Introduction to Gaussian Processes

Smartstart Student Retreat

July 2, 2017

Introduction to Gaussian Processes

Coding: *Sampling from a Gaussian Process*

Gaussian Process Regression

Coding: *Curve fitting with a Gaussian Process*

Bayesian Optimization

Coding: *Gaussian Processes for Global Optimization*

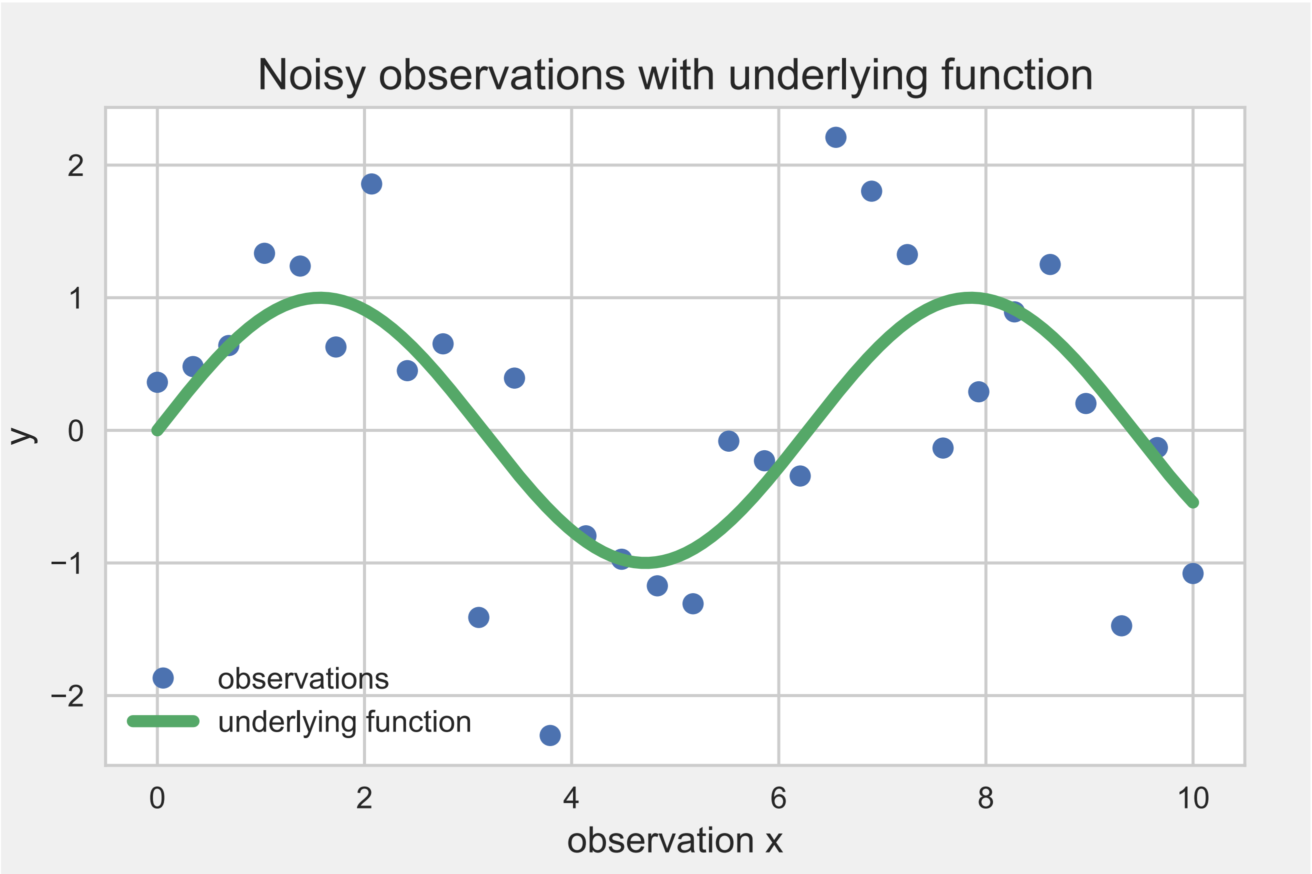
Supervised learning

- Input-output mappings from empirical data
  - robotic control, digit classification, spike sorting
- Input vector **x**, output *y*

  mapping $y = f(\mathbf{x}) + \epsilon$

  data set

  $$\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \ldots, n\}$$



Two common approaches

1. Restrict the class of functions we consider:

   linear - or quadratic functions, sum of basis functions

   Problem: How to decide for the correct class?

2. Give a prior probability to **any possible function**

   Problem: How to evaluate infinitely many functions?

Solution: **Gaussian Processes**

The Gaussian Process

- generalization of the Gaussian distribution:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- not over a scalar variable: $x \sim \mathcal{N}(\mu, \sigma)$

  or a vector, $\mathbf{x} \sim \mathcal{N}(\mu, \mathbf{\Sigma})$

  but over functions $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$

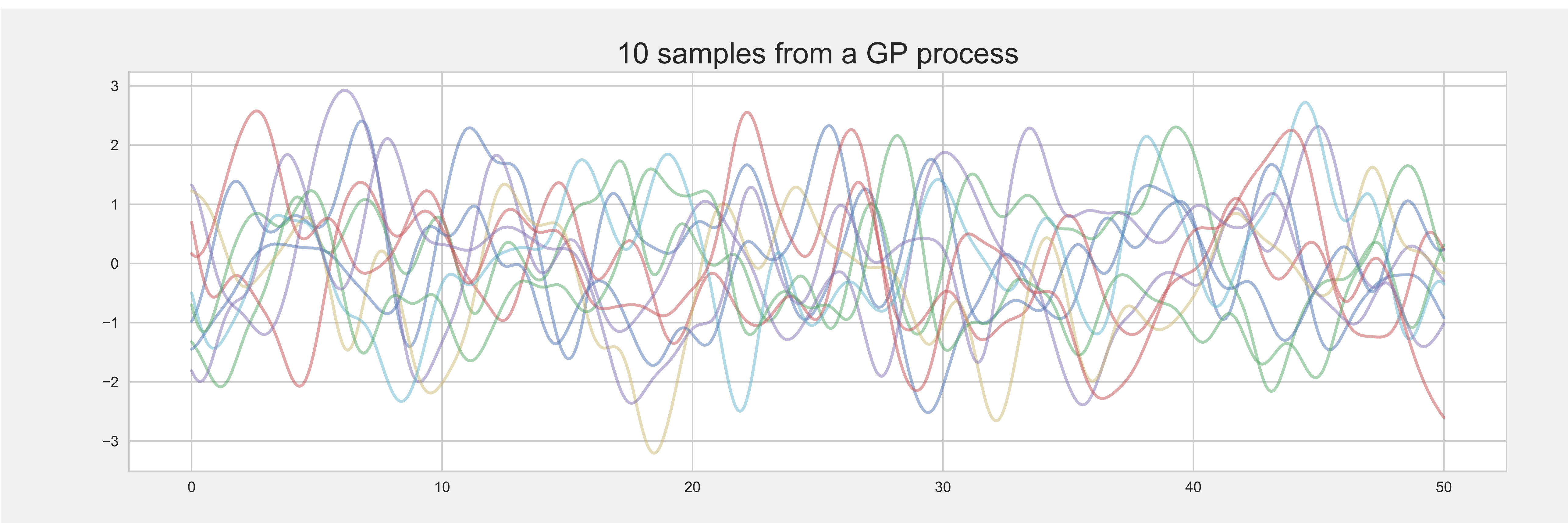
Definition:
A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

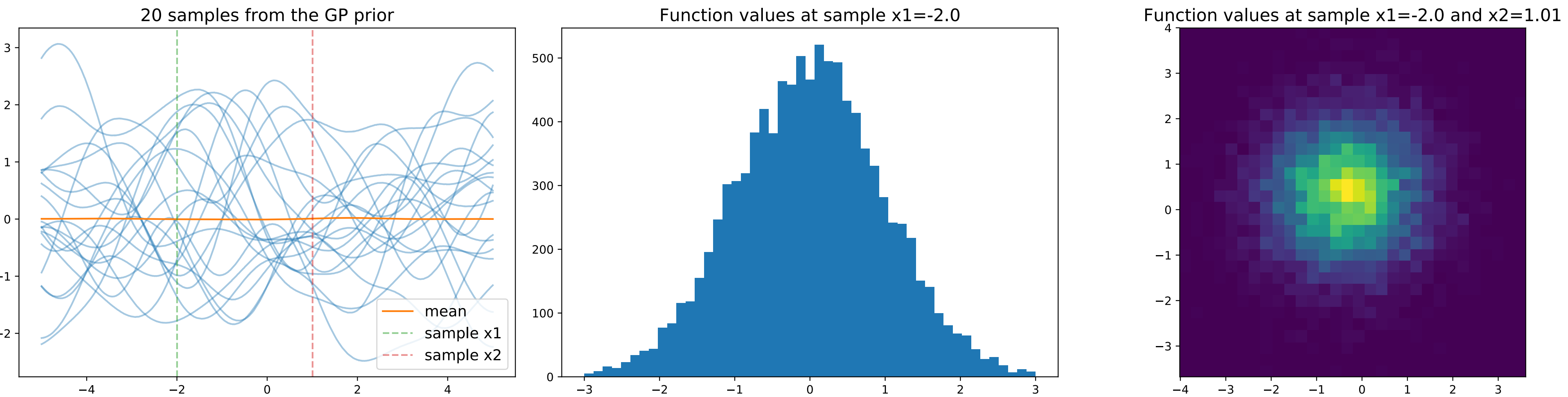
$$m(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})]$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}[(f(\mathbf{x}) - m(\mathbf{x}))(f(\mathbf{x}') - m(\mathbf{x}'))]$$

$$k(x, x') = \exp(-\frac{1}{2}|x - x'|^2)$$



Coding


20 samples from the GP prior
mean
sample x1
sample x2
Function values at sample x1=-2.0
Function values at sample x1=-2.0 and x2=1.01

Introduction to Gaussian processes

Coding: *Sampling from a Gaussian process*

Gaussian process regression

Coding: *Curve fitting with a Gaussian process*

Bayesian Optimization

Coding: *Gaussian processes for global optimization*

Bayesian linear regression

- Data set

$$\mathcal{D} = \{(\mathbf{x}_i, y_i) | i = 1, \dots, n\}$$

- Standard linear model:

$$f(\mathbf{x}) = \mathbf{x}^T \mathbf{w} \qquad y = f(\mathbf{x}) + \epsilon \qquad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

- Bayesian approach:

$$posterior = \frac{likelihood \times prior}{normalization} \qquad p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathbf{w}) p(\mathbf{w})}{p(\mathbf{y}|X)}$$

Bayesian linear regression

- Make predictions $f(\mathbf{x}_*)$ for new data $\mathbf{x}_*$

$$p(\mathbf{y}|X, \mathbf{w}) \sim \mathcal{N}(X^T\mathbf{w}, \sigma^2 I) \qquad p(\mathbf{w}) \sim \mathcal{N}(0, \Sigma)$$

$$p(\mathbf{w}|\mathbf{y}, X) = \frac{p(\mathbf{y}|X, \mathbf{w})p(\mathbf{w})}{p(\mathbf{y}|X)}$$

- The crucial Bayesian step: average over all possible parameters

$$p(f_*|\mathbf{x}_*, X, \mathbf{y}) = \int p(f_*|\mathbf{x}_*, \mathbf{w})p(\mathbf{w}|X, \mathbf{y})d\mathbf{w}$$

Gaussian process regression

- Define the linear model as a Gaussian process, noise free

$$f(\mathbf{x}) = \phi(\mathbf{x})^T \mathbf{w} \qquad\qquad p(\mathbf{w}) \sim \mathcal{N}(0, \Sigma)$$

$$f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}'))$$

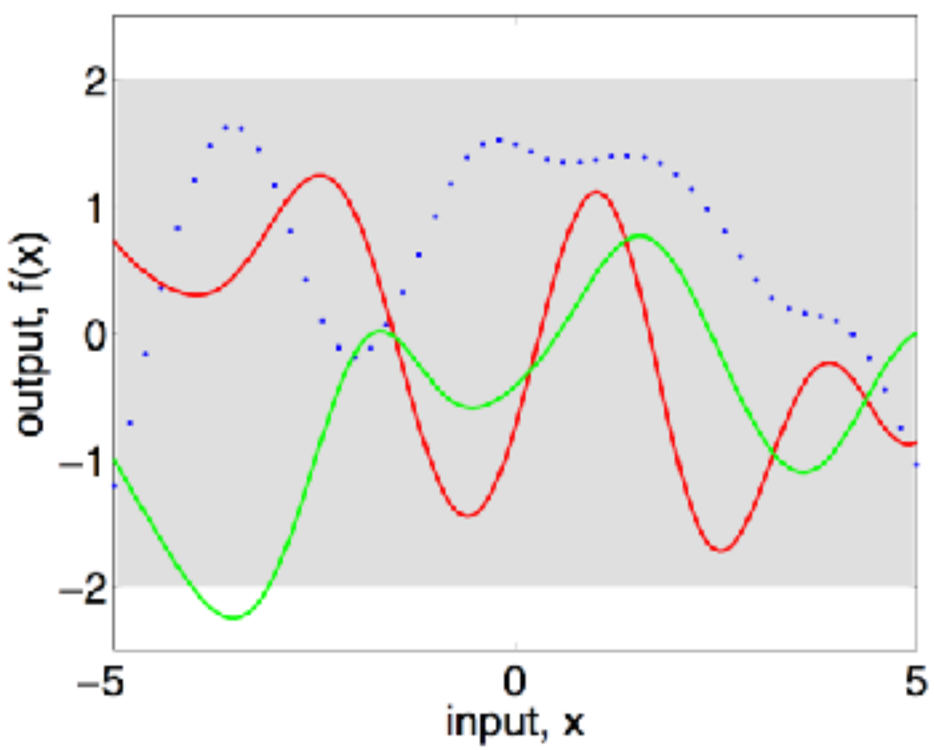
- The mean is zero: $\mathbb{E}[f(\mathbf{x})] = \phi(\mathbf{x})^T \mathbb{E}[\mathbf{w}] = 0$

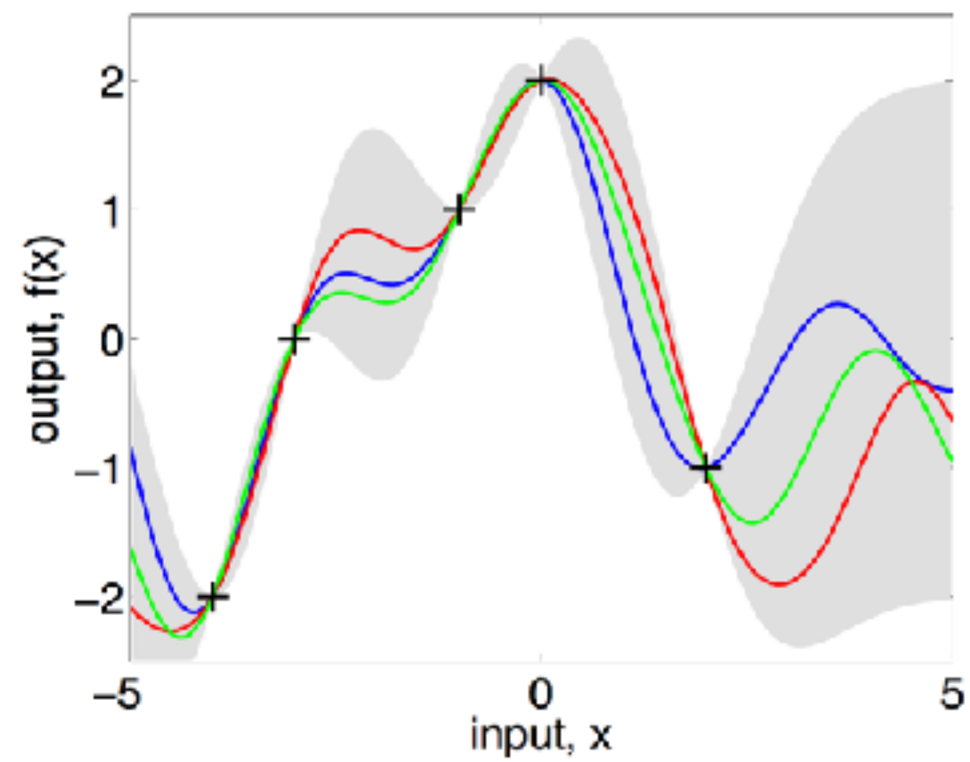
- Define a covariance function: $k(\mathbf{x}, \mathbf{x}') = \exp(-\frac{1}{2l}|\mathbf{x} - \mathbf{x}'|^2)$

- This gives a Gaussian process prior:

$$f \sim \mathcal{N}(0, K(X, X))$$

$$f_* | X_*, X, \mathbf{f}$$



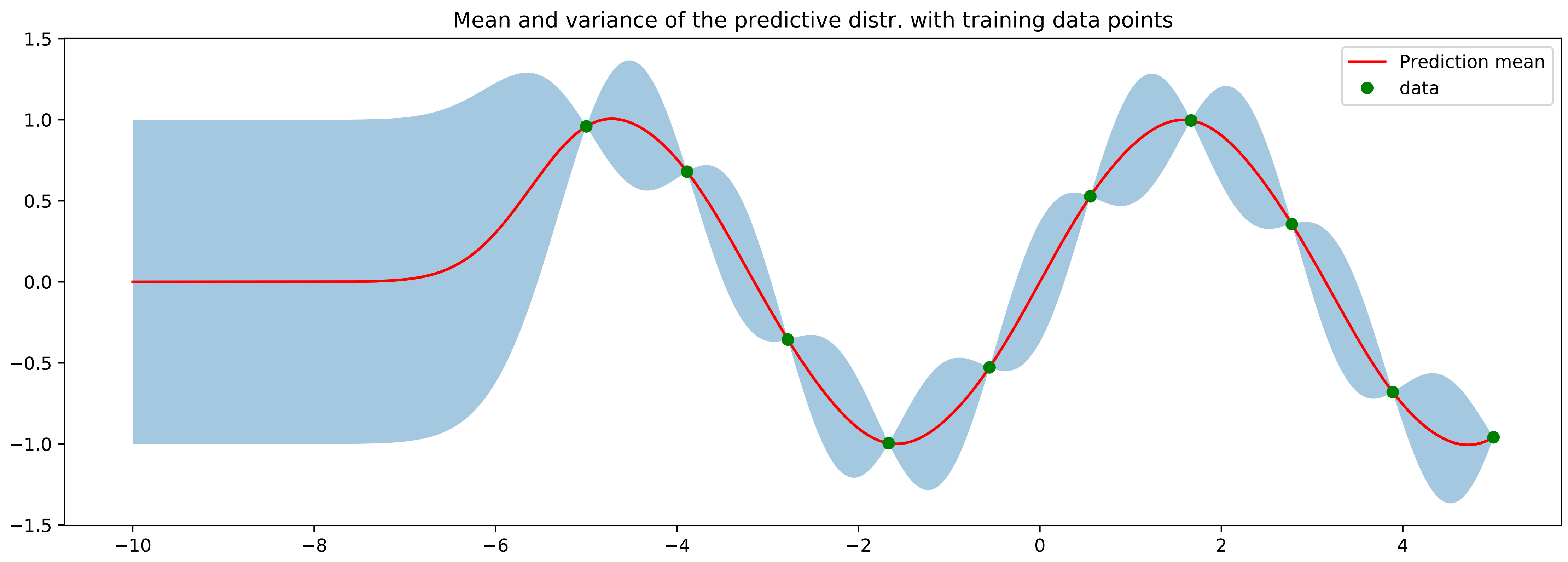


- But we want a posterior with incorporated training data

→ condition the prior on the training data

→ make predictions by evaluating the mean

Coding


Mean and variance of the predictive distr. with training data points
Prediction mean
data
1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−10
−8
−6
−4
−2
0
2
4

Introduction to Gaussian processes

Coding: *Sampling from a Gaussian process*

Gaussian process regression

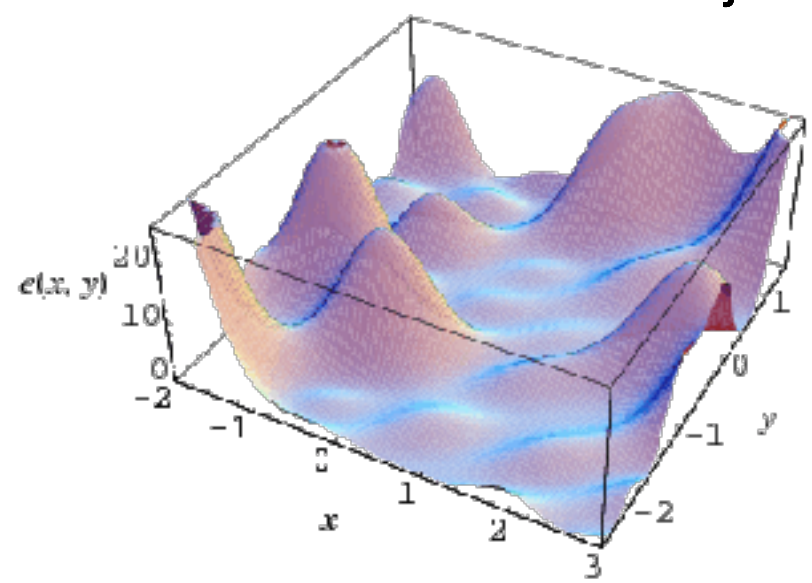
Coding: *Curve fitting with a Gaussian process*

Bayesian Optimization

Coding: *Gaussian processes for global optimization*

Bayesian Optimization

- Global optimization: find extrema of an objective function



- What if it is costly to evaluate or non-convex or unknown?
  - minimize evaluations
  - build a model of the function to evaluate effectively
  - combine prior knowledge with current evidence

- Estimate $f(x)$ given accumulated observations

$$\mathcal{D} = \{\mathbf{x}_{1:t}, f(\mathbf{x}_{1:t})\}$$

and prior information $p(f)$

to calculate the posterior over $f(x)$

$$p(f|\mathcal{D}_{1:t}) \propto p(\mathcal{D}_{1:t}|f)p(f)$$

- Use this **model** of the objective function to find the next sampling location

- What kind of model…? **Gaussian Process**

GP for Bayesian optimization

- Define a GP with prior information in the covariance fun

$$f \sim \mathcal{N}(0, K(X, X))$$

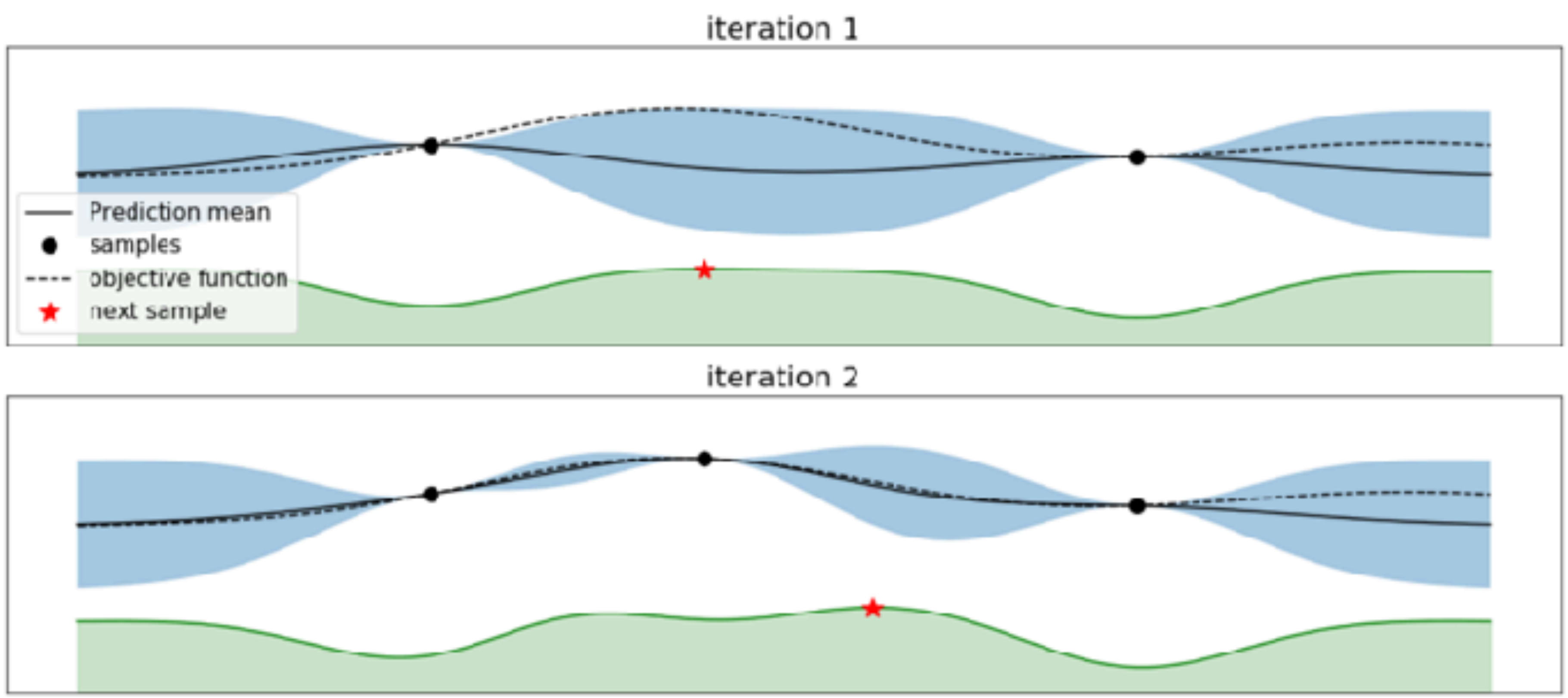
- Define an acquisition function to decide where to sample:

$$\mathcal{U}(\mathbf{x}) = \mu(\mathbf{x}) + \kappa \cdot \sigma(\mathbf{x})$$

Algorithm 1 Bayesian optimization

1: **for** $t = 1, 2, \ldots$ **do**
2: Find $\mathbf{x}_t$ by maximizing the acquisition function: $\mathbf{x}_t = \mathrm{argmax}_{\mathbf{x}} \mathcal{U}(\mathbf{x}|\mathcal{D}_{1:t-1})$.
3: Sample the objective function at the new location: $\mathbf{y}_t = f(\mathbf{x}_t)$.
4: Add the new data point to $\mathcal{D}$ and update the GP model.
5: **end for**

Coding


iteration 1
Prediction mean
samples
objective function
next sample
iteration 2

Thank you!

References:

Rasmussen and Williams. *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).* The MIT Press, 2005.

Resources:

Book link: http://www.gaussianprocess.org/gpml/

Podcast: http://www.thetalkingmachines.com/blog/2016/1/28/openai-and-gaussian-processes

Lecture: https://www.youtube.com/watch?v=4vGiHC35j9s