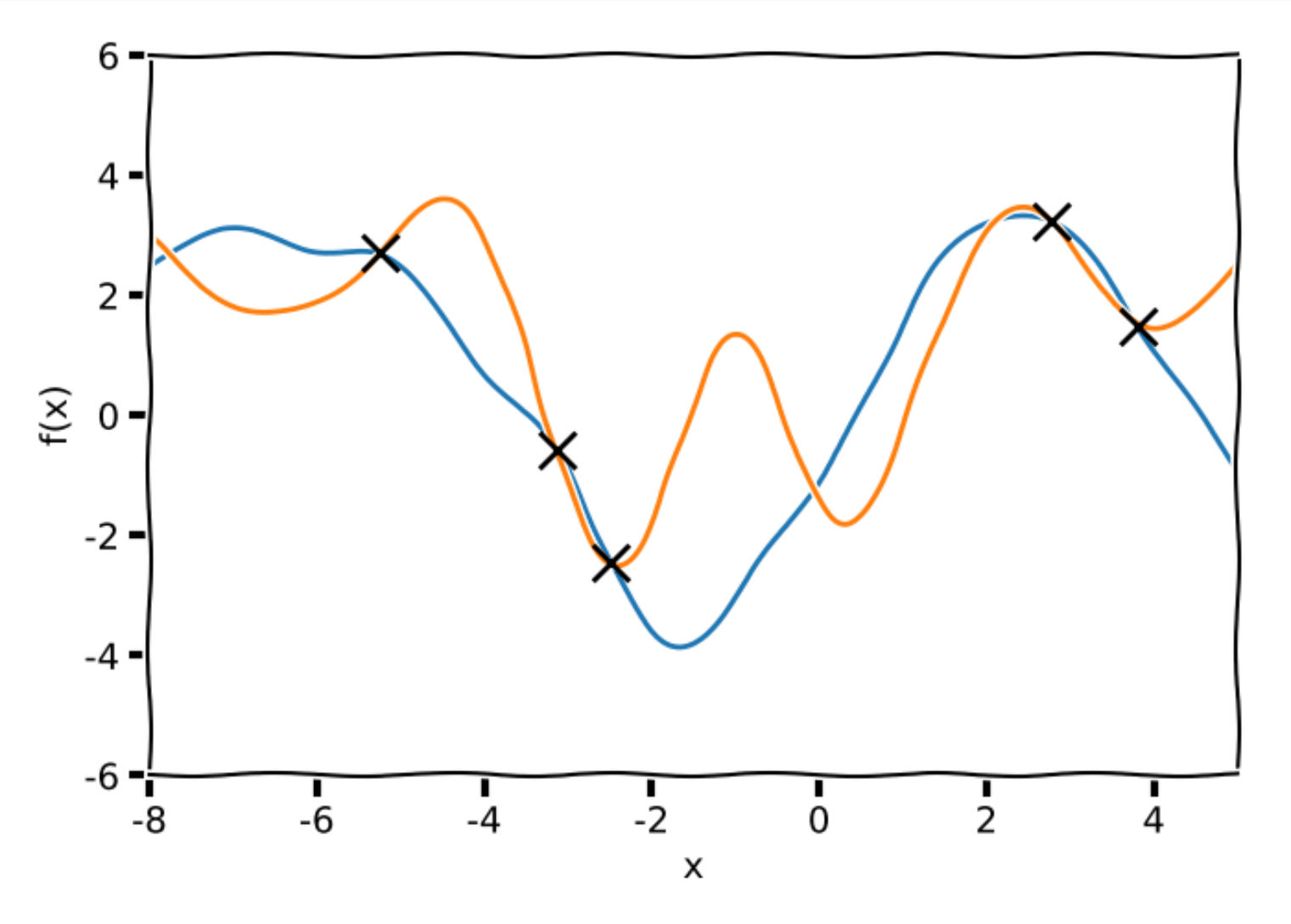
BO in practice



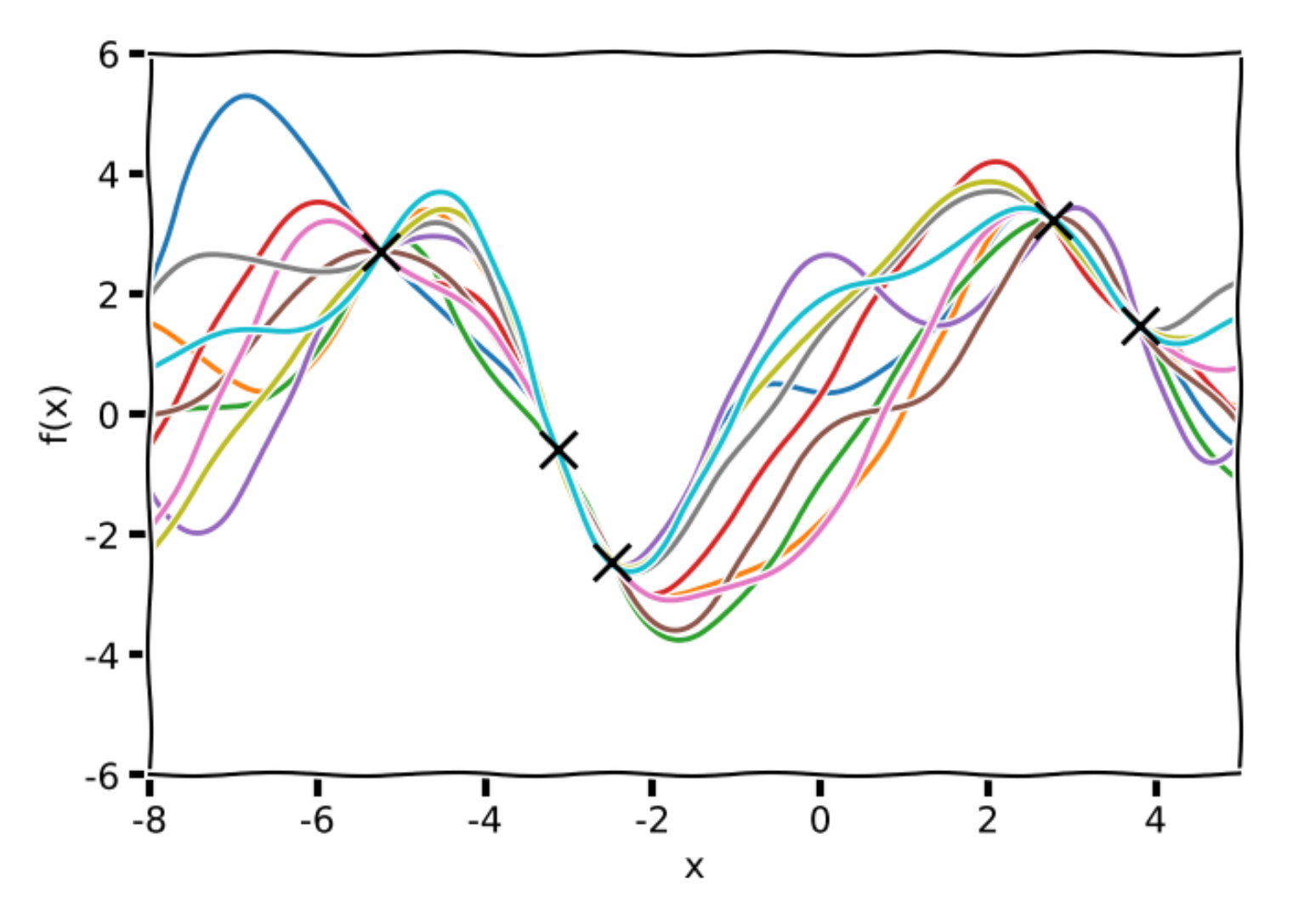
BO in practice



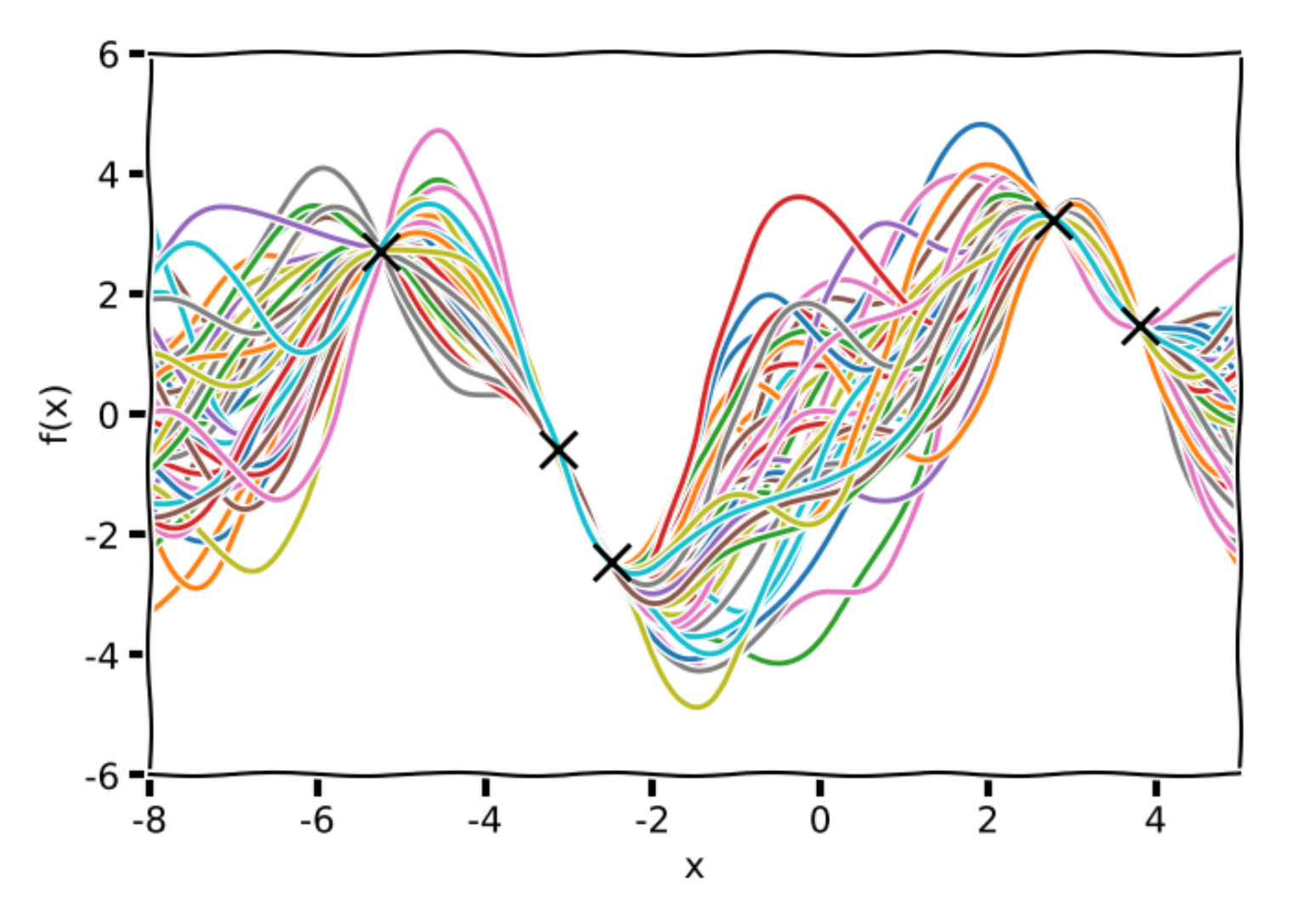
BO in practice



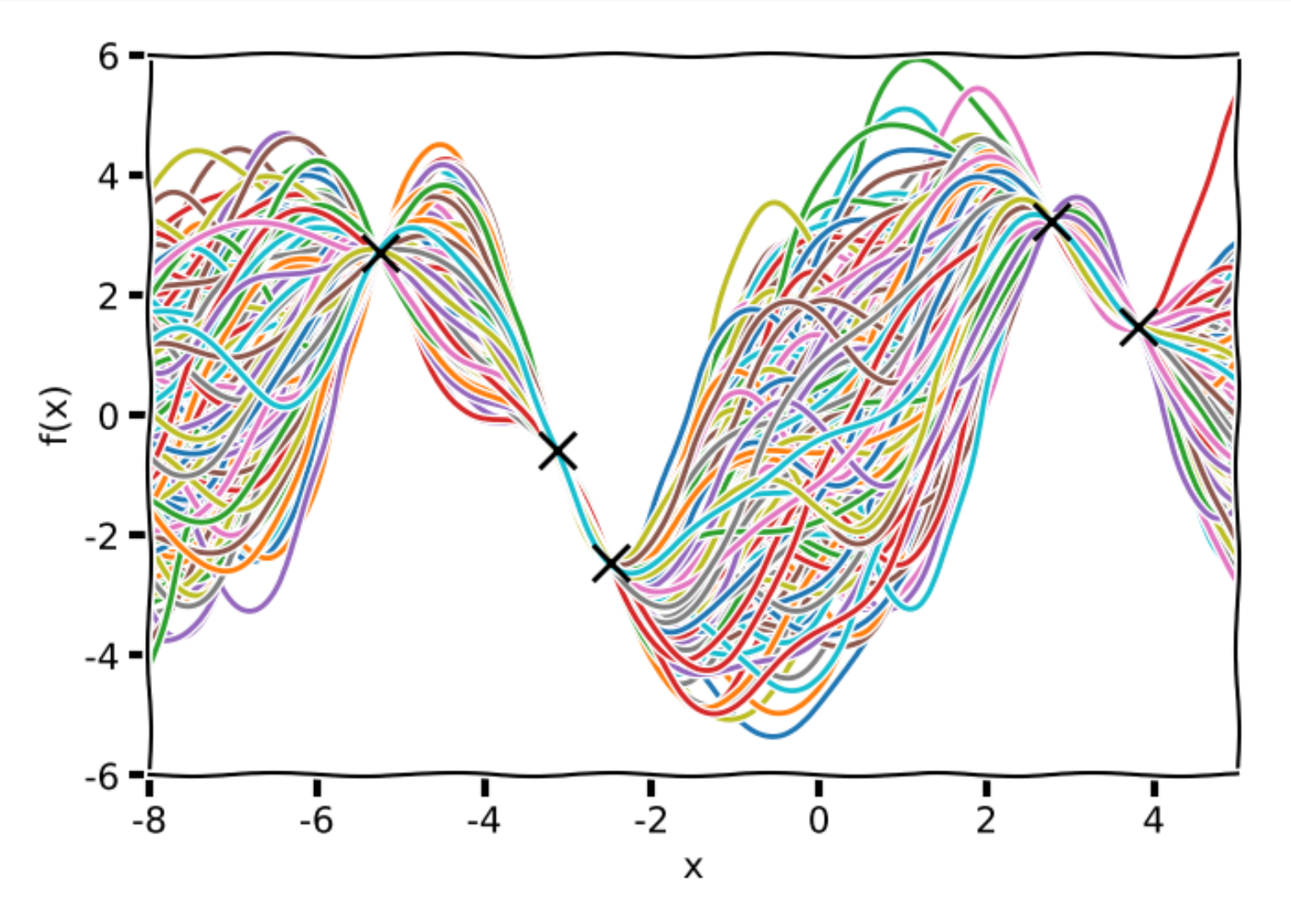
BO in practice



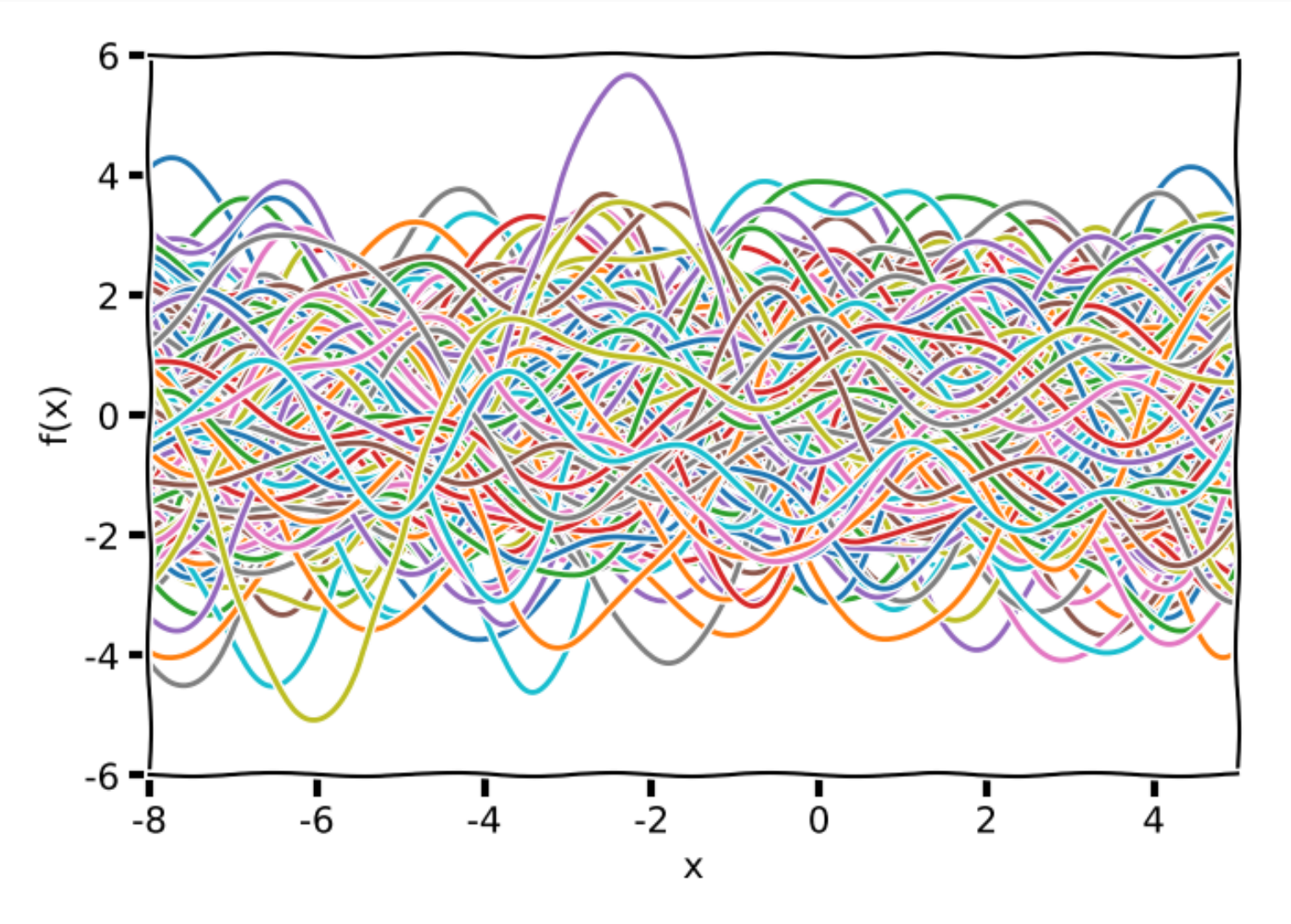
BO in practice



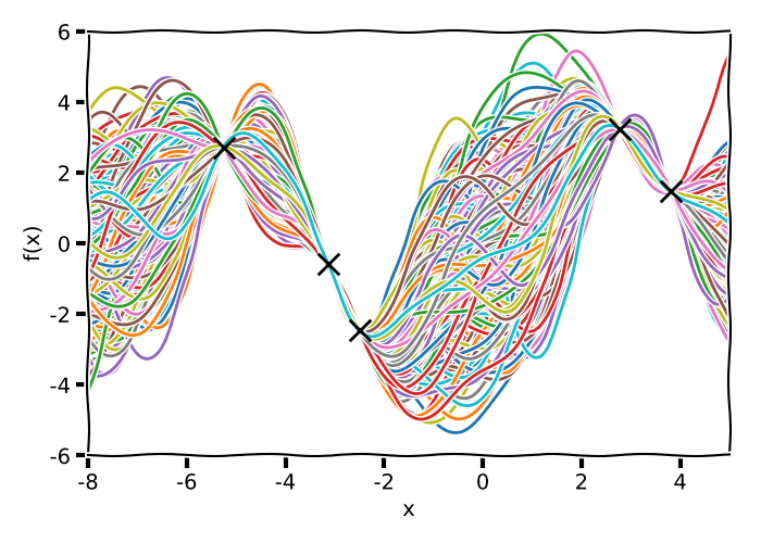
BO in practice

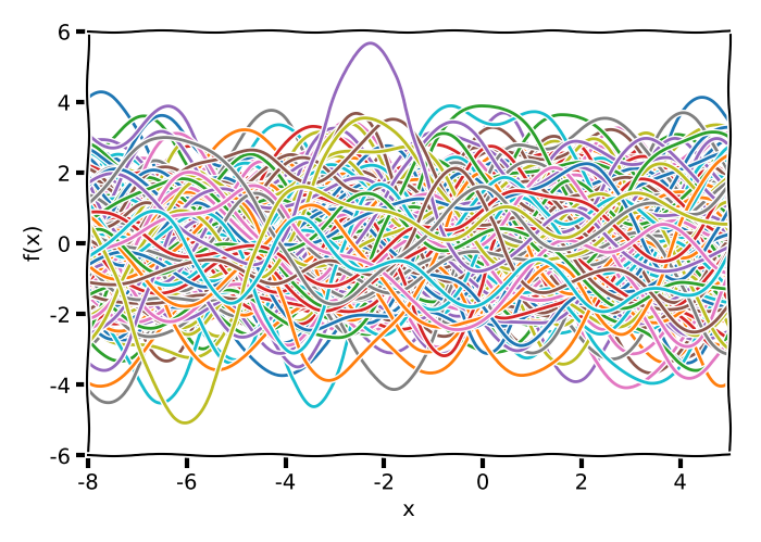


BO in practice



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

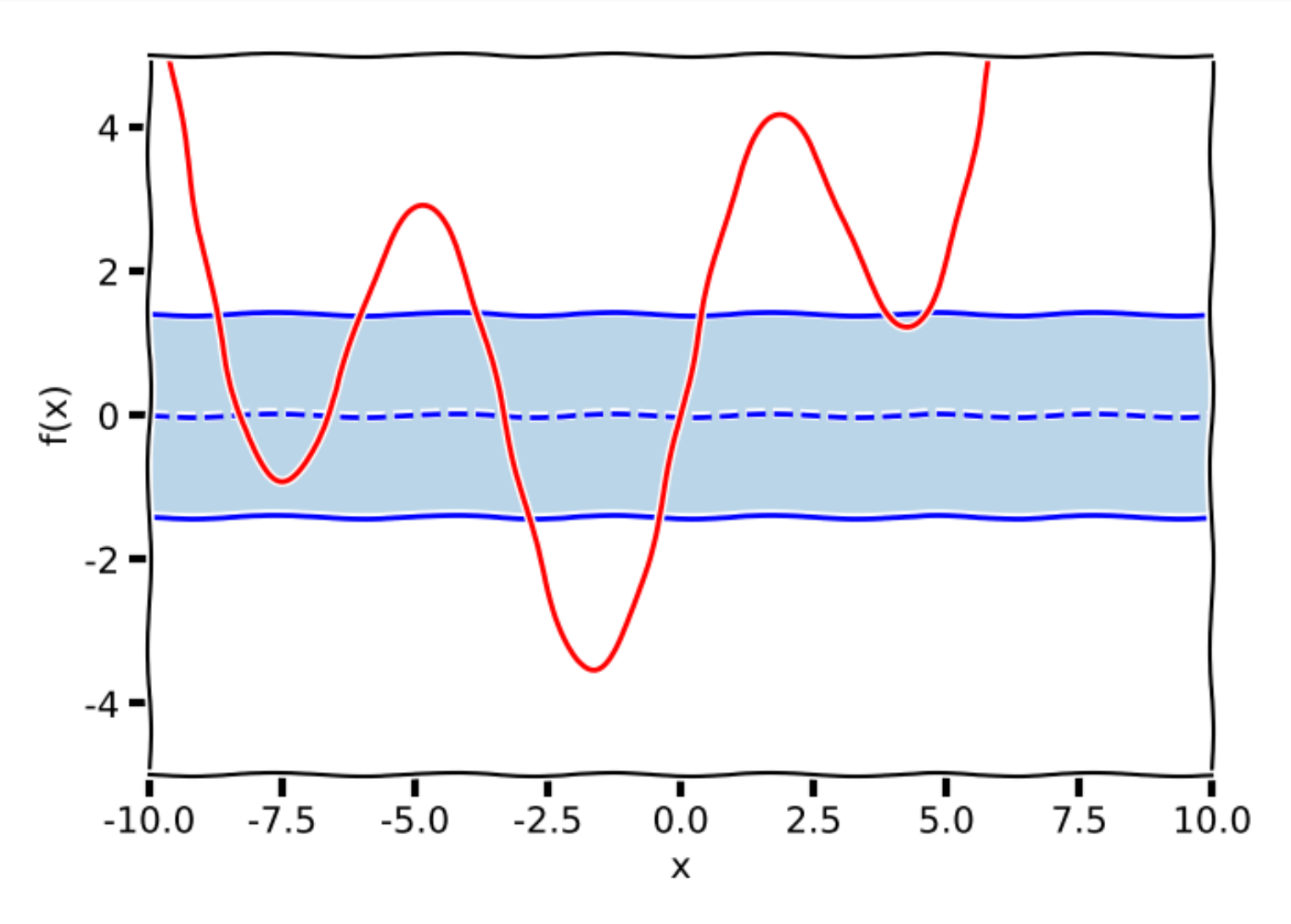
Bayesian Optimisation

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

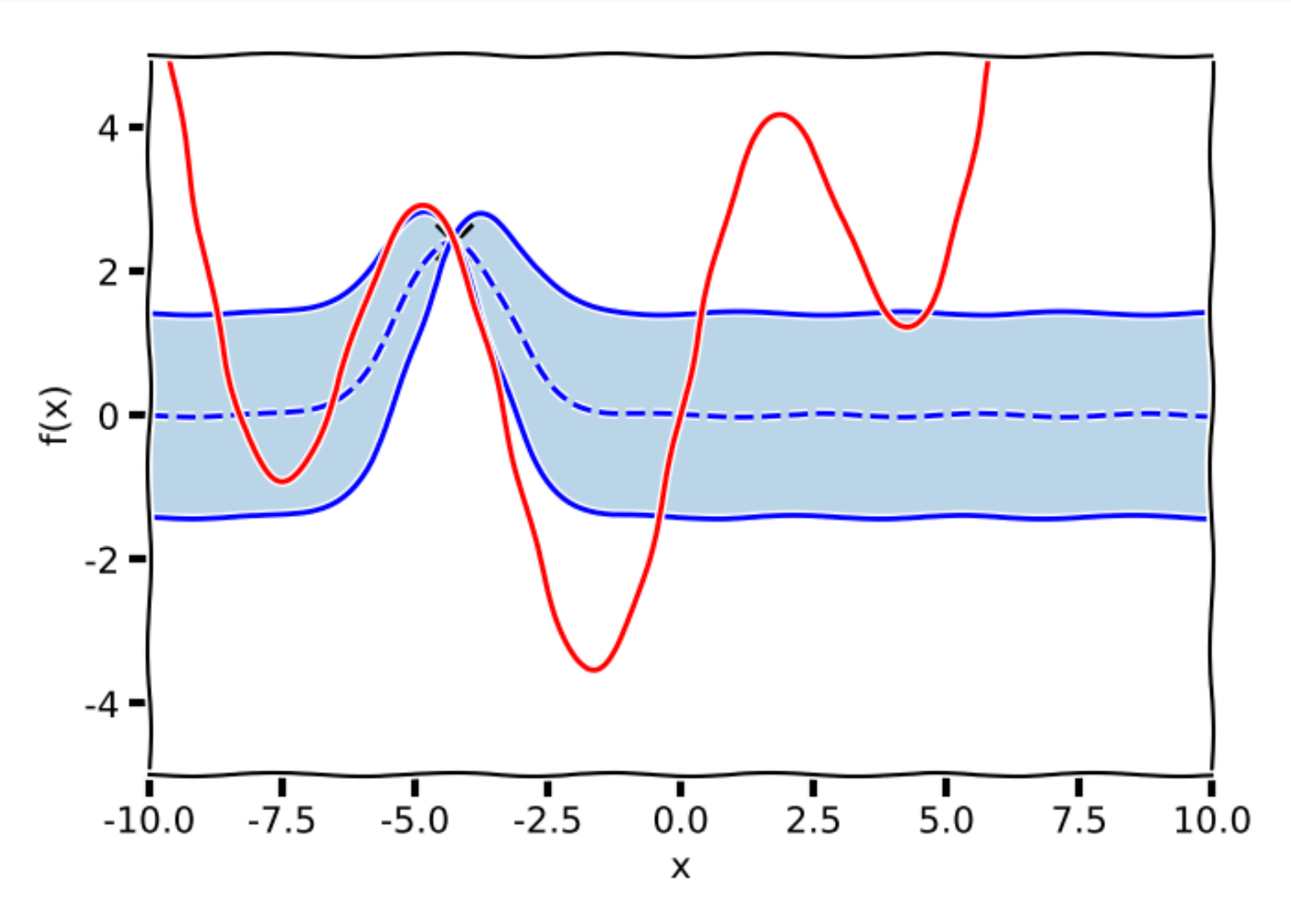
Bayesian Optimisation

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

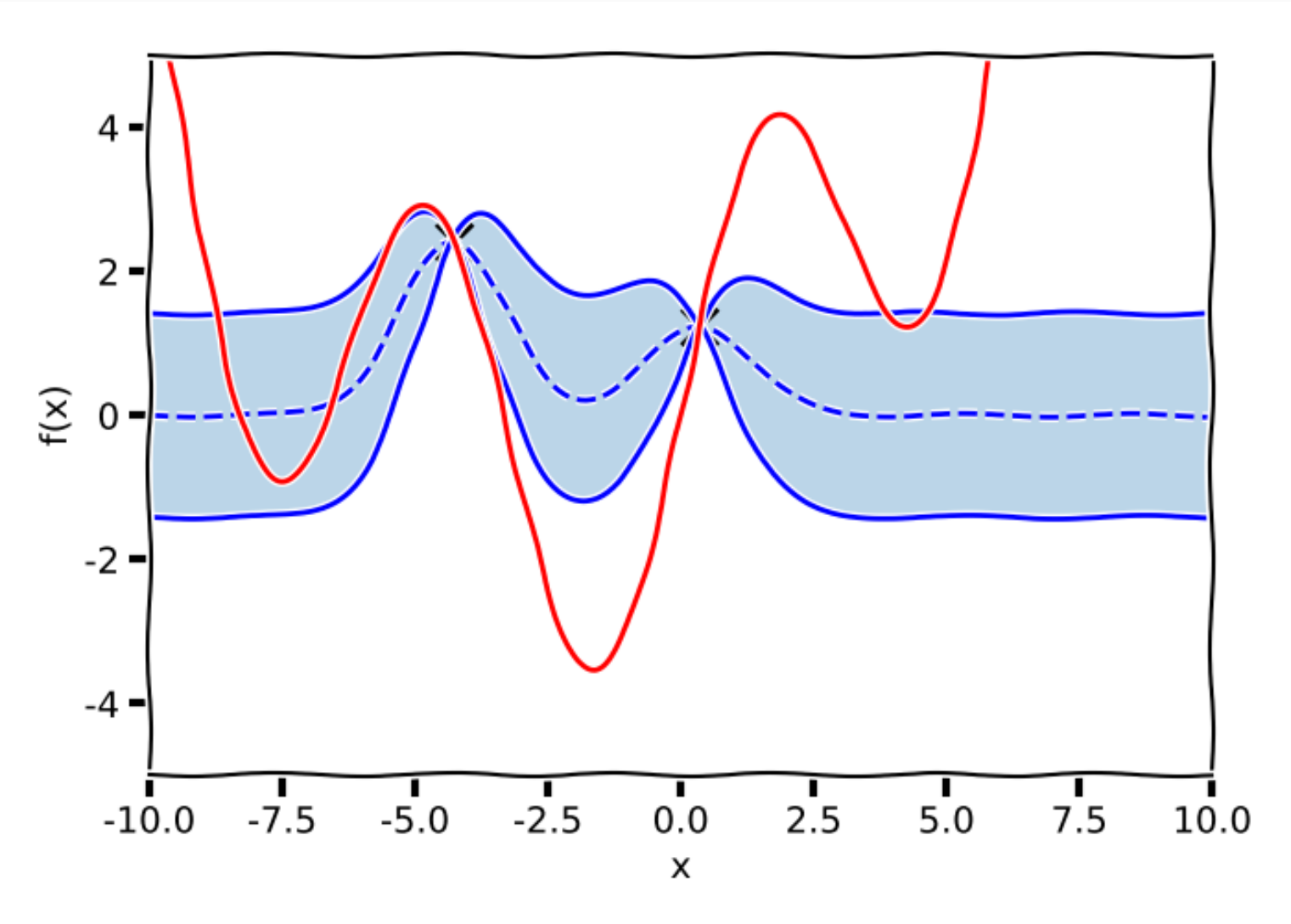
Gaussian Processes



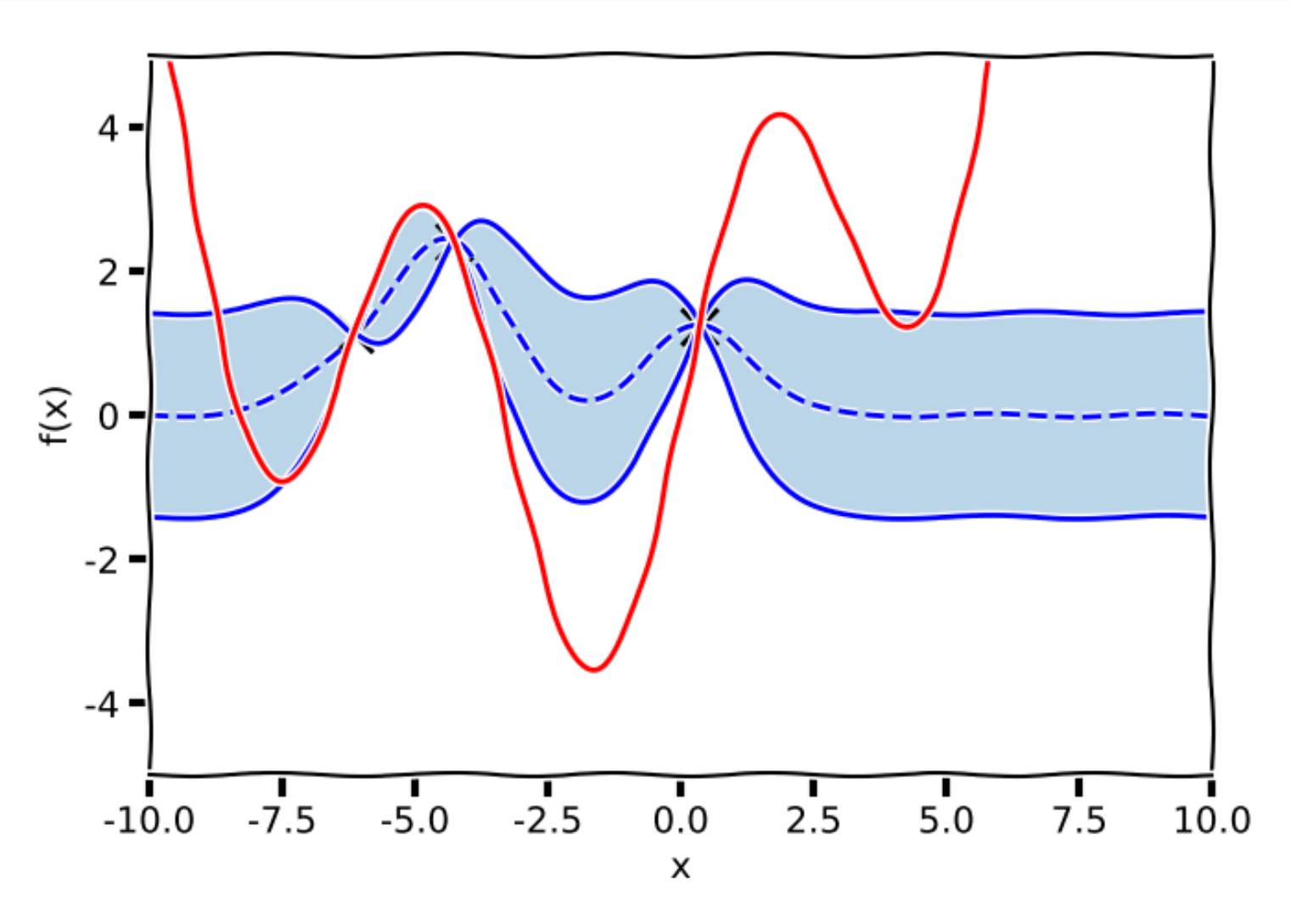
Gaussian Processes



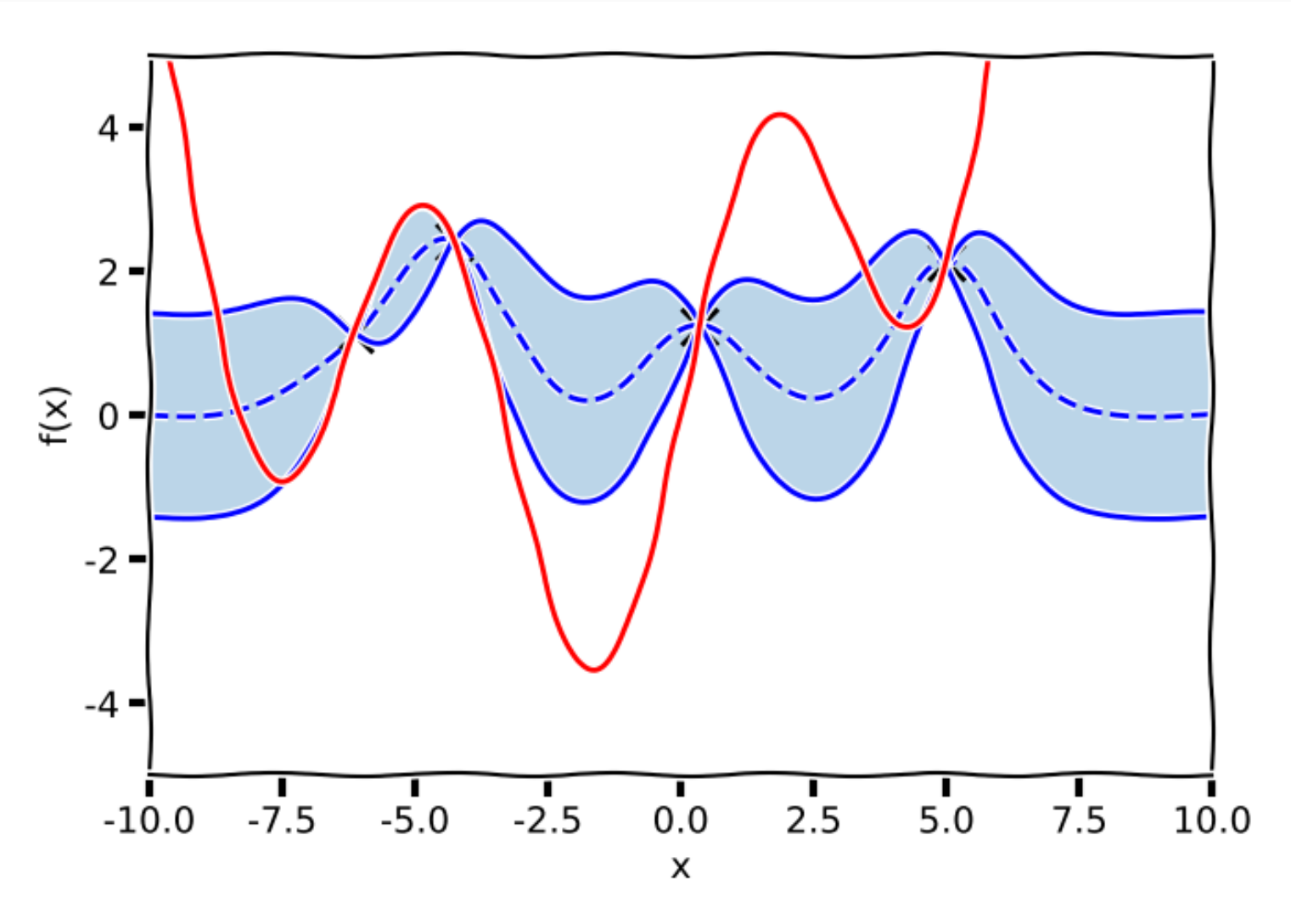
Gaussian Processes



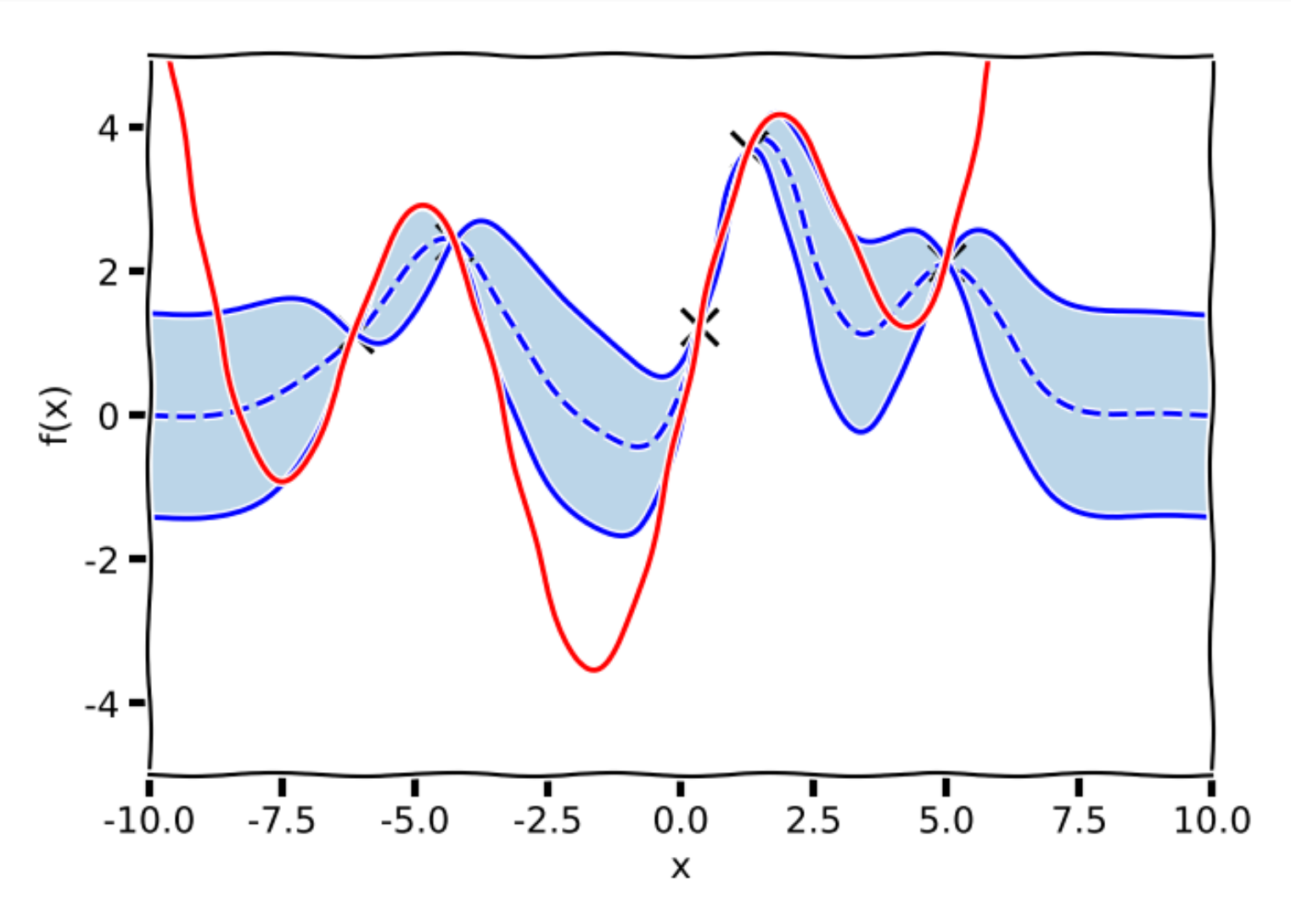
Gaussian Processes



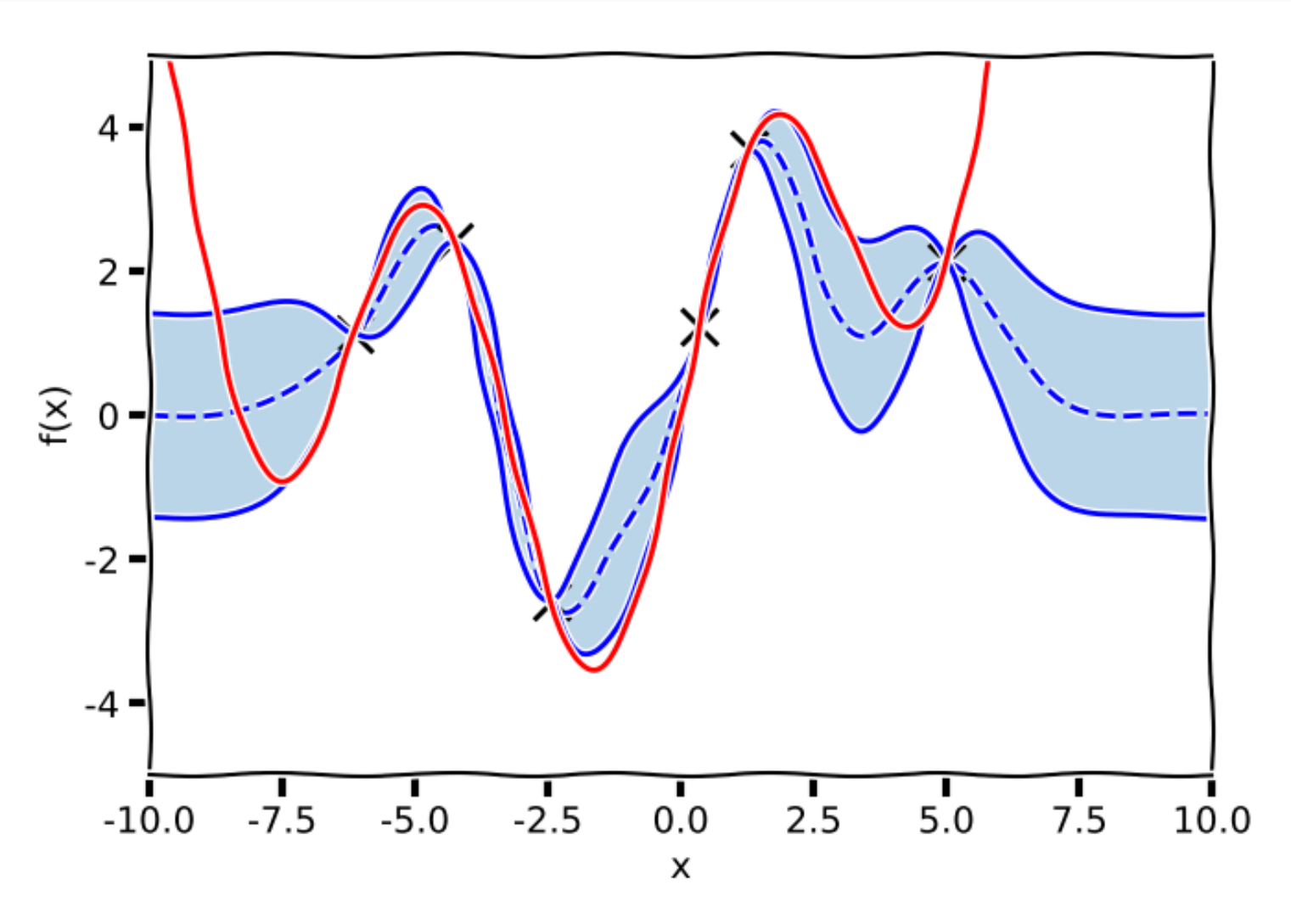
Gaussian Processes



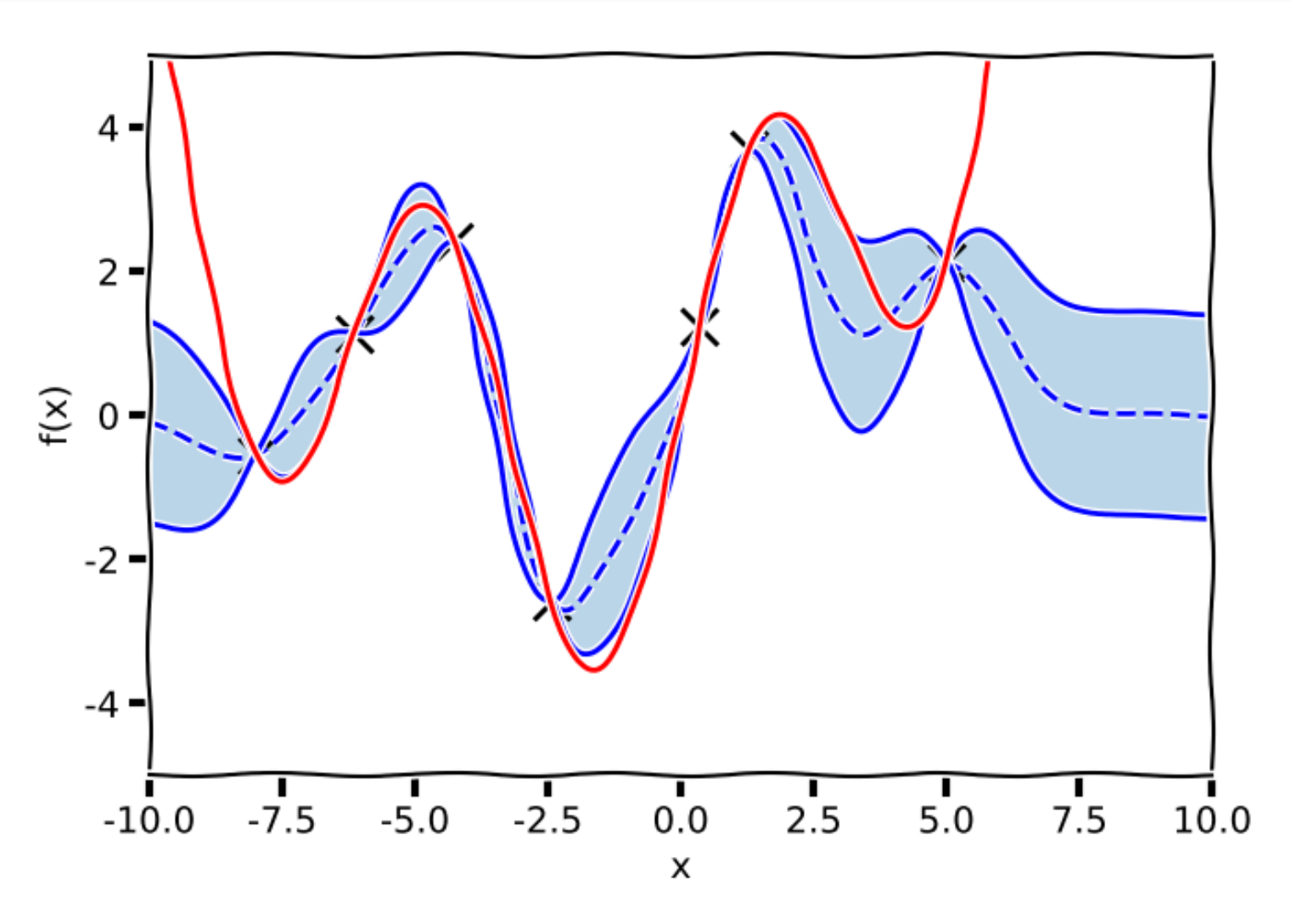
Gaussian Processes



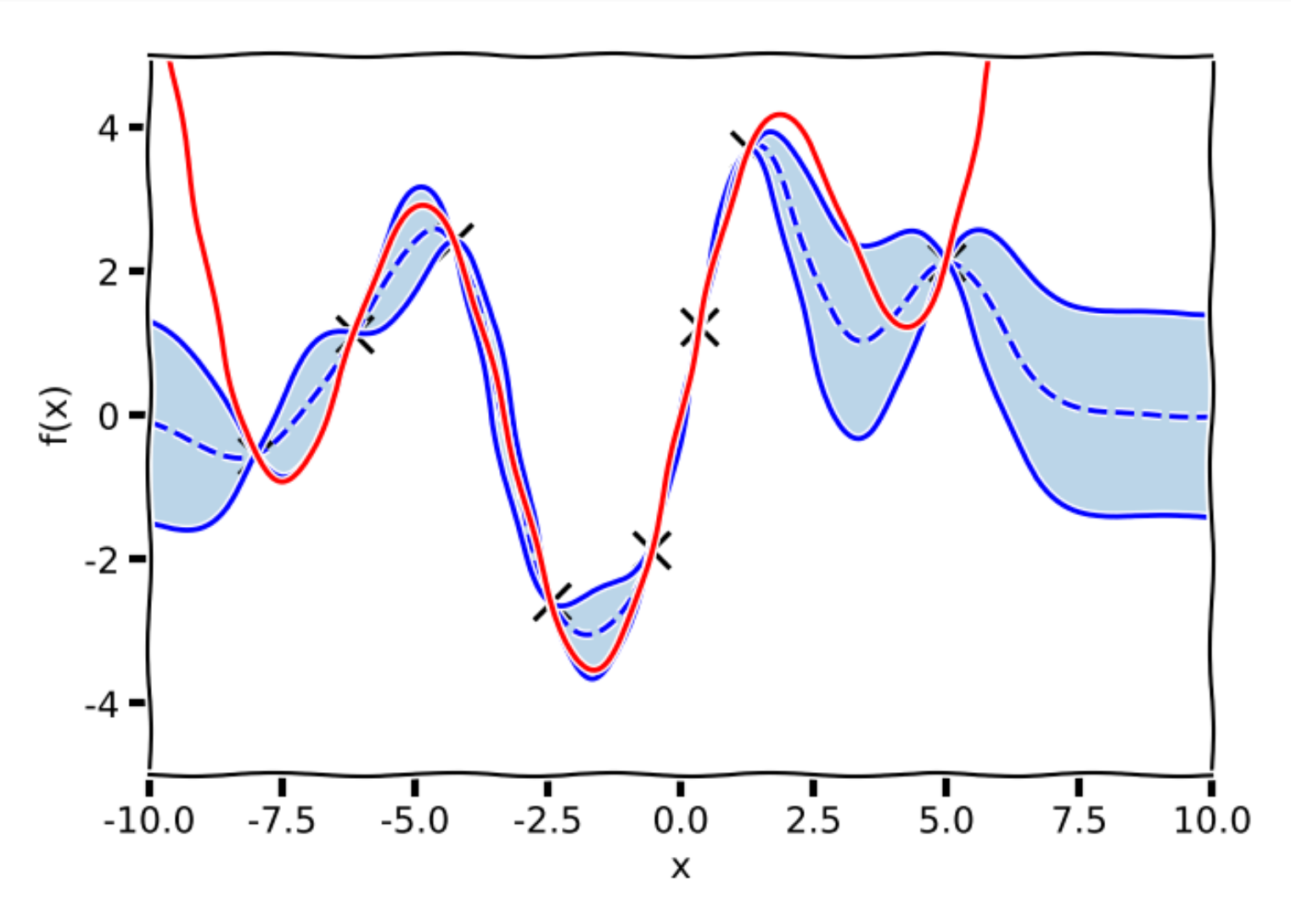
Gaussian Processes



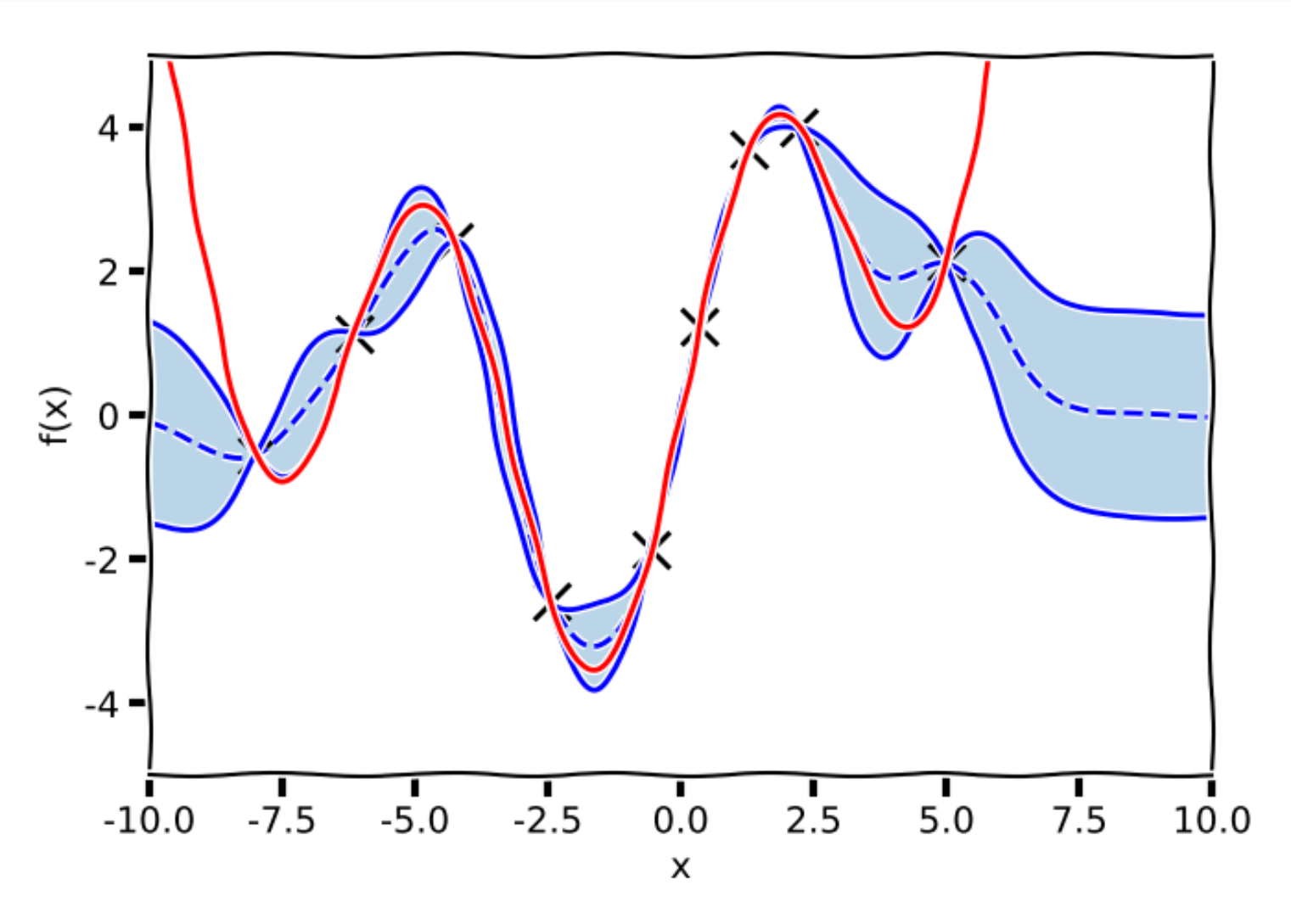
Gaussian Processes



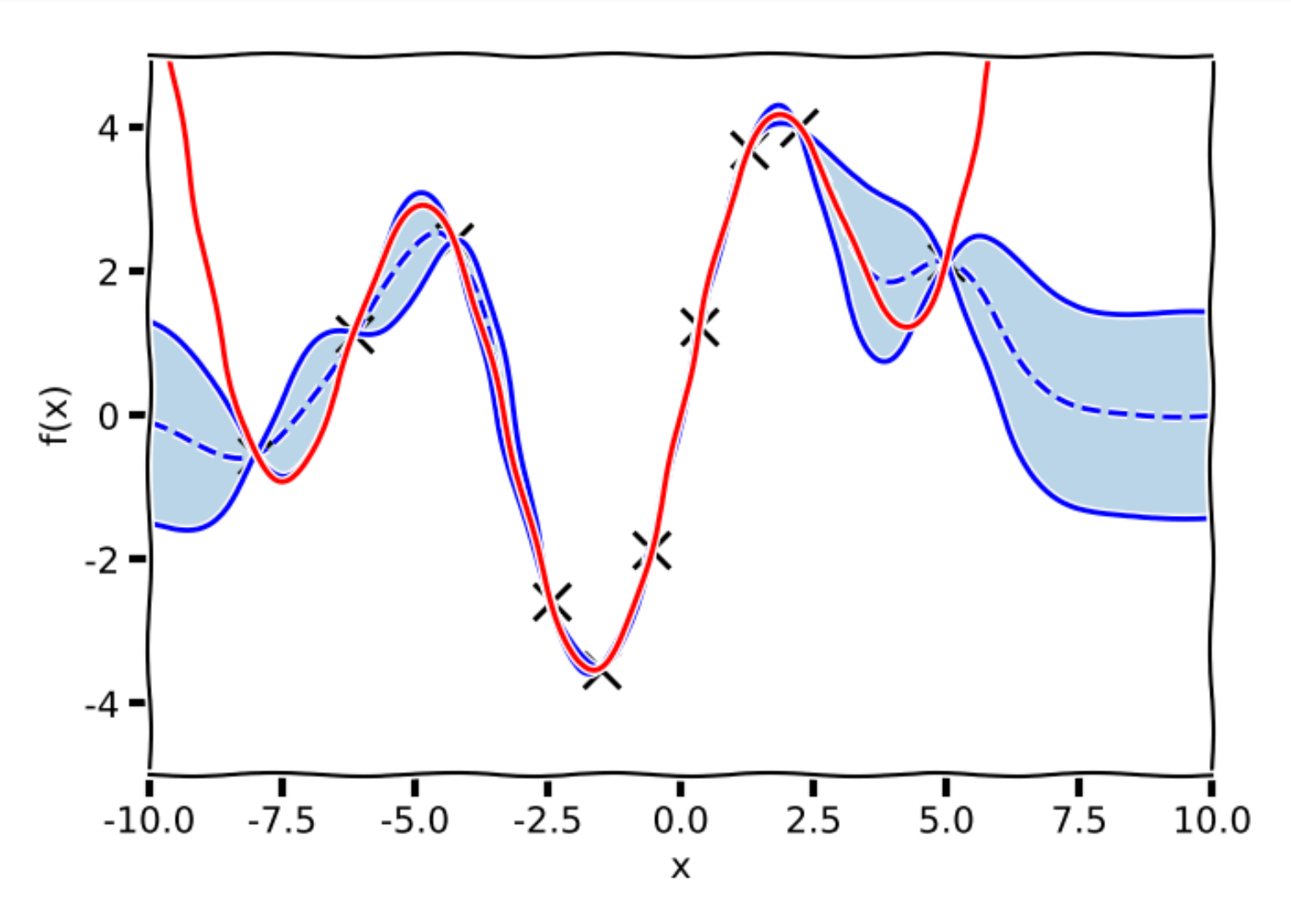
Gaussian Processes



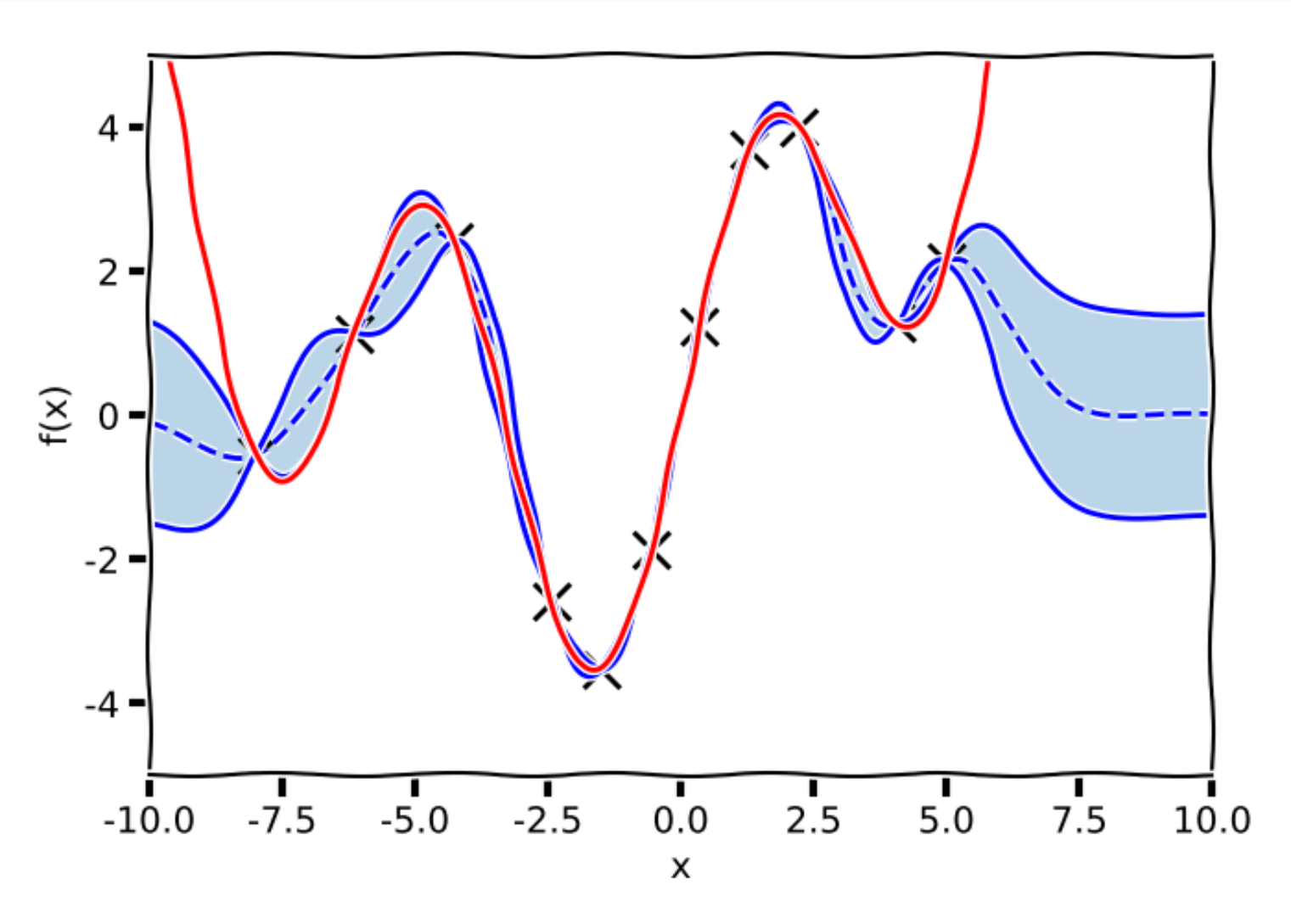
Gaussian Processes



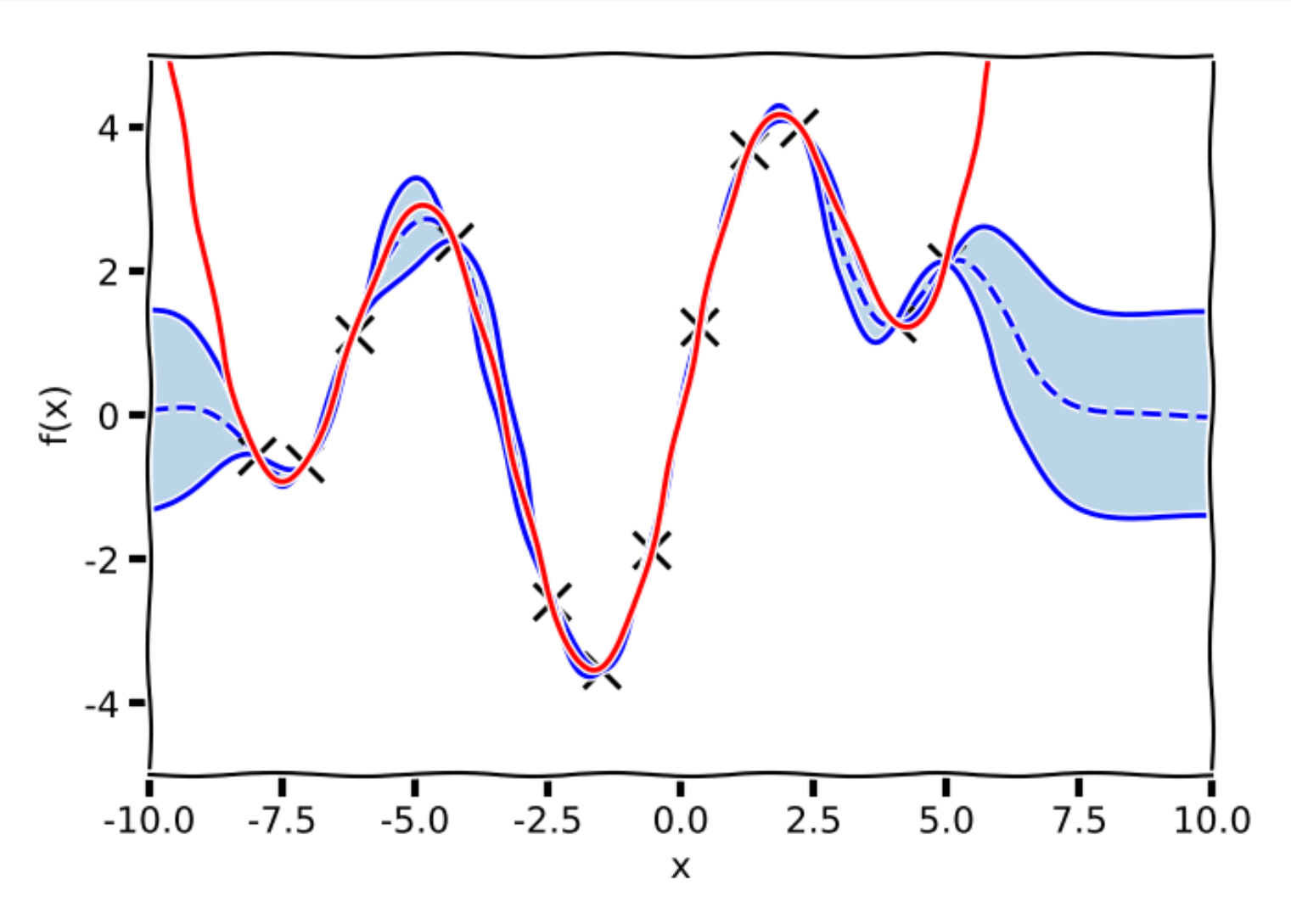
Gaussian Processes



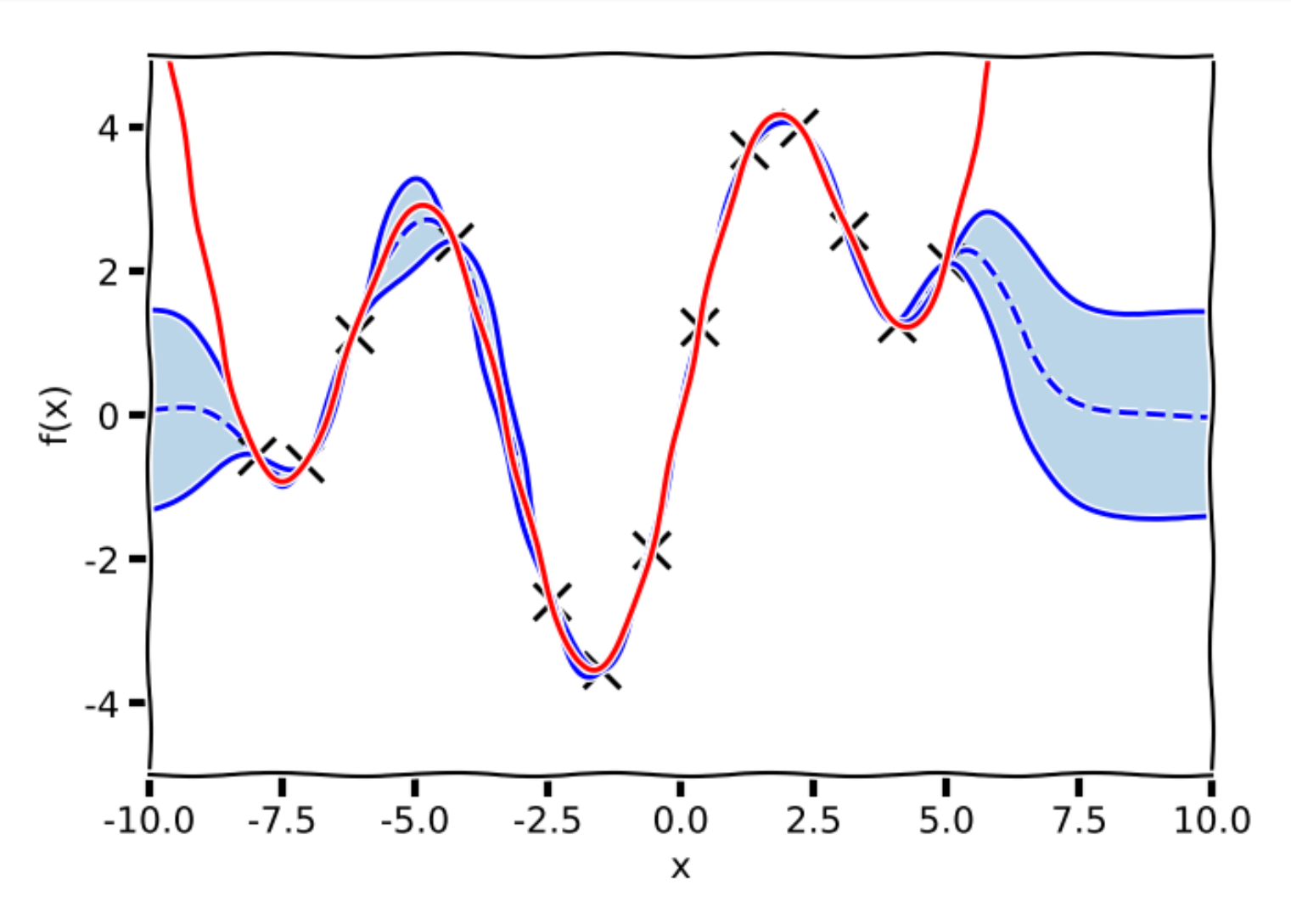
Gaussian Processes



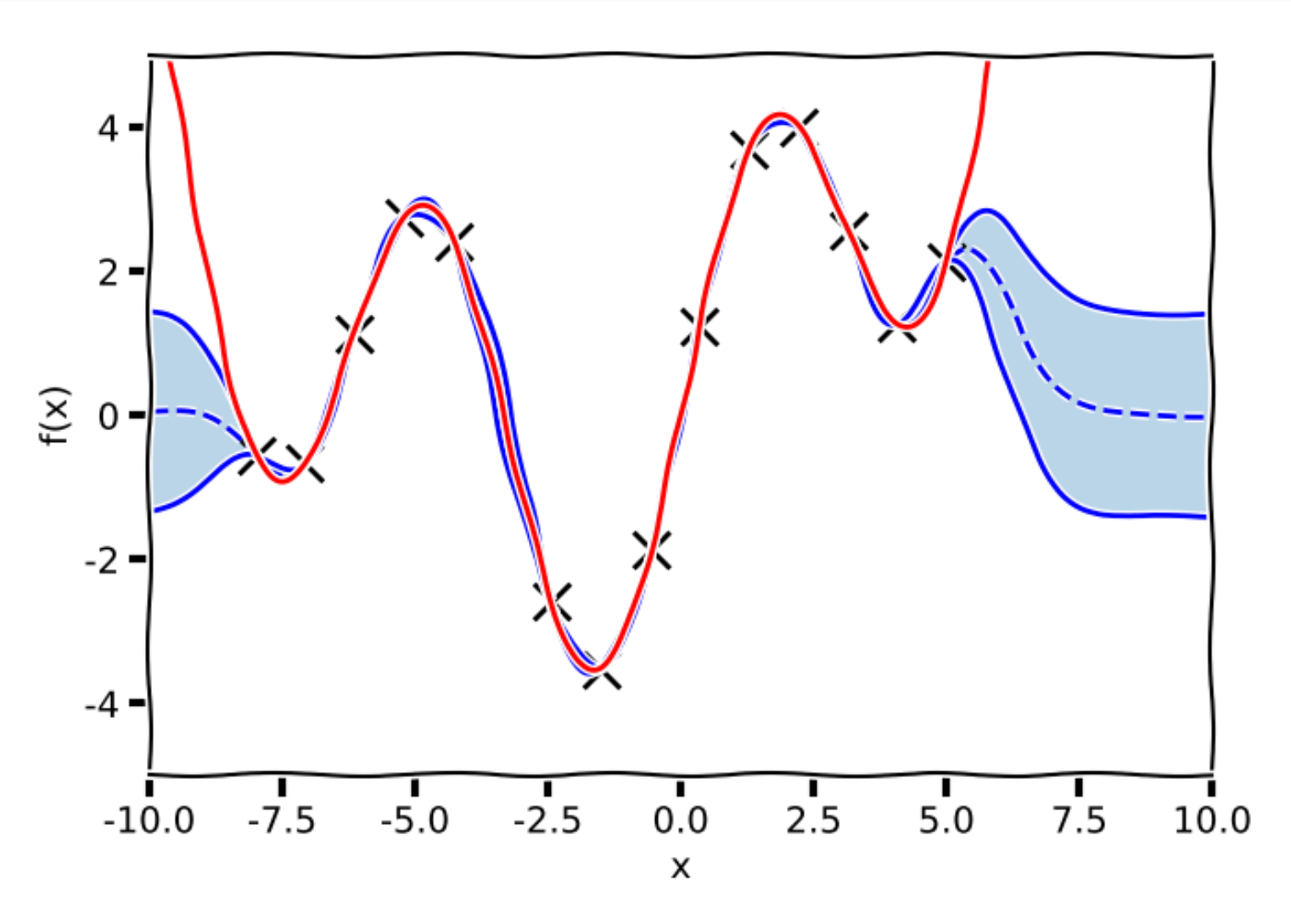
Gaussian Processes



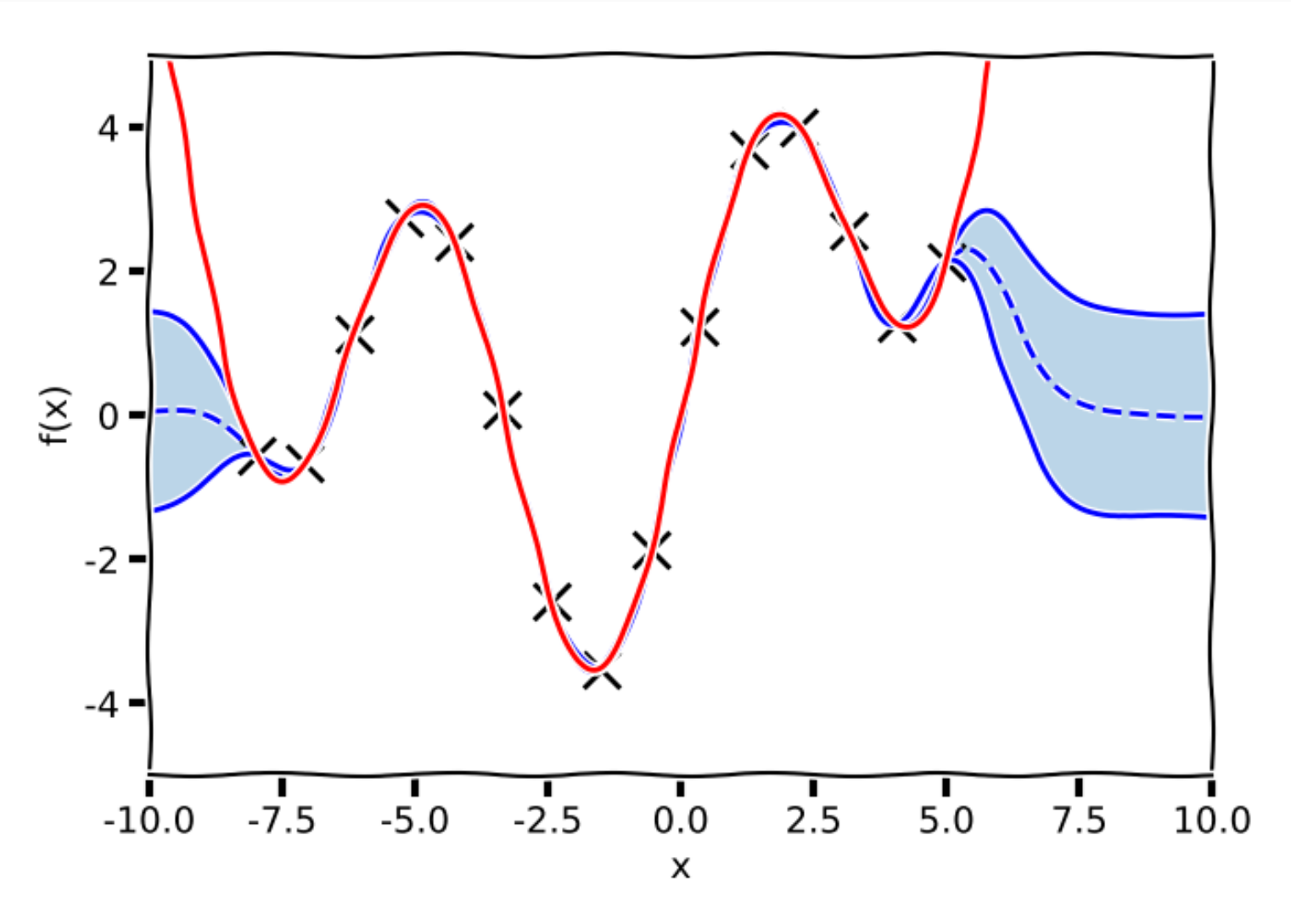
Gaussian Processes



Gaussian Processes



Gaussian Processes



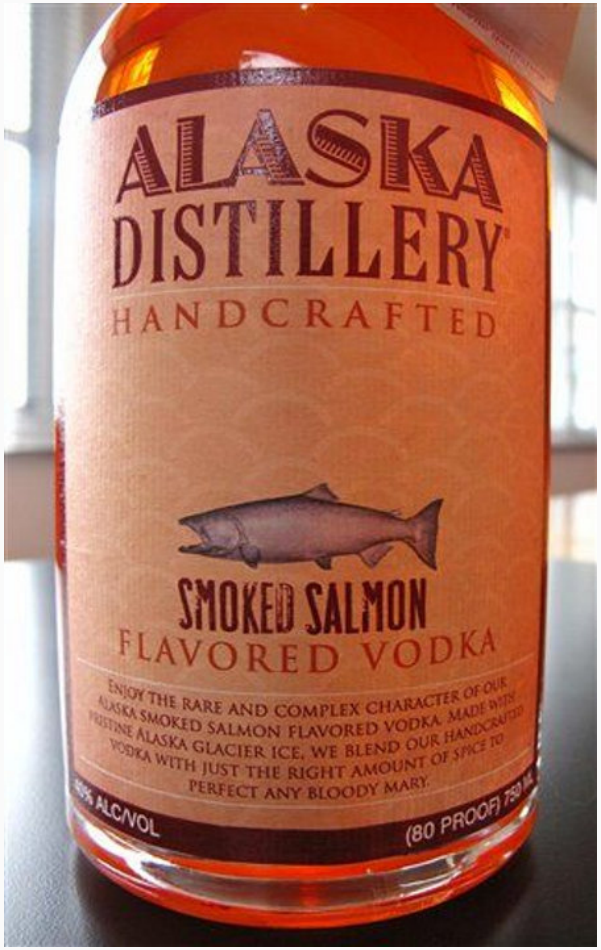
Bayesian Optimisation

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

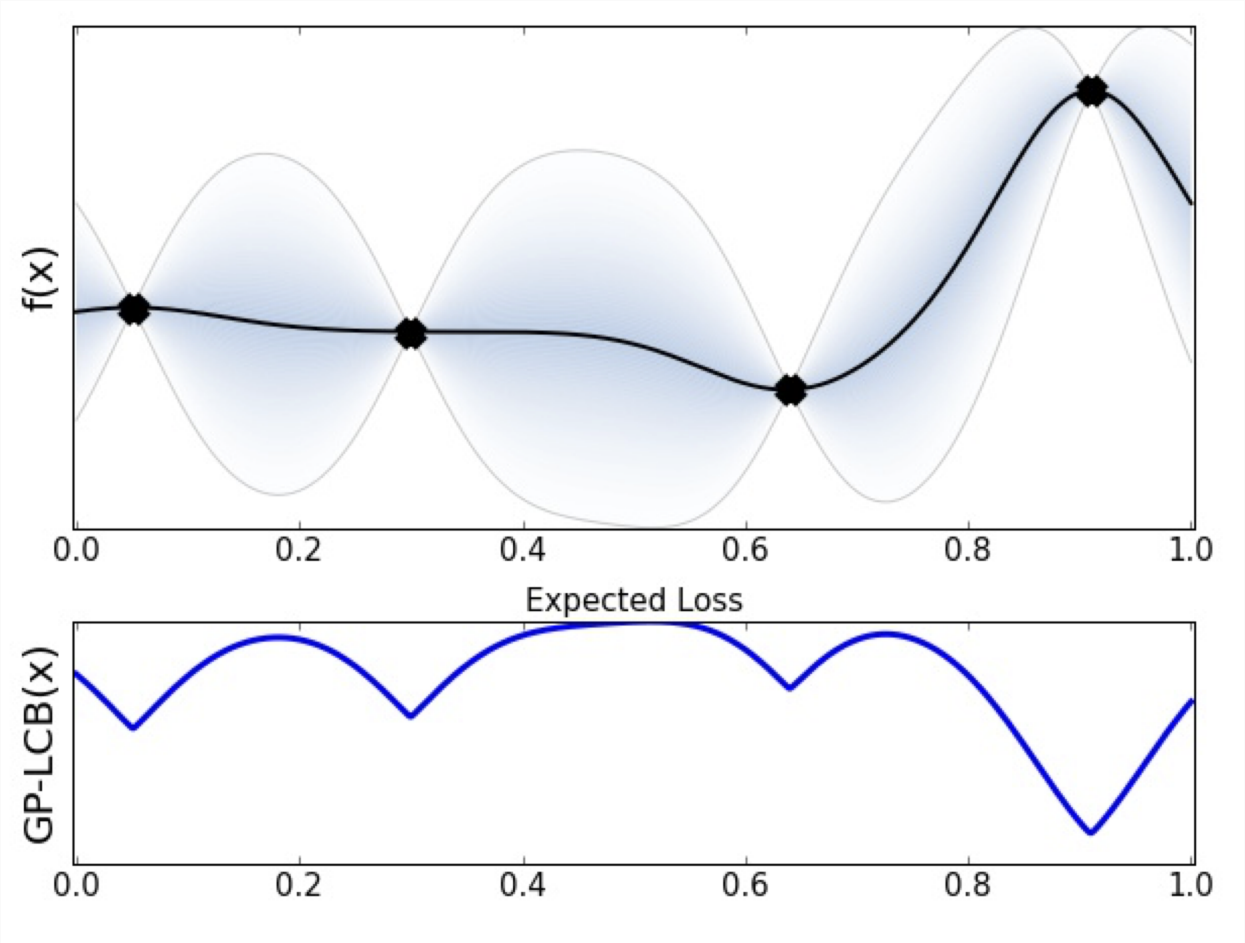
Balance is the key



Aquisition function

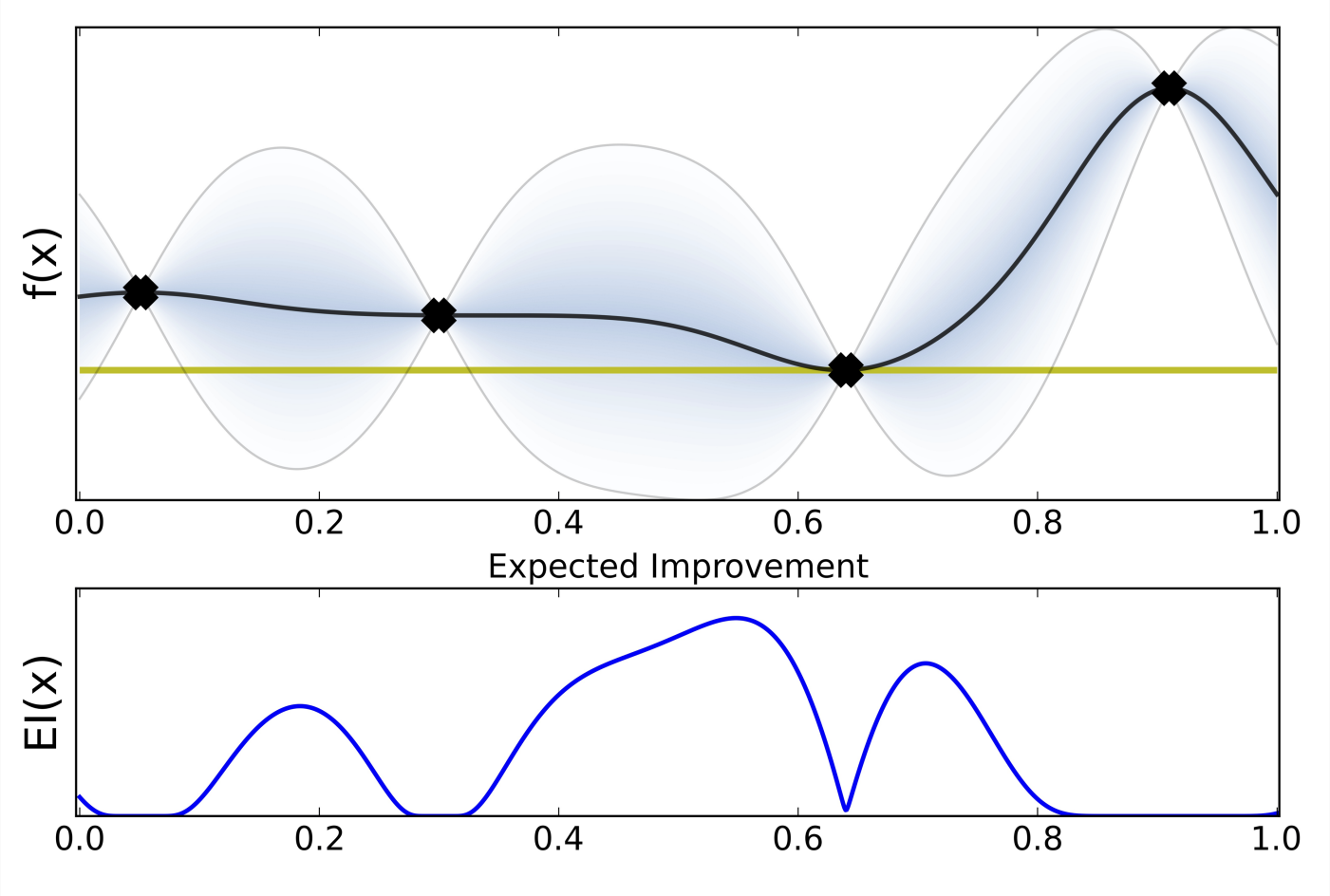
- how should we explore the space?
- humans do active search?
    - studying for exam
-

Confidence



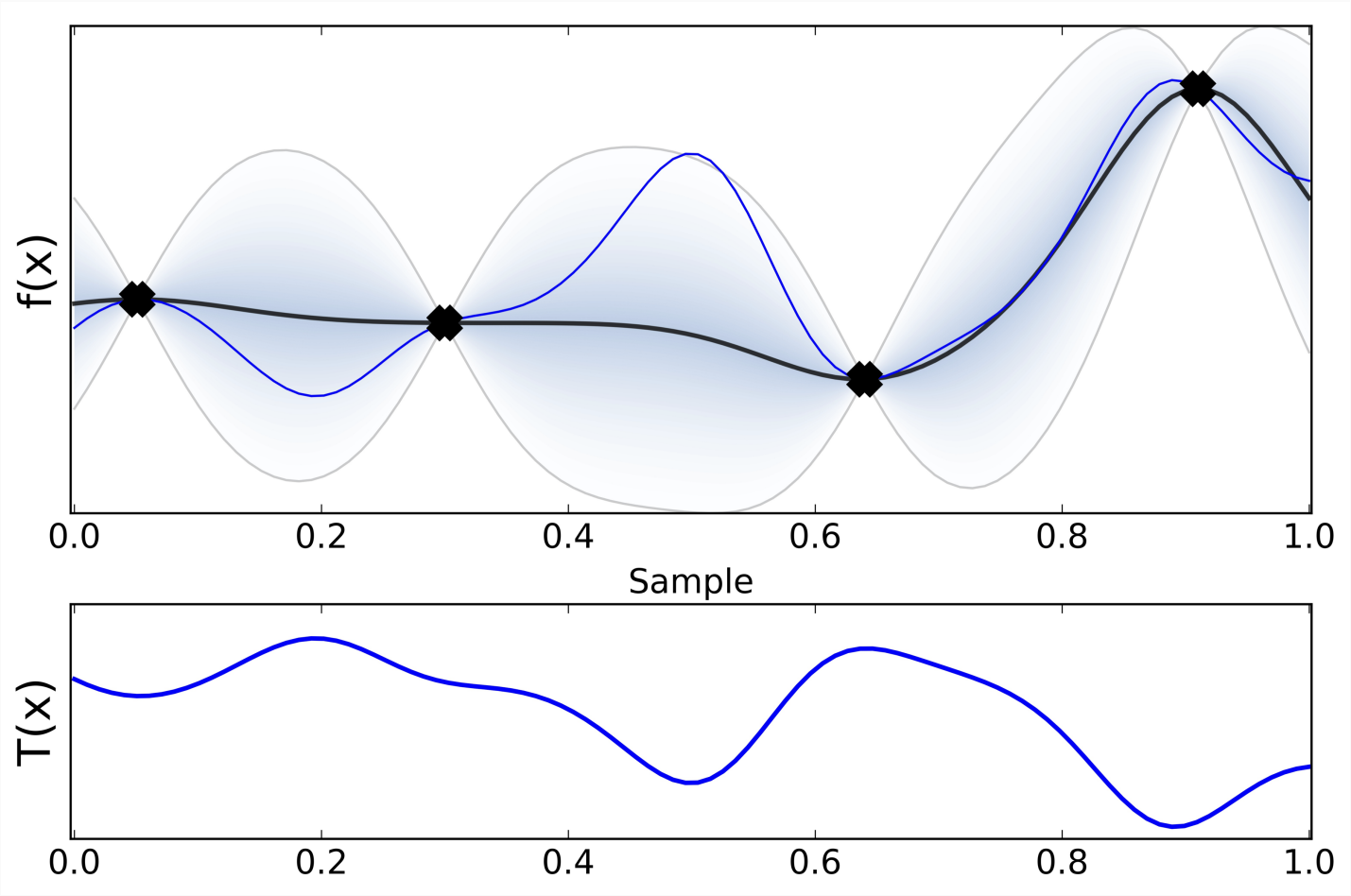
$$\alpha(\mathbf{x}; \theta, \mathcal{D}) = -\mu(\mathbf{x}; \theta, \mathcal{D}) + \beta_t \sigma(\mathbf{x}; \theta, \mathcal{D})$$

Expected Improvement



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

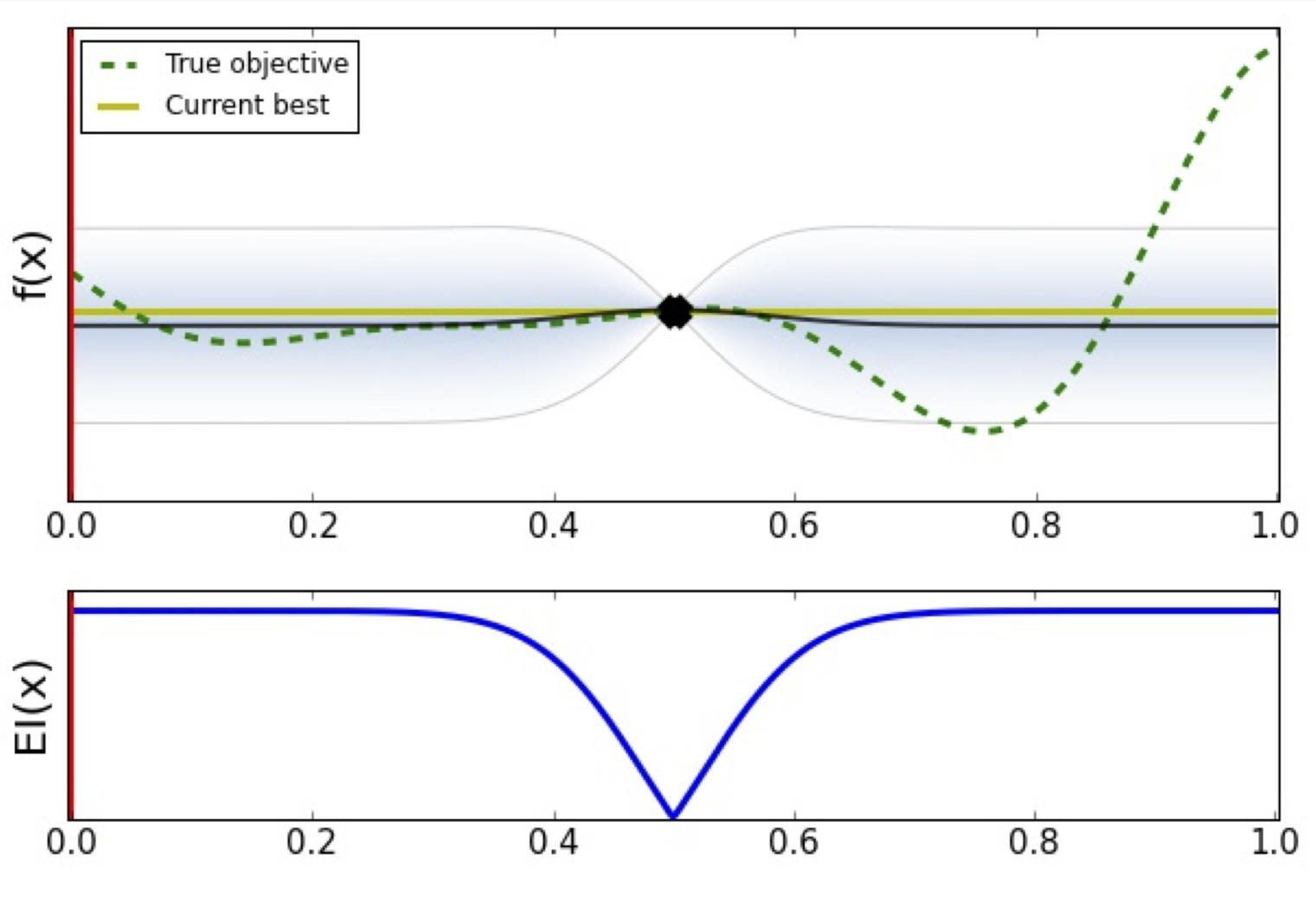
Thomson Sampling



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

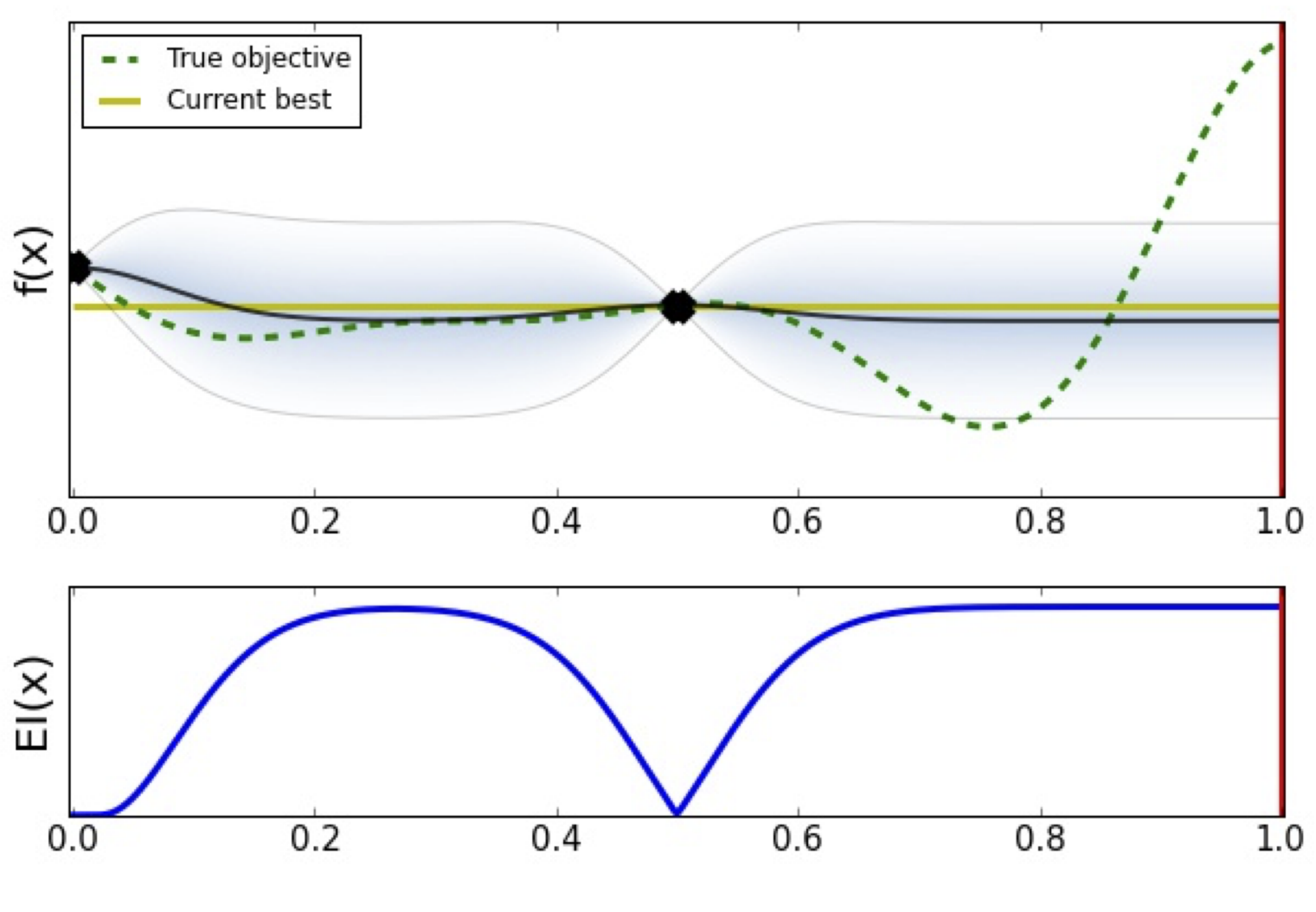
$$g(\mathbf{x}) \sim p(y|\mathbf{x}, \theta, \mathcal{D})$$

Bayesian Optimisation [1]



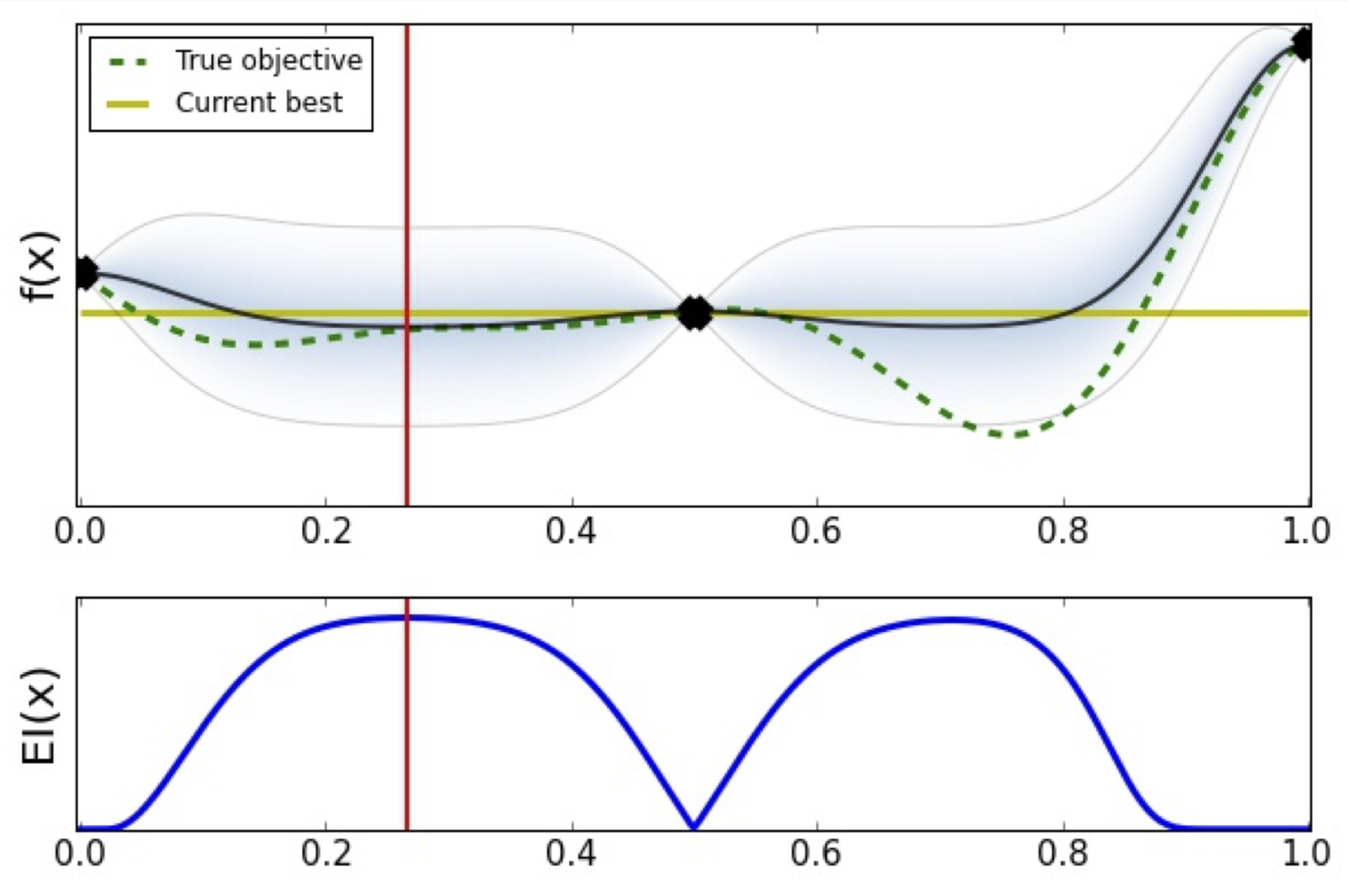
[1] Slides courtesy of Javier Gonzales

Bayesian Optimisation [1]



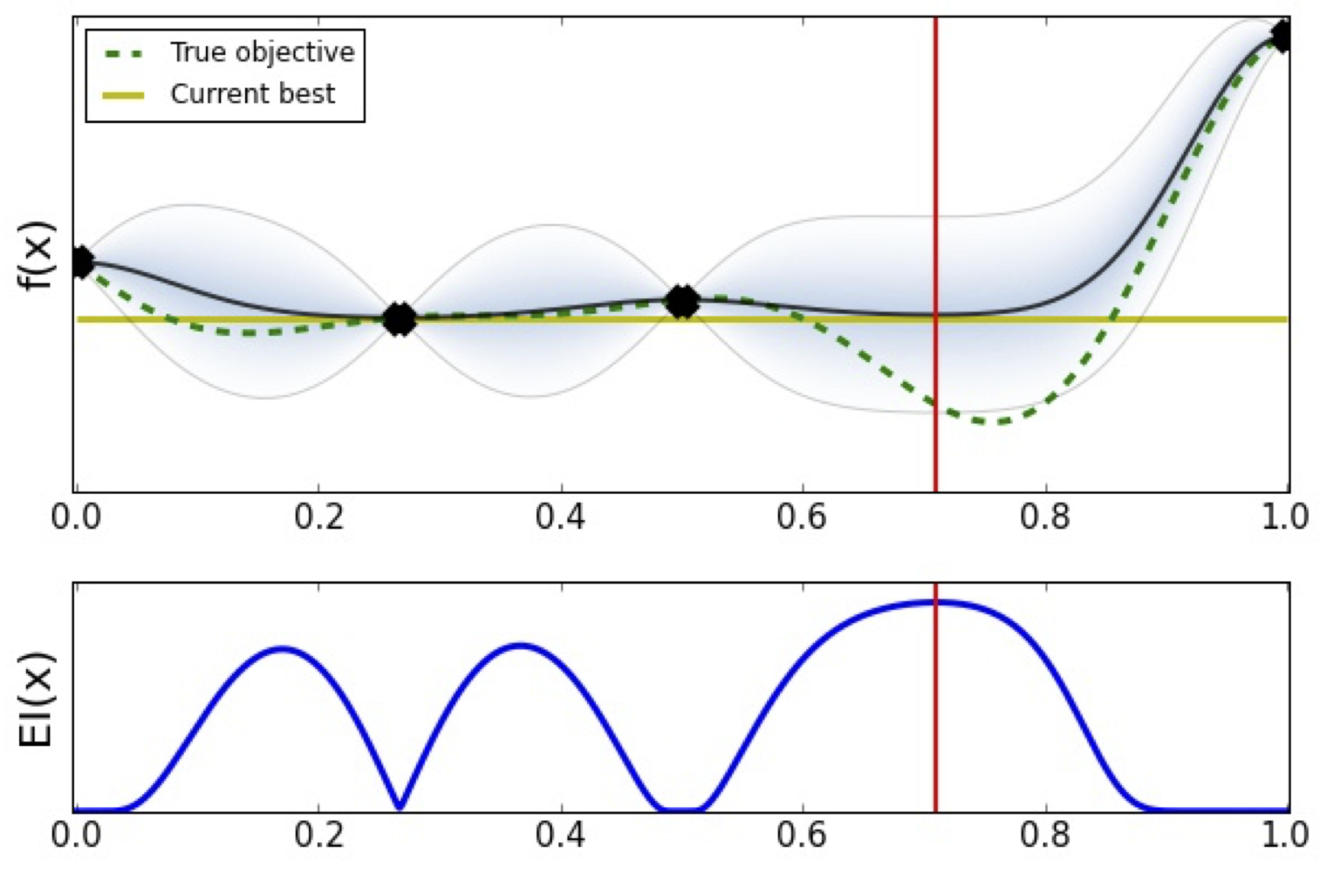
[1] Slides courtesy of Javier Gonzales

Bayesian Optimisation [1]



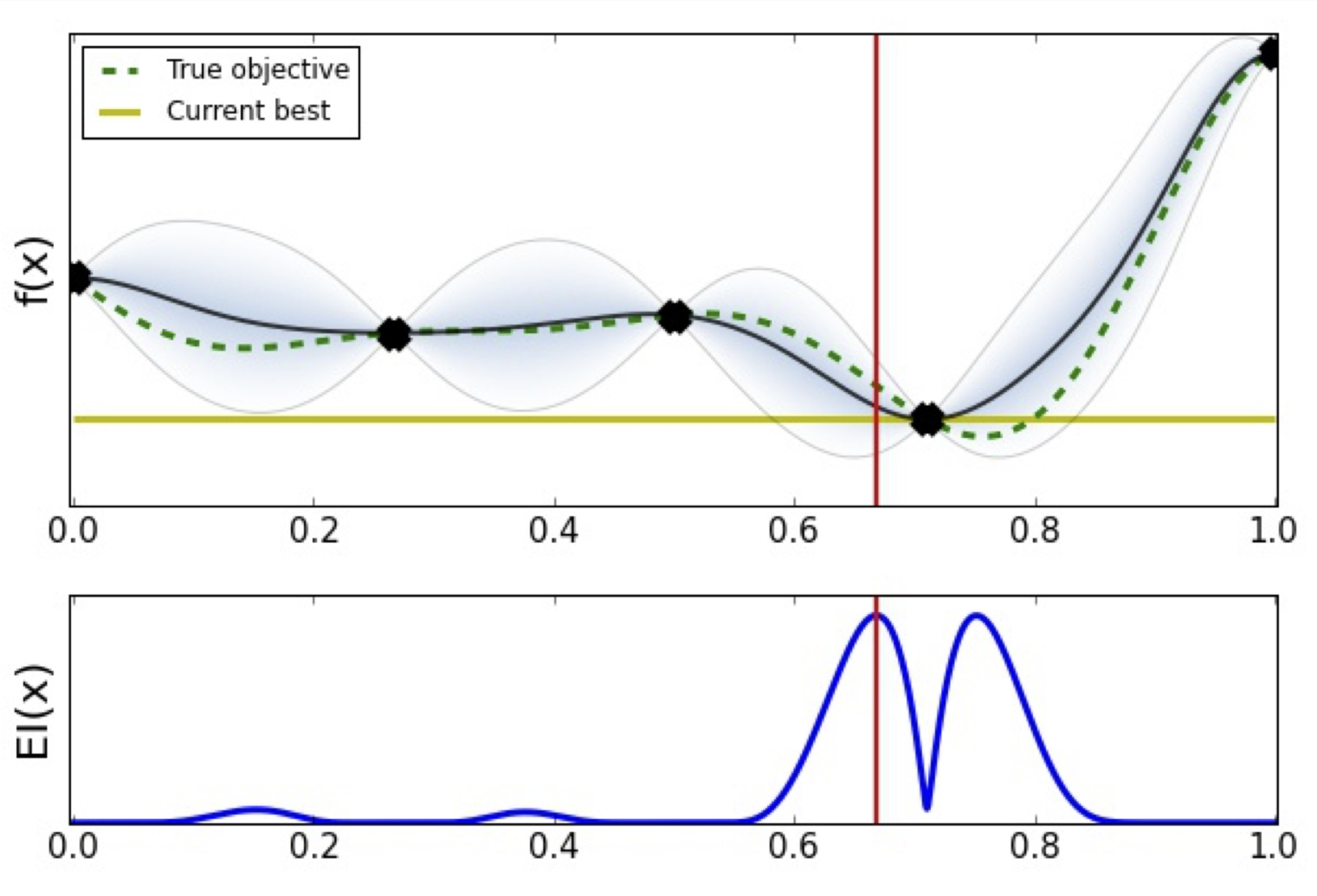
[1] Slides courtesy of Javier Gonzales

Bayesian Optimisation [1]



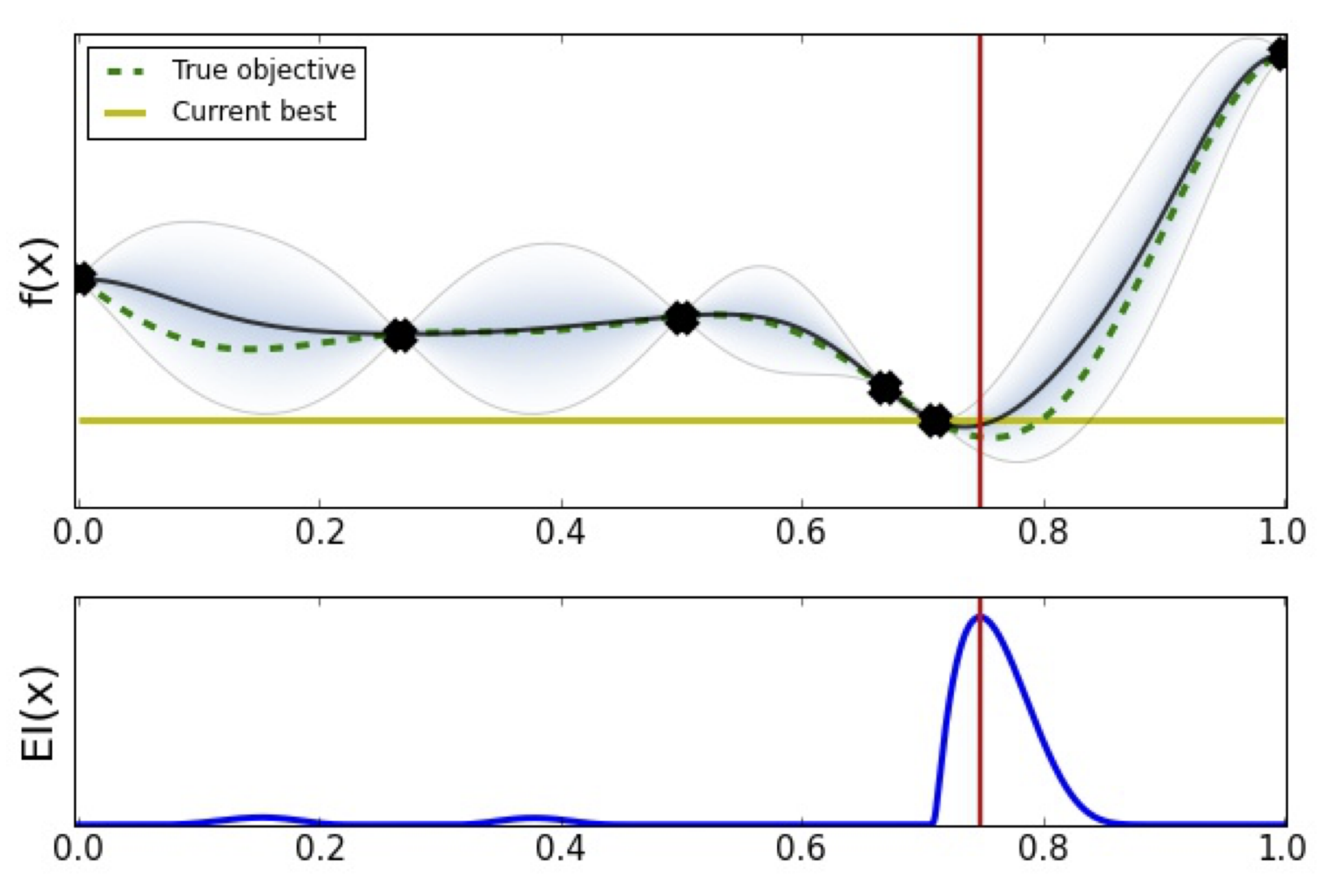
[1] Slides courtesy of Javier Gonzales

Bayesian Optimisation [1]



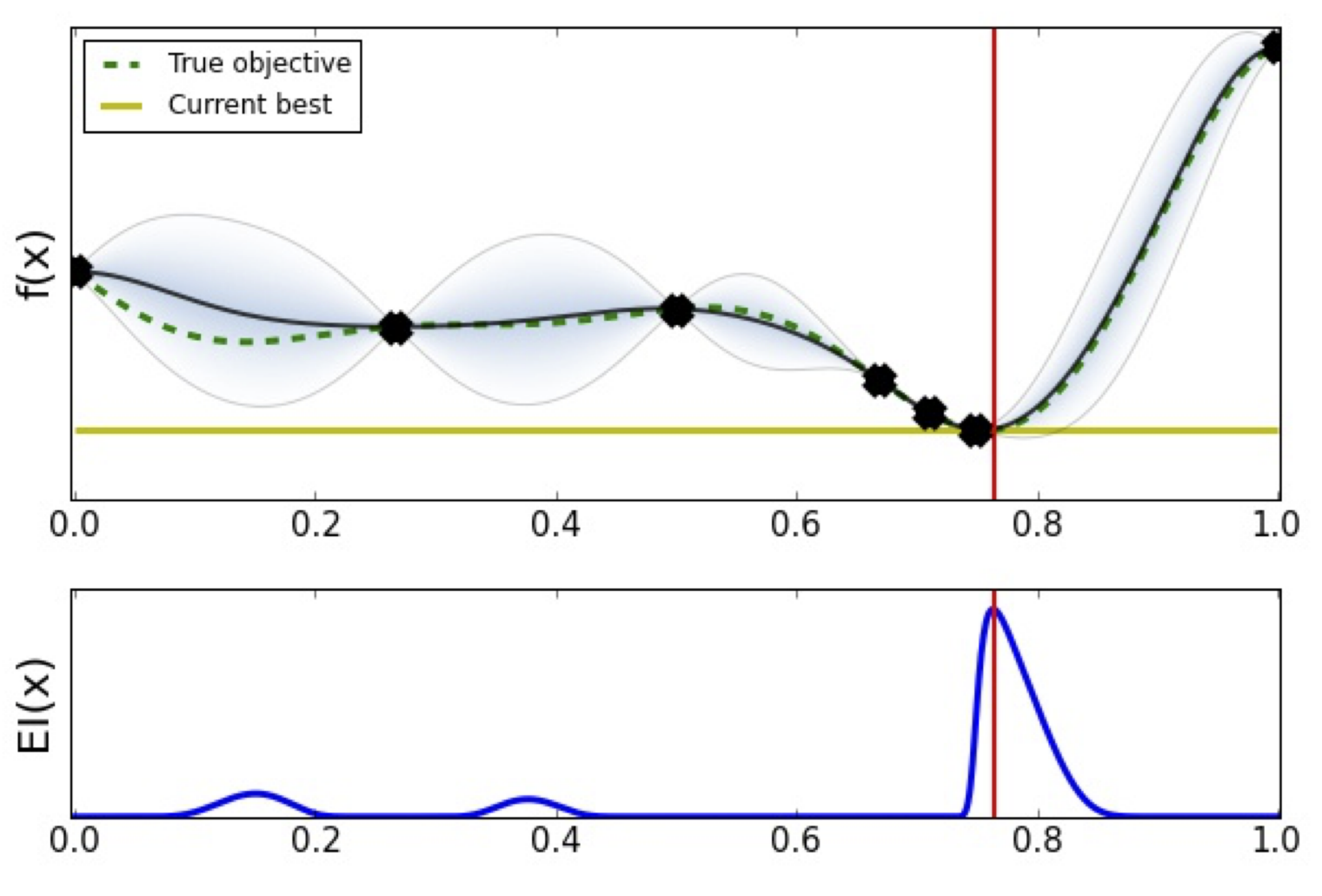
[1] Slides courtesy of Javier Gonzales

Bayesian Optimisation [1]



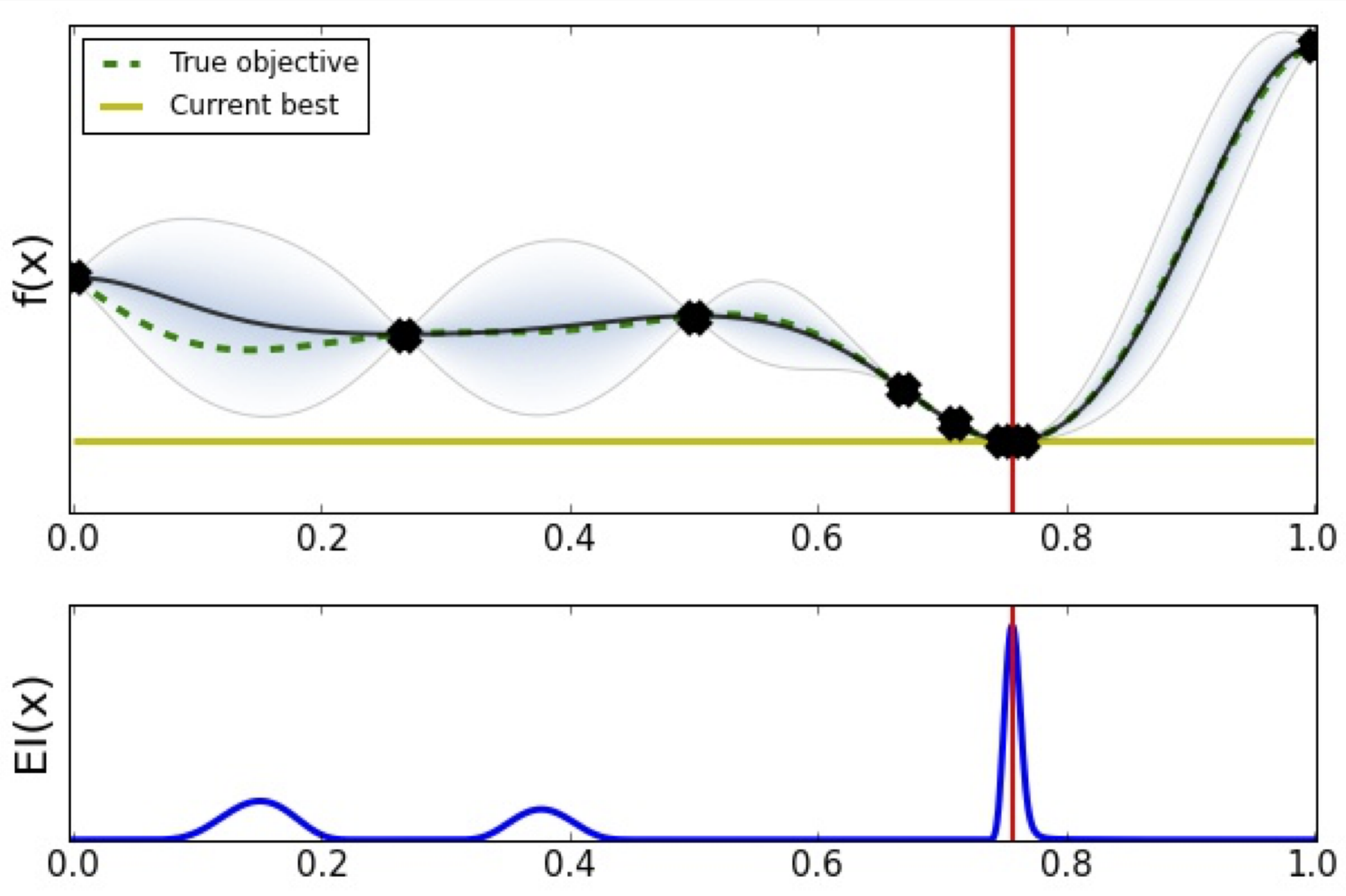
[1] Slides courtesy of Javier Gonzales

Bayesian Optimisation [1]



[1] Slides courtesy of Javier Gonzales

Bayesian Optimisation [1]



[1] Slides courtesy of Javier Gonzales

Bayesian Optmisation

- We cannot solve the direct problem

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

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

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

- 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
  - $\alpha(x)$ is cheap to compute

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
  - $\alpha(x)$ is cheap to compute
  - we can get gradients of αx

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

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.