Gaussian Process Models

AE 6310: Optimization for the Design of Engineered Systems

Spring 2017

Dr. Glenn Lightsey

Lecture Notes Developed By Dr. Brian German

Georgia Institute of Technology

Daniel Guggenheim School of Aerospace Engineering

Primary References

- *Gaussian Processes for Machine Learning* by Rasmussen & Williams
  - Contains nearly everything you could want to know about Gaussian process models
  - It's FREE!!!!
  - Download: http://www.gaussianprocess.org/gpml/

- *Efficient Global Optimization of Expensive Black-box Functions* by Jones et. al.
  - 1998 Journal of Global Optimization

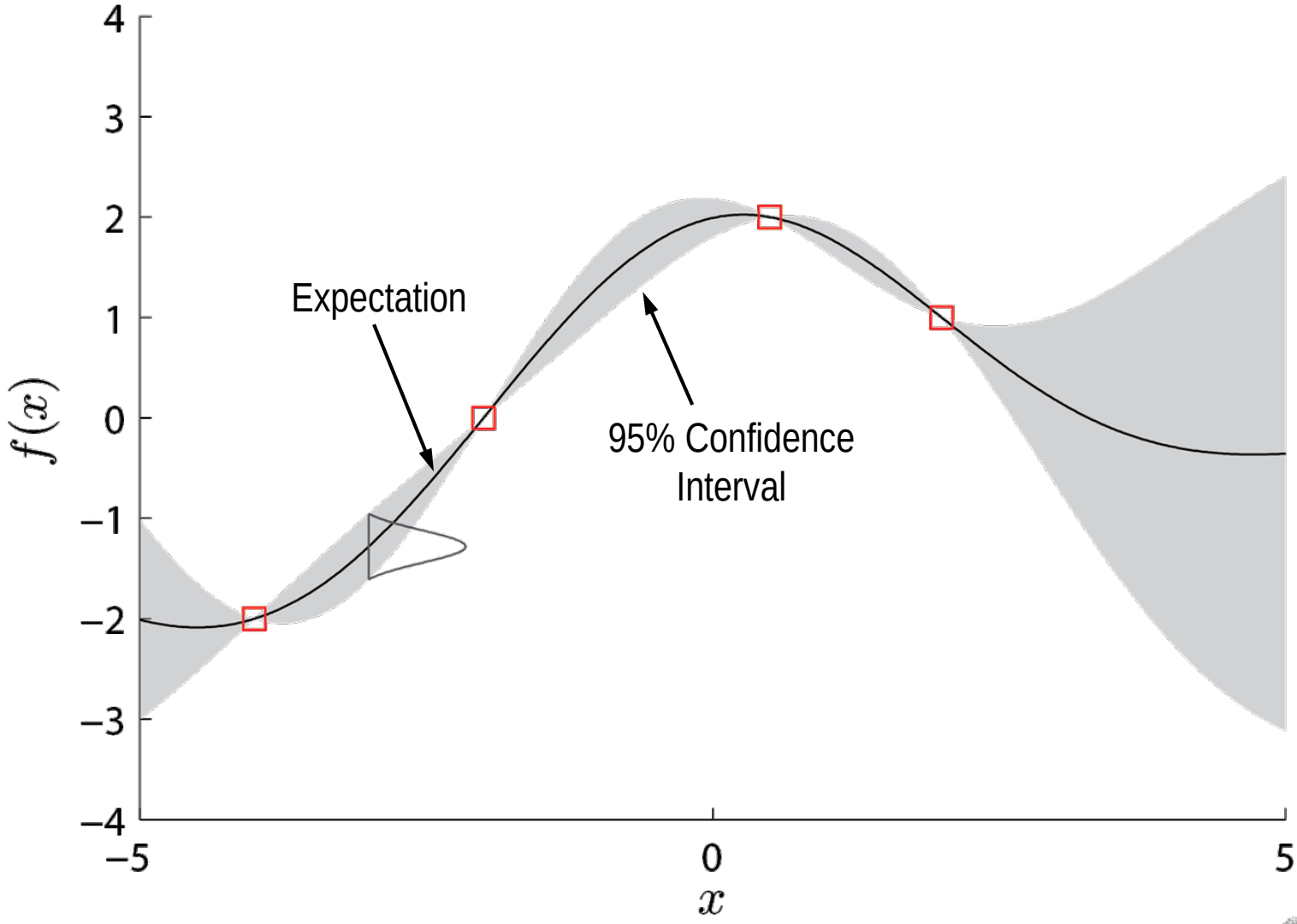
Also Known As…

- Gaussian process models
  - This name is primarily used in the statistical and robotics communities
- Kriging
  - This name is primarily used in the optimization and geostatistics communities
- DACE stochastic process models
  - DACE is an acronym for “Design and Analysis of Computer Experiments”

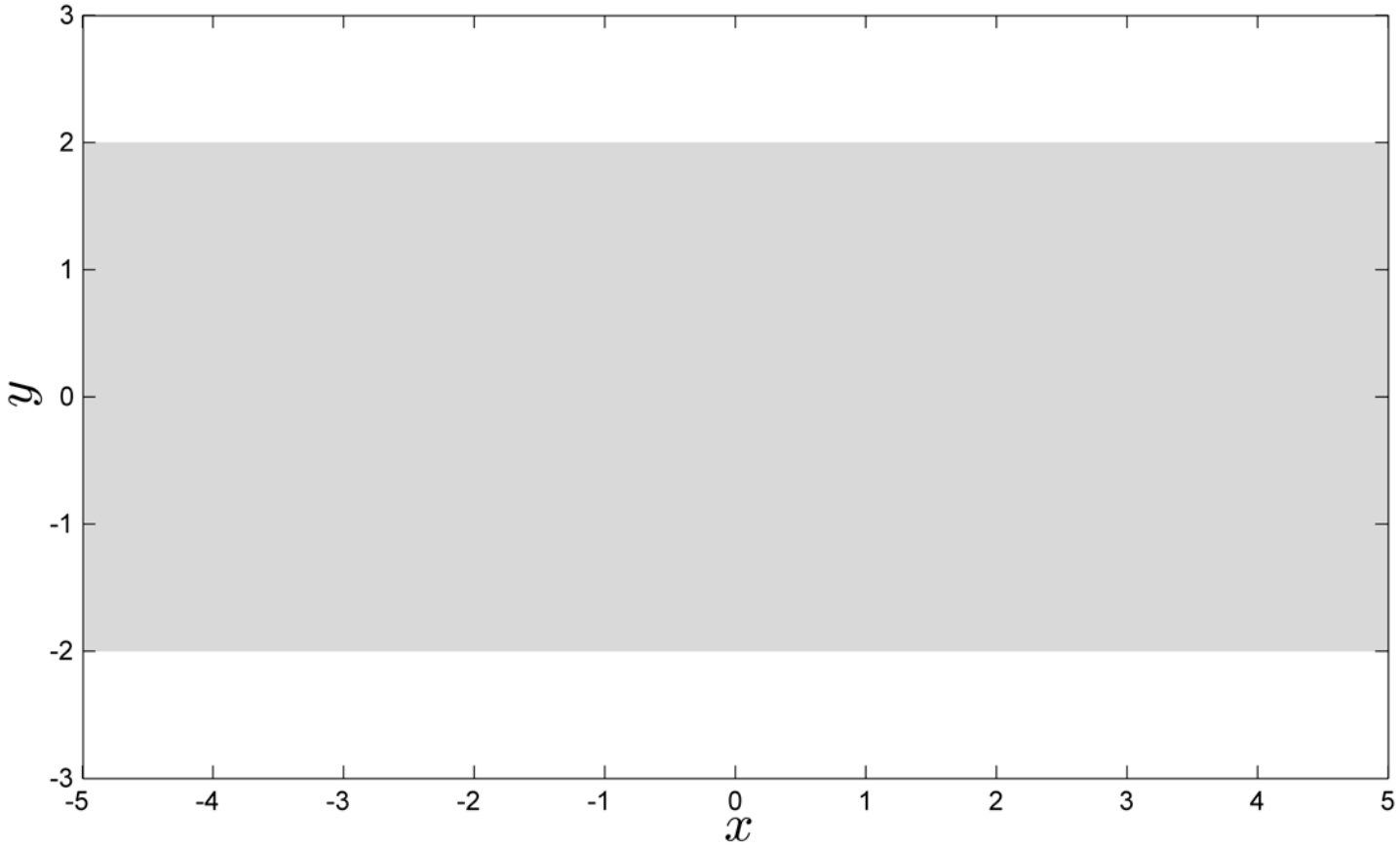
What are Gaussian process models?

- Used for regression
  - Infers the mapping between variables
- Used for classification
  - Is it a car or a truck?
- Supervised learning
  - Uses training data to infer a function
- Non-parametric
  - The model does not assume a functional form
- Stochastic processes
  - Returns a probability distribution for the function value rather than a deterministic value

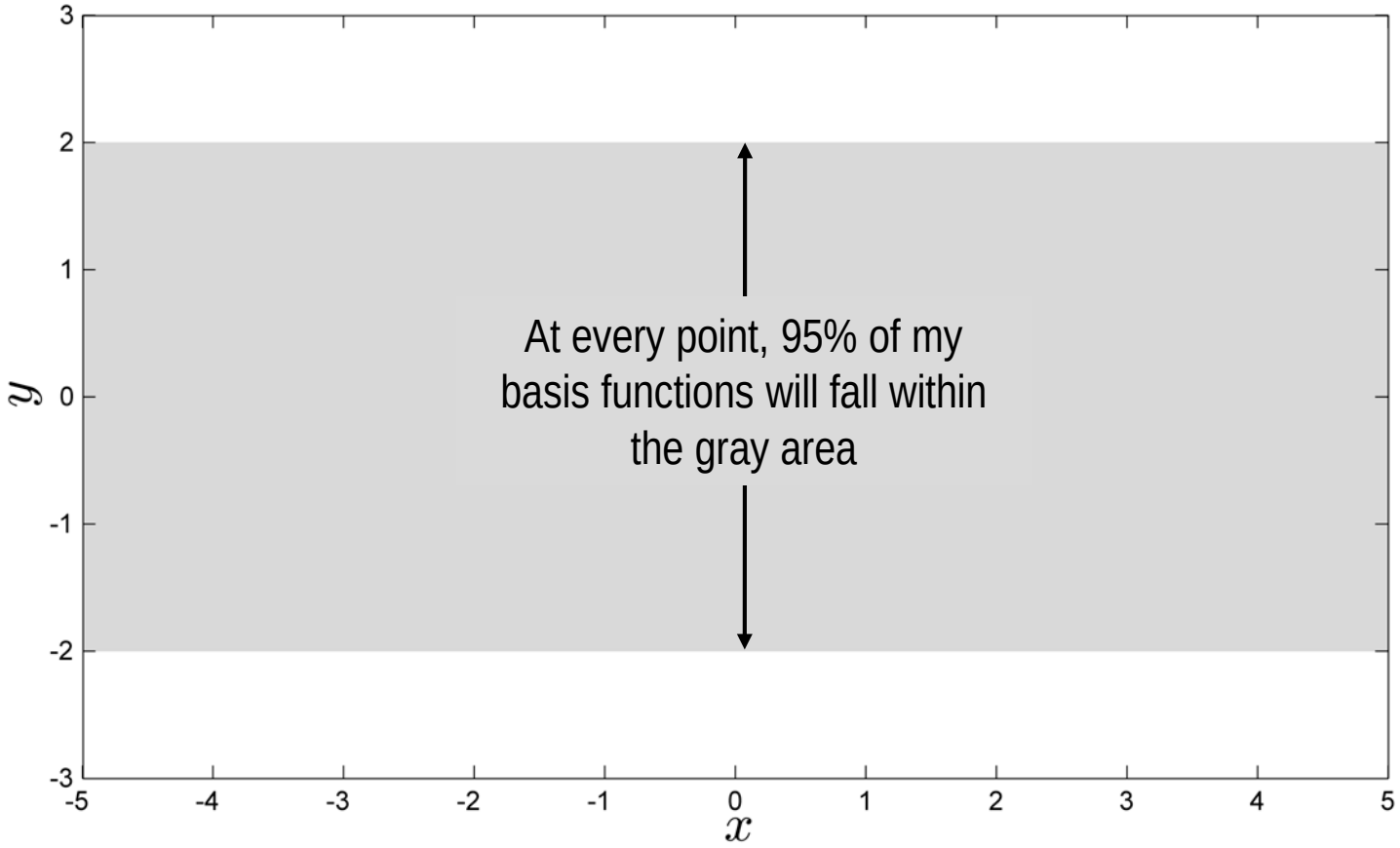
What are Gaussian process models?



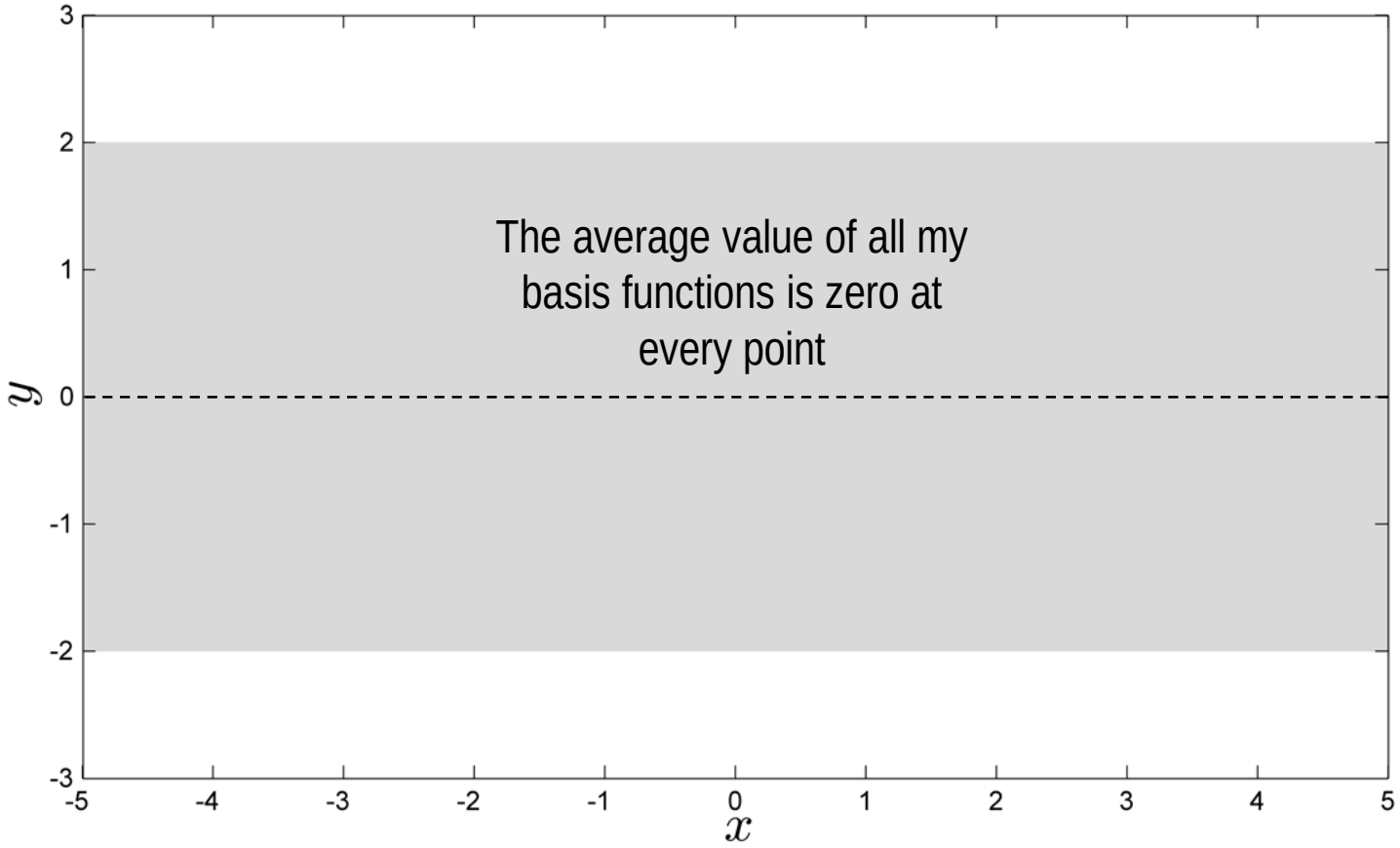
Prior Distribution



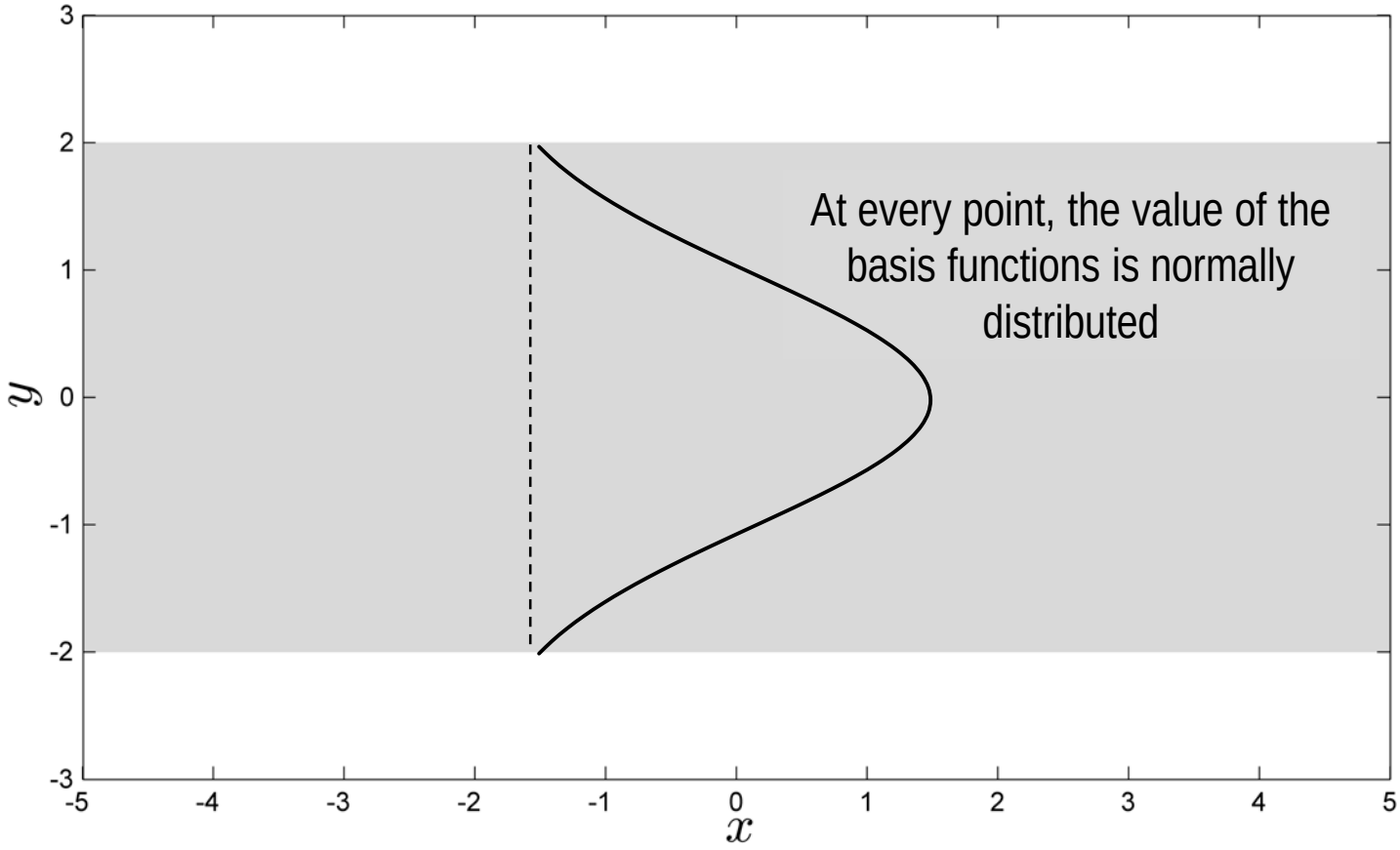
Prior Distribution



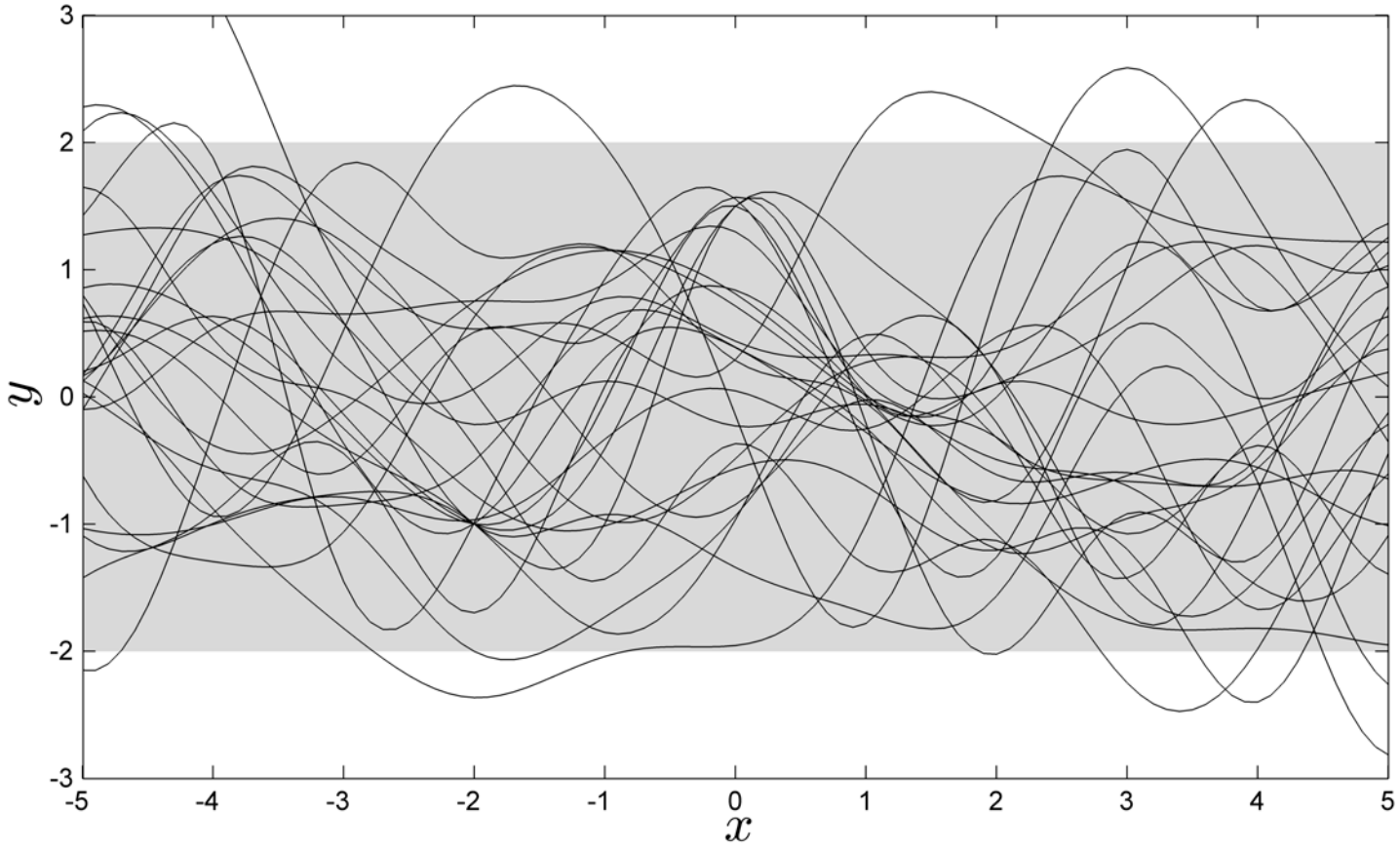
Prior Distribution



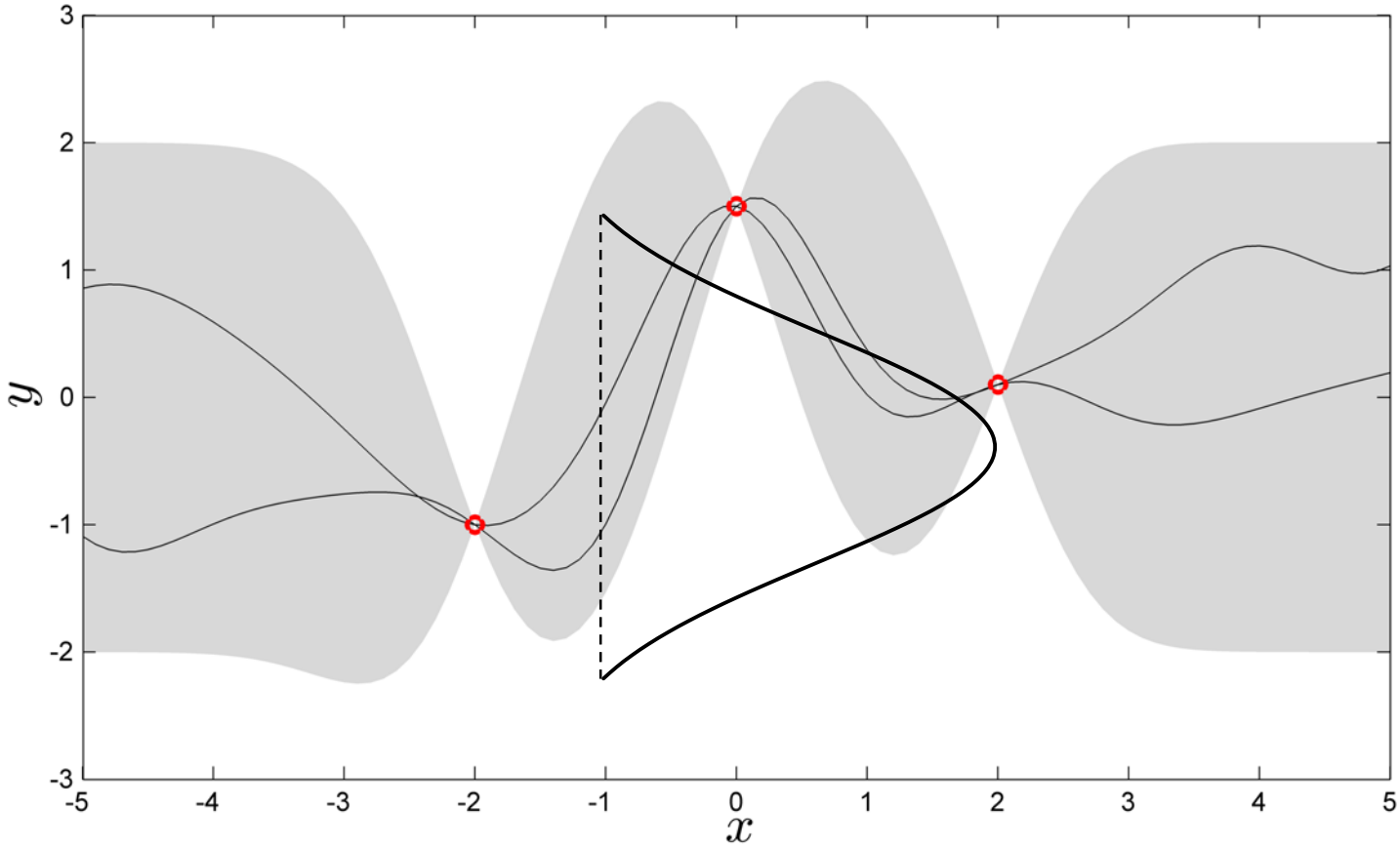
Prior Distribution



Prior Distribution



Posterior Distribution



What's happening?

- The prior distribution is determined entirely by a:
  - Mean function (we will assume this is zero for now)
  - Covariance Function

- Suppose we have n m-dimensional training points and p test points
  - X_{tr} is an m x n matrix of training points
  - Y_{tr} is a row vector of length n
  - X is an m x p matrix of test (prediction) points (we want to know the function value at X)

Covariance Function

❖ Covariance Function:

- Loosely speaking, it describes the relationship between the *function values* of two points based on the *coordinate locations*
- In the earlier example, I used the *squared exponential covariance function:*

$$k(x, x^*) = \sigma^2 \exp\left(-\frac{(x - x^*)^2}{2l^2}\right)$$

- We can create a covariance matrix, *K*, between two sets of points, X and X *
- The (i^{th},j^{th}) term of K equals the covariance of the i^{th} point in X and the j^{th} point X *

Calculating Distributions

- We assume the basis functions are distributed as follows:

$$\begin{bmatrix} Y \\ Y_{tr} \end{bmatrix} \sim \mathcal{N}\left(0, \begin{bmatrix} K(X,X) & K(X,X_{tr}) \\ K(X_{tr},X) & K(X_{tr},X_{tr}) \end{bmatrix}\right)$$

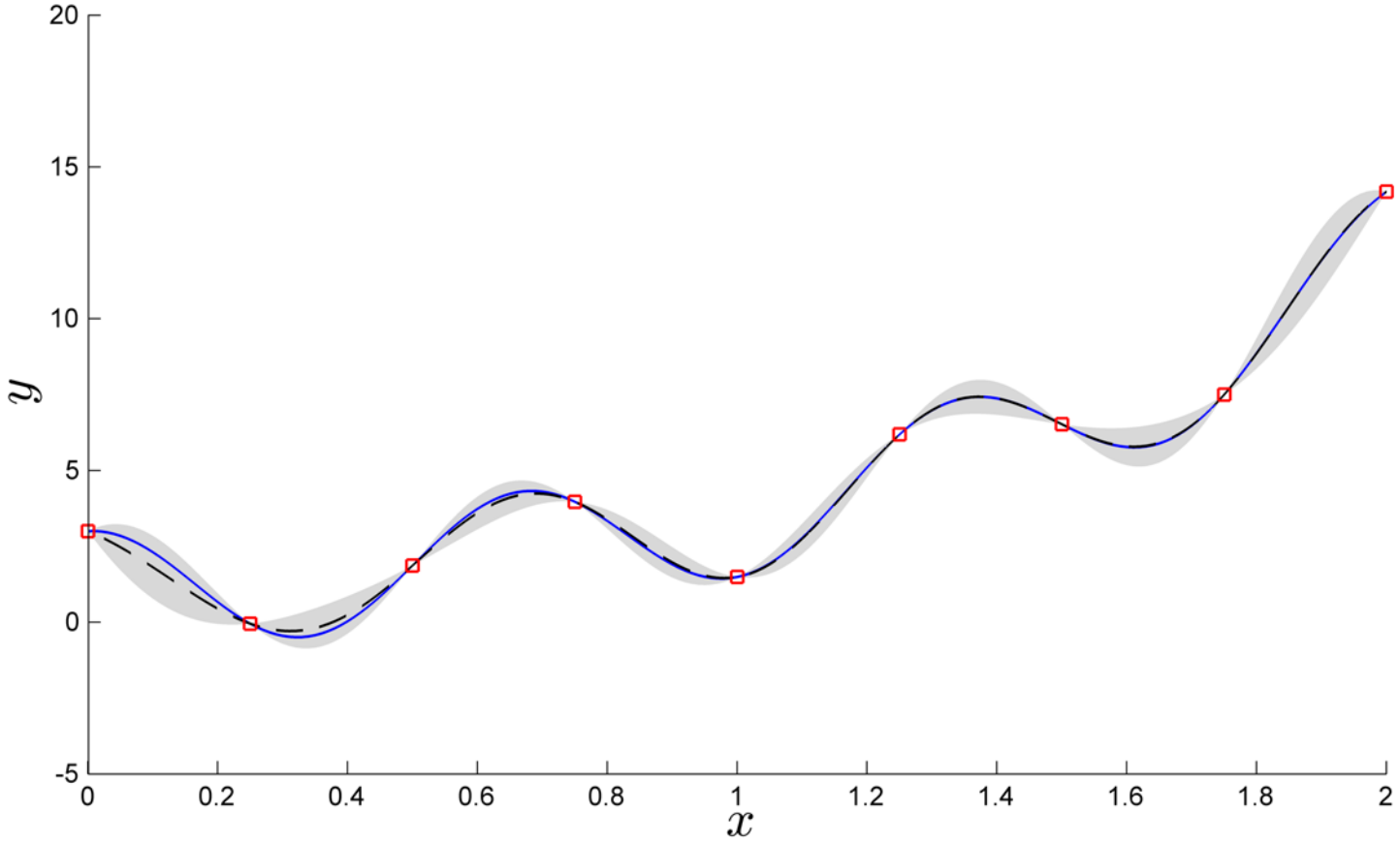
- Using Bayes' Theorem, we can solve for:

$$P(Y|X, X_{tr}, Y_{tr})$$

- The result:

$$Y|X, X_{tr}, Y_{tr} \sim \mathcal{N}\Big(Y_{tr}K(X_{tr},X_{tr})^{-1}K(X_{tr},X), \\ K(X,X) - K(X,X_{tr})K(X_{tr},X_{tr})^{-1}K(X_{tr},X)\Big)$$

Example Gaussian Process



Covariance Functions

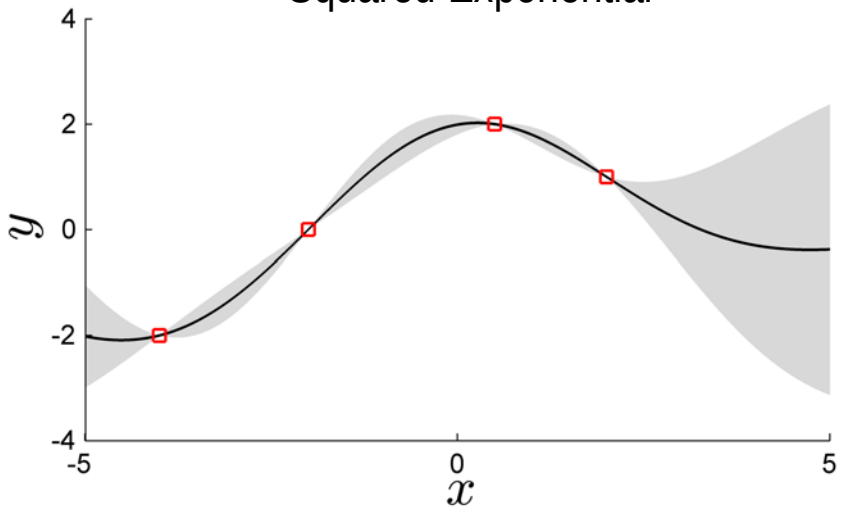
- Squared Exponential

$$k(x, x^*) = \sigma^2 \exp\left(-\frac{(x - x^*)^2}{2l^2}\right)$$

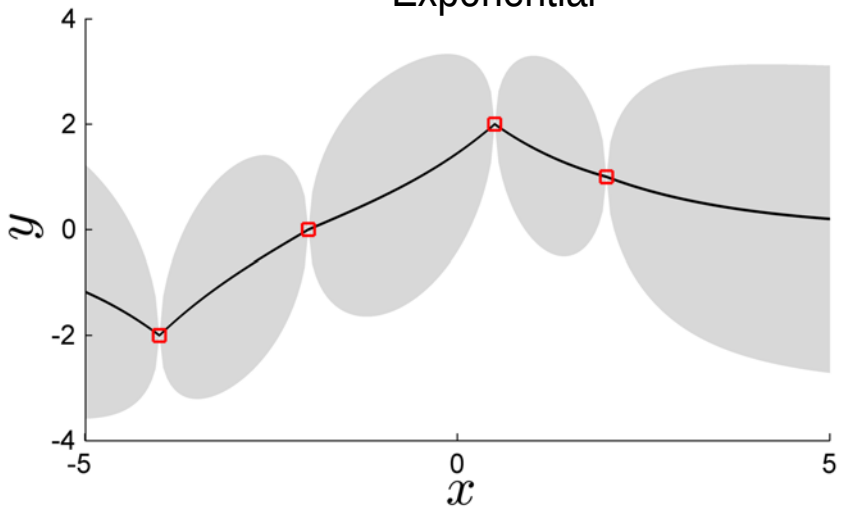
- Assumed Properties:
  - Stationary: function of only the distance between the two points (invariant to translations)
  - Infinitely Differentiable: implies the underlying function is continuous and infinitely differentiable

Other Covariance Functions

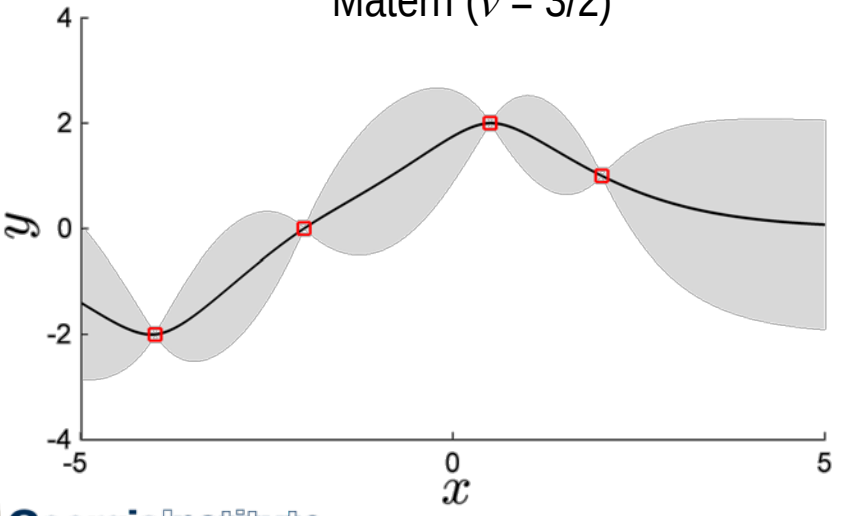
Squared Exponential



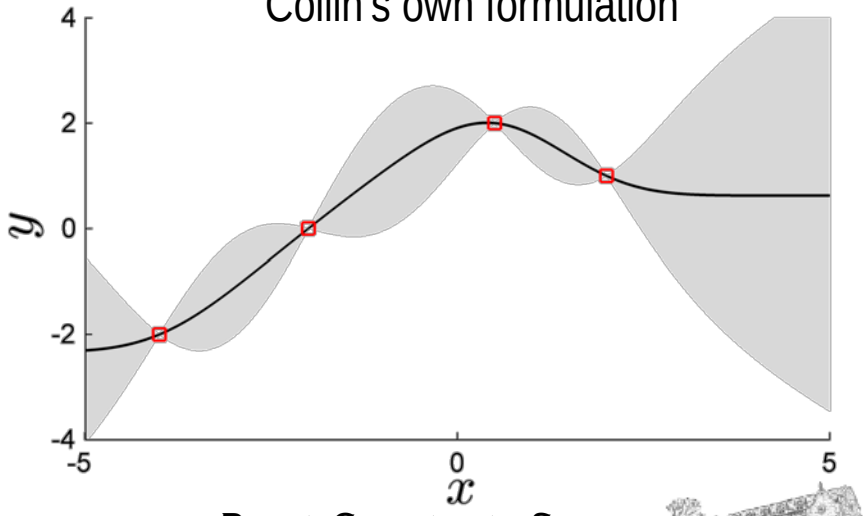
Exponential



Matern (ν = 3/2)



Collin's own formulation

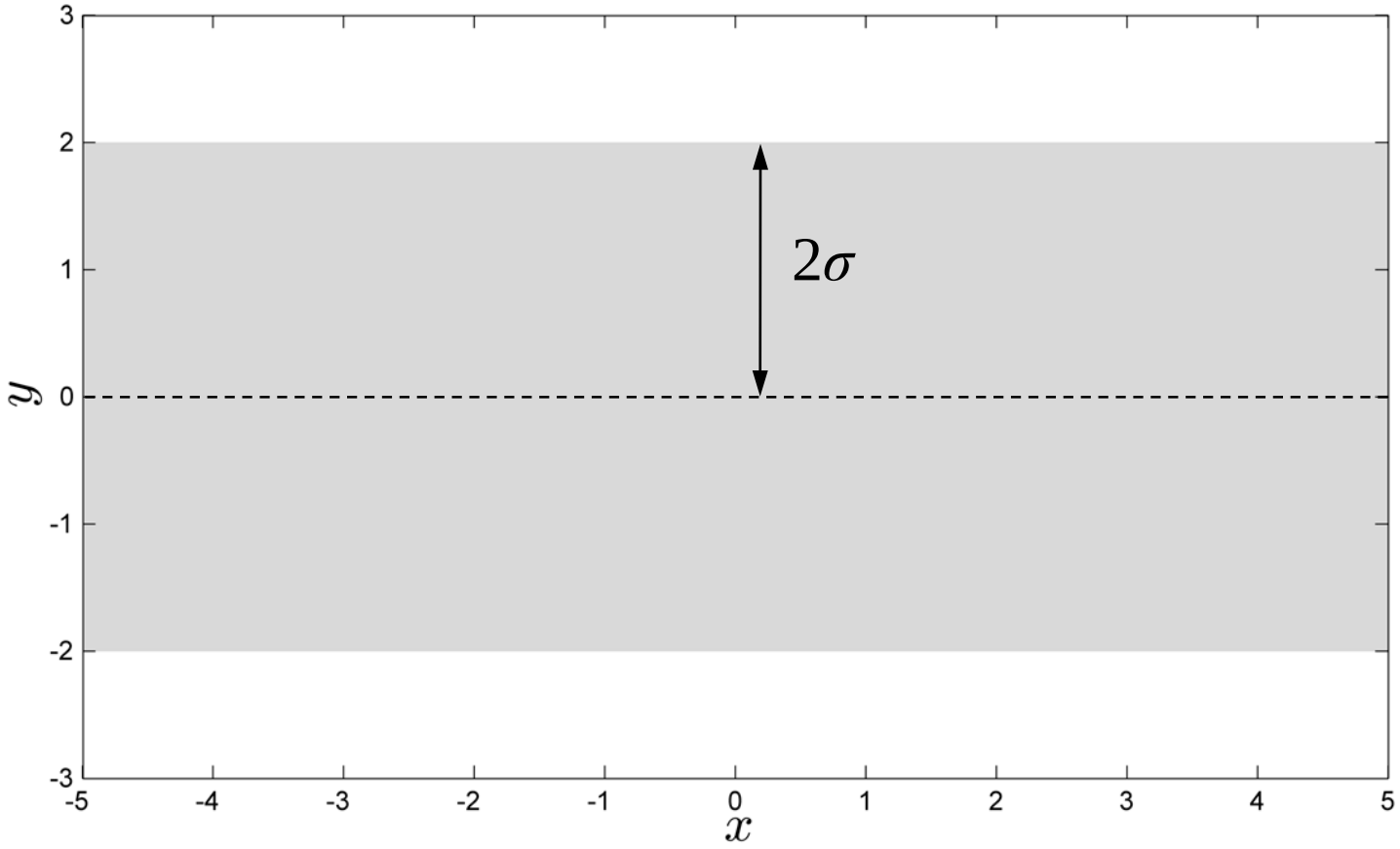


Hyperparameters

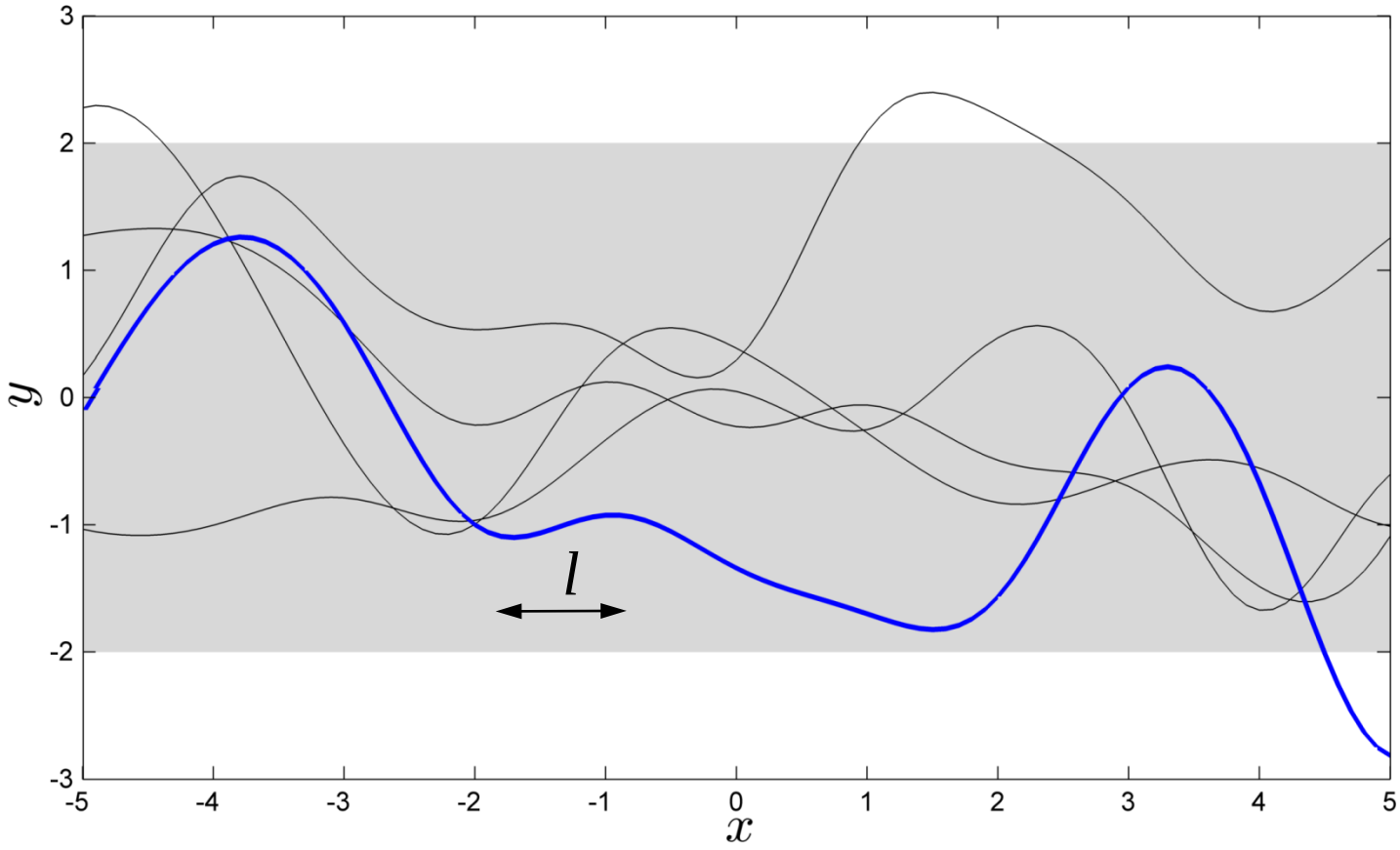
$$k(x, x^*) = \sigma^2 \exp\left(-\frac{(x - x^*)^2}{2l^2}\right)$$

- What are σ and l?
  - Hyperparameters: they affect the shape of the Gaussian process model

Hyperparameters



Hyperparameters



Hyperparameters

$$k(x, x^*) = \sigma^2 \exp\left(-\frac{(x - x^*)^2}{2l^2}\right)$$

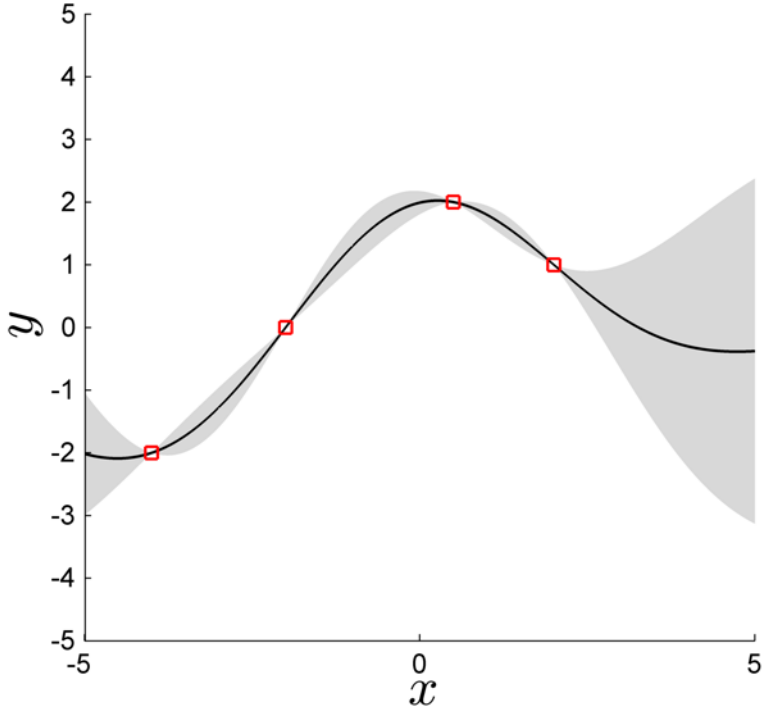
- What are σ and l?
  - Hyperparameters: they affect the shape of the Gaussian process model
- How do we determine their values?
  - Maximize ***marginal likelihood*** : $p(Y_{tr}|X_{tr}, \Theta)$
  - Most people minimize the negative log transform:

Hyperparameters

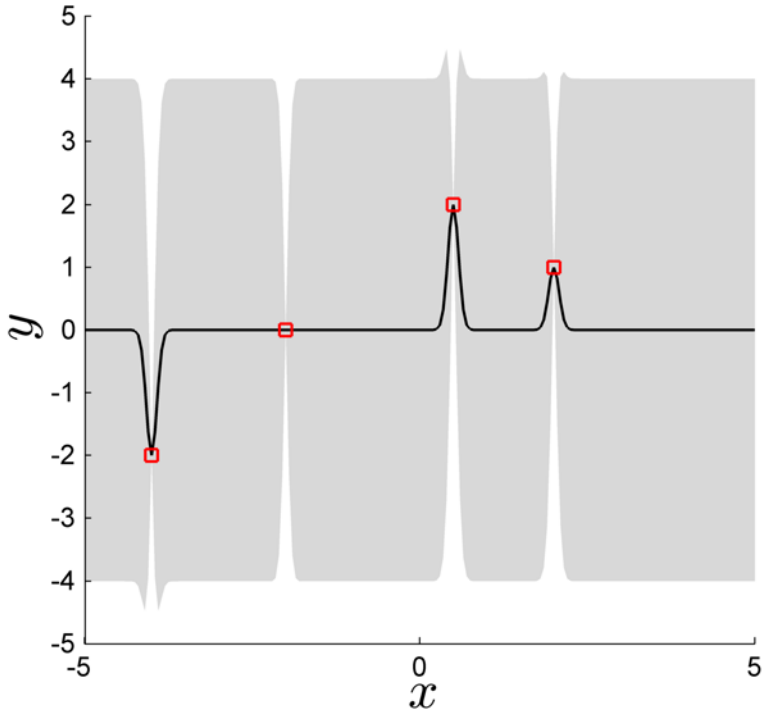
$$-\log p(Y_{tr}|X_{tr}, \Theta) = \frac{1}{2}\log(\det(K(X_{tr}, X_{tr}))) - \frac{1}{2}Y_{tr}K(X_{tr}, X_{tr})^{-1}Y_{tr}^T - \frac{n}{2}\log 2\pi$$

Hyperparameters

Good Hyperparameters



Bad Hyperparameters



What about the mean?

- If $\mu = 0$ is a bad assumption, there are other options:
  - Simple Kriging
    - Mean is zero
  - Ordinary Kriging
    - Mean is stationary but unknown
    - Maximum likelihood estimator is often used
  - Universal Kriging
    - Mean is assumed to be a polynomial
  - Many other methods exist…

Gaussian Processes in Optimization

- Efficient Global Optimization (EGO):
  - "Warm start:" DOE is performed on the design space to obtain training points
  - Gaussian process model is fit to the training data
  - The next sample point is chosen by maximizing ***expected improvement***
  - After each iteration, we refit the Gaussian process model
  - Branch-and-bound methods exist that guarantee finding the global maximum of expected improvement
  - Intuitively balances *exploration* and *exploitation*

Expected Improvement

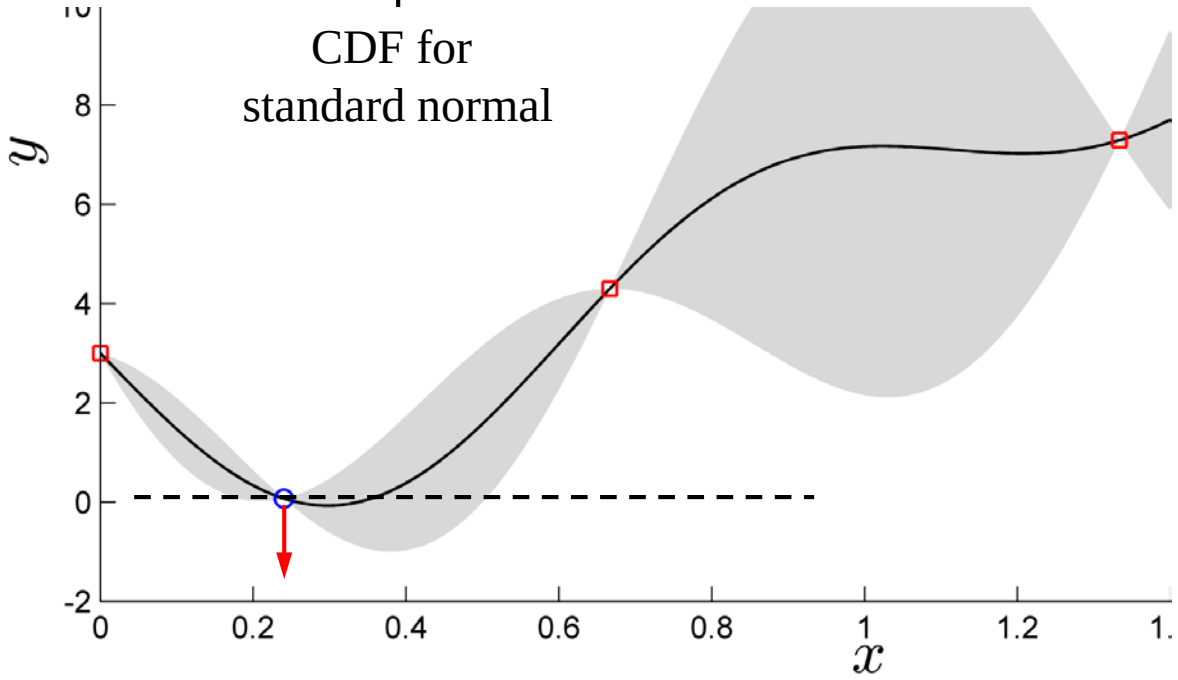
- Expected improvement:

PDF for standard normal

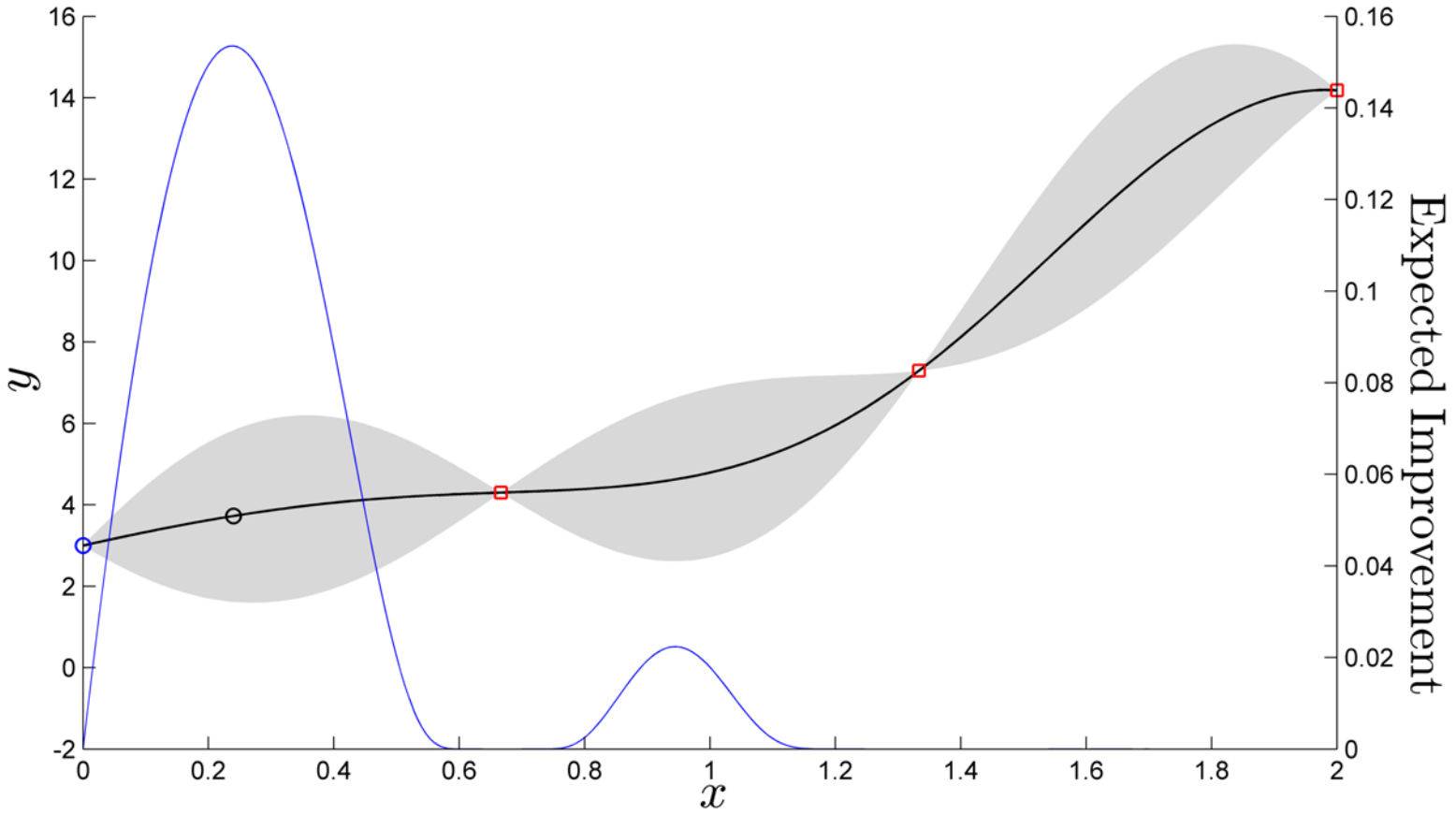
$$\mathrm{E}[I(x)] = \mathrm{E}[\max(f_{\min} - Y, 0)]$$

$$= (f_{\min} - \mu)\Phi\left(\frac{f_{\min} - \mu}{\sigma}\right) + \sigma\phi\left(\frac{f_{\min} - \mu}{\sigma}\right)$$

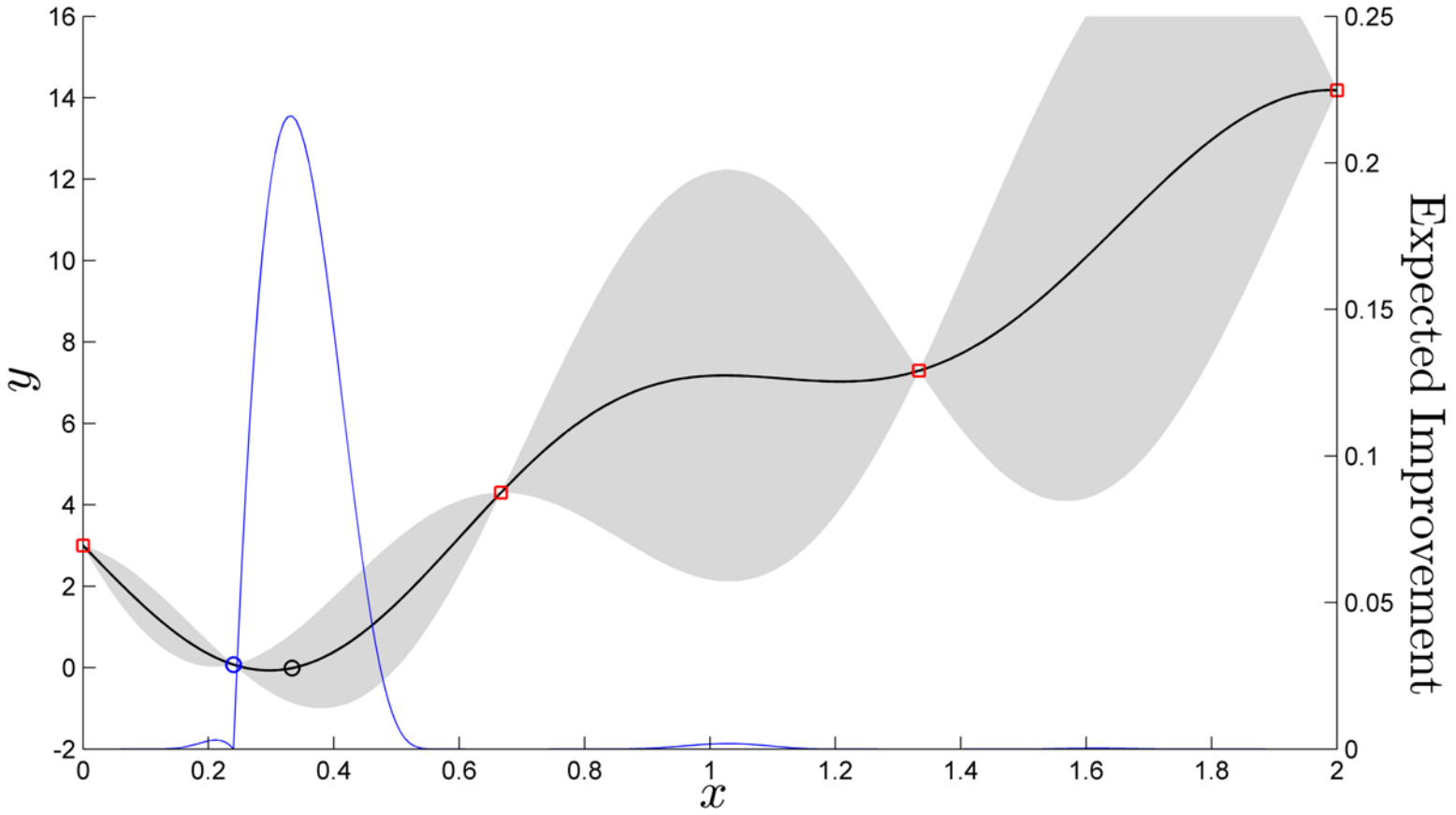
CDF for standard normal



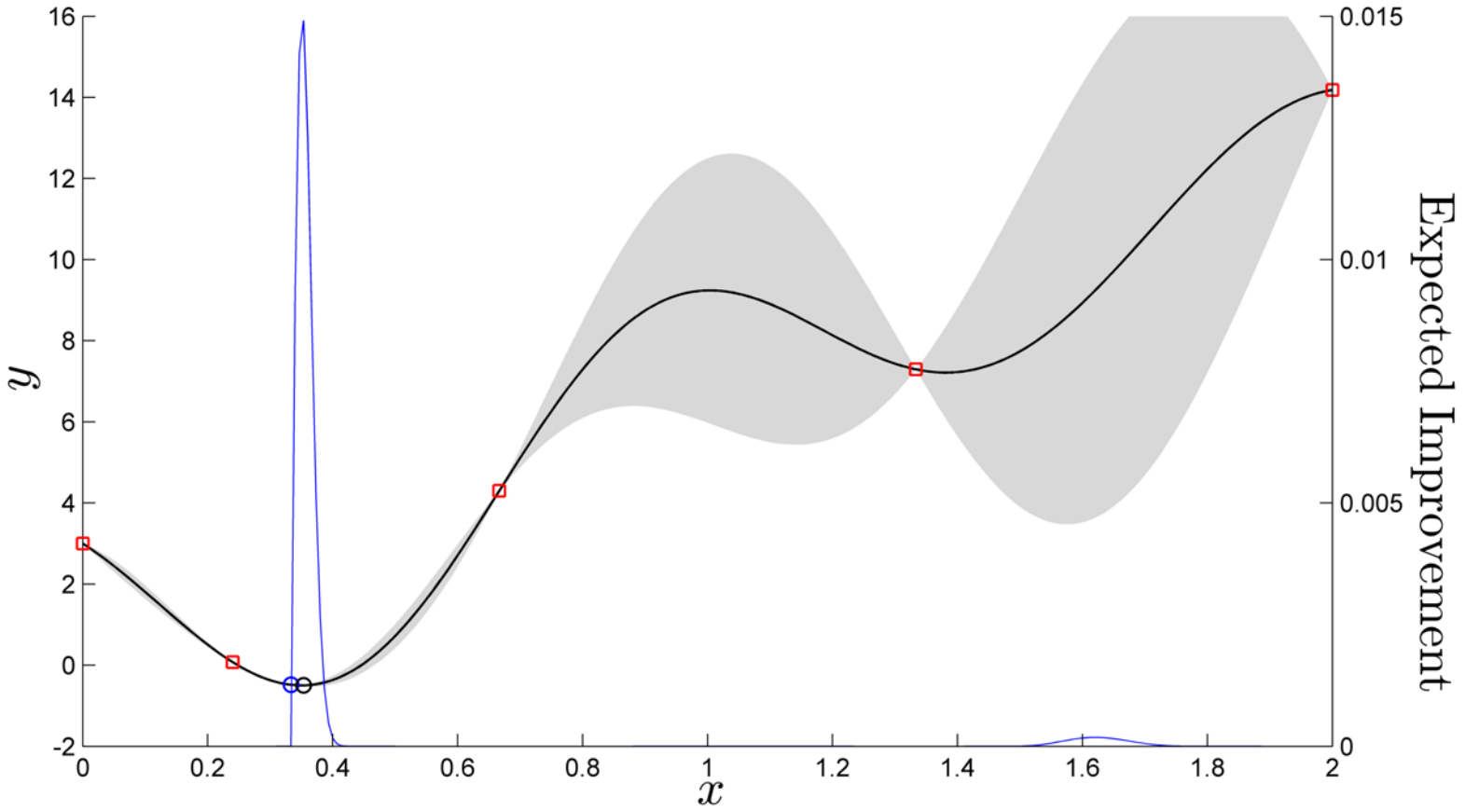
EGO Example


16
14
12
10
8
6
4
2
0
-2
y
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x
0.16
0.14
0.12
0.1
0.08
0.06
0.04
0.02
0
Expected Improvement

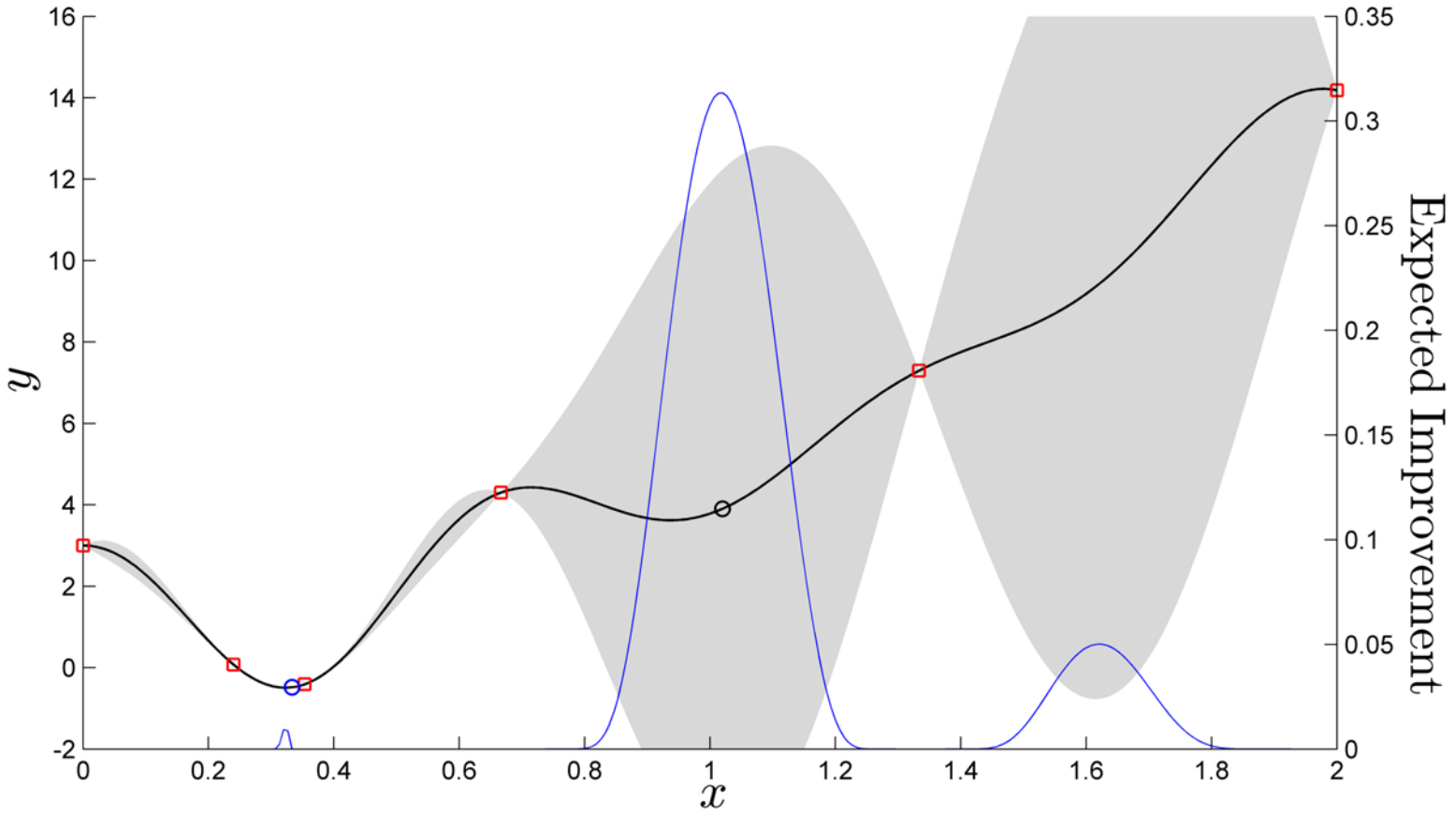
EGO Example


16
14
12
10
8
6
4
2
0
-2
y
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x
0.25
0.2
0.15
0.1
0.05
0
Expected Improvement

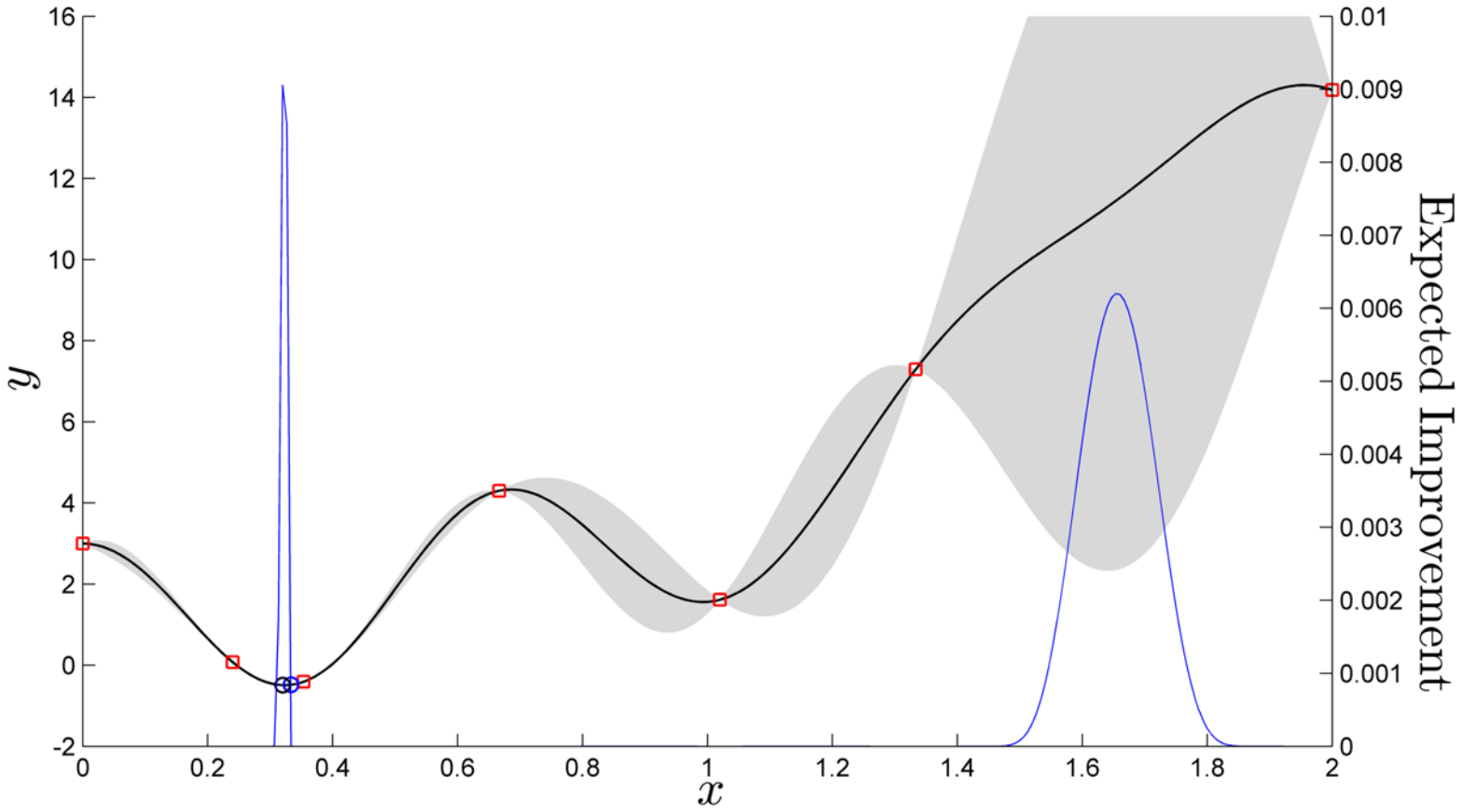
EGO Example


16
14
12
10
8
6
4
2
0
-2
y
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x
0.015
0.01
0.005
0
Expected Improvement

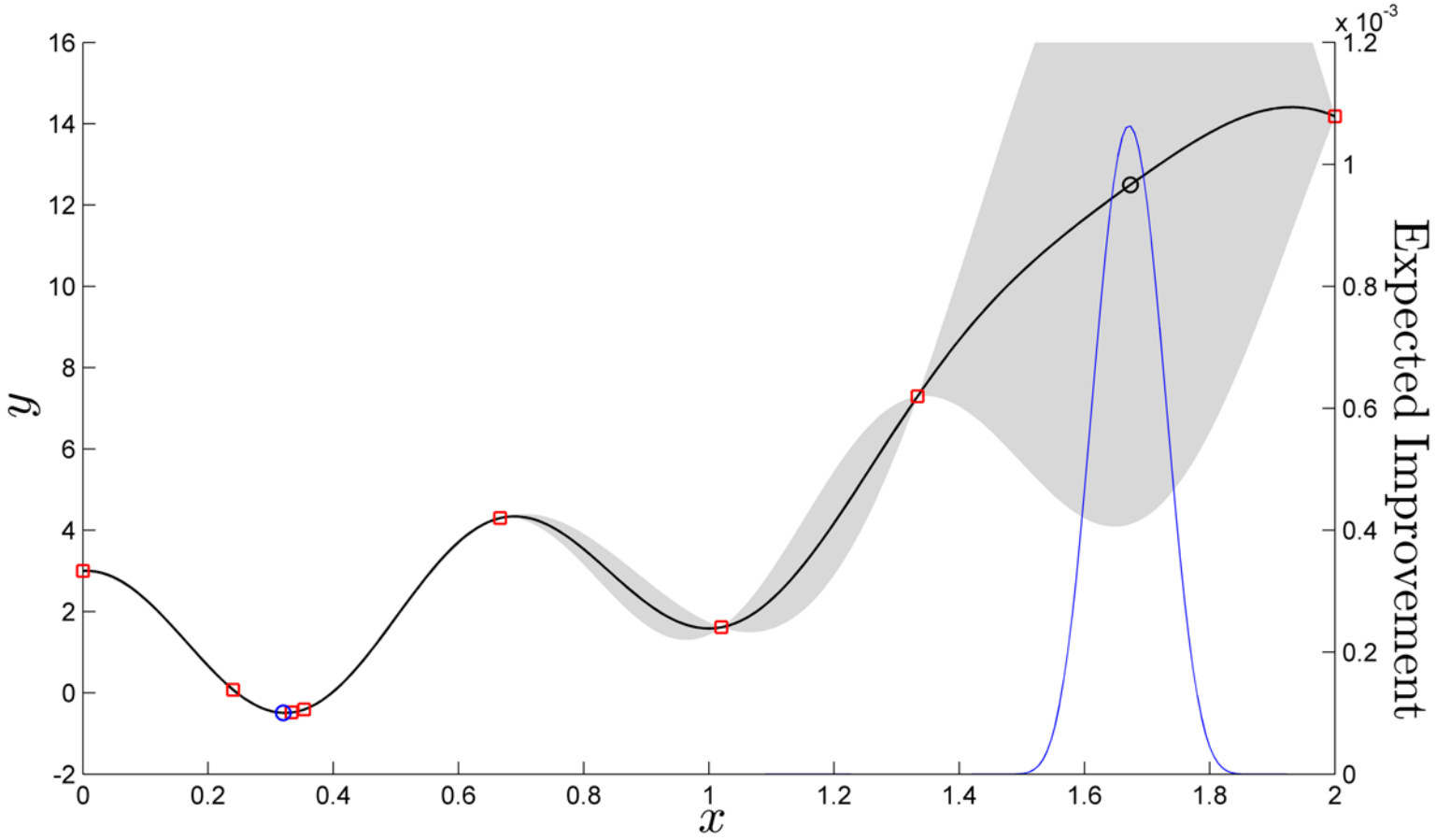
EGO Example


16
14
12
10
8
6
4
2
0
-2
y
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
Expected Improvement

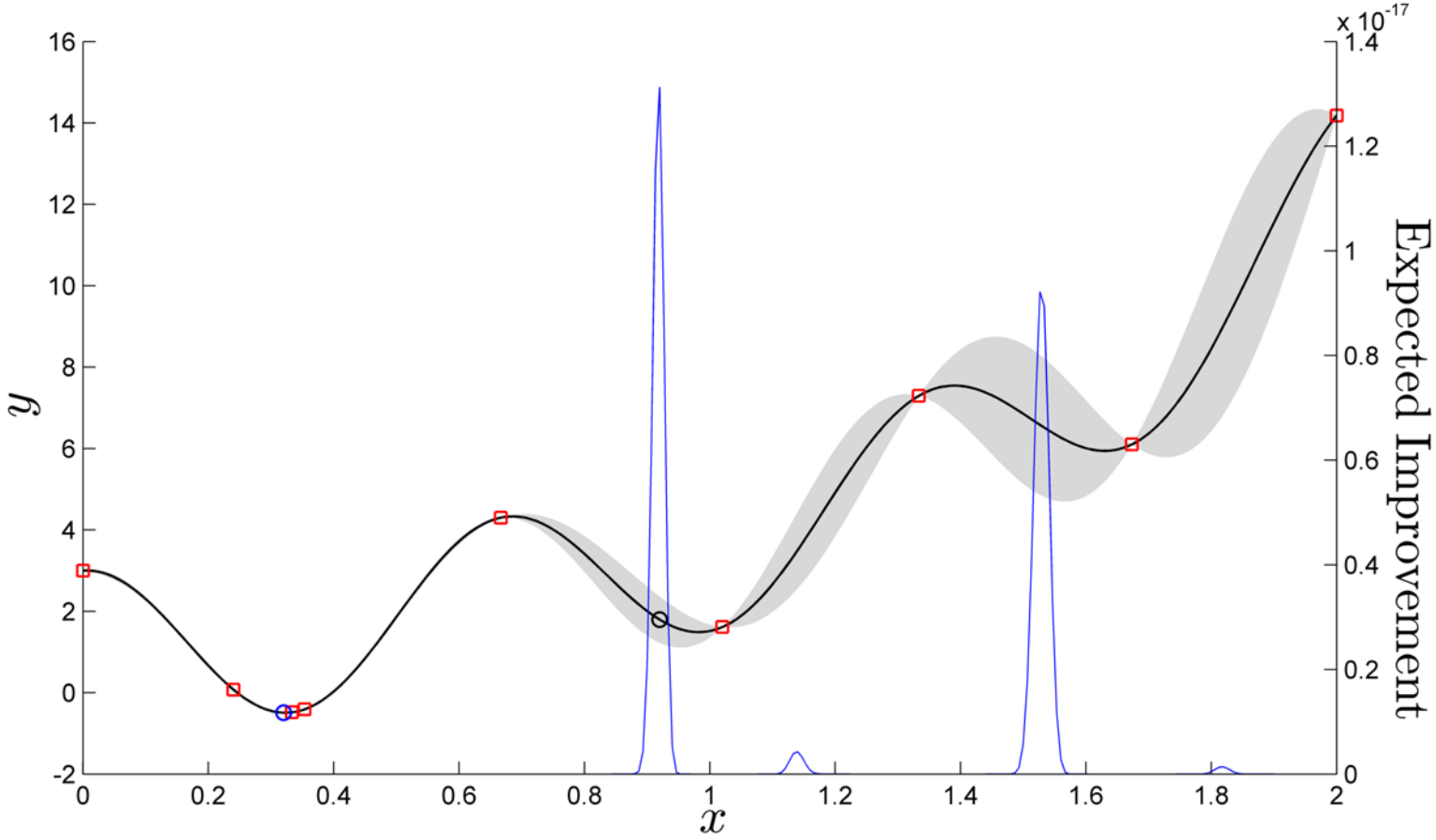
EGO Example


16
14
12
10
8
6
4
2
0
-2
y
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x
0.01
0.009
0.008
0.007
0.006
0.005
0.004
0.003
0.002
0.001
0
Expected Improvement

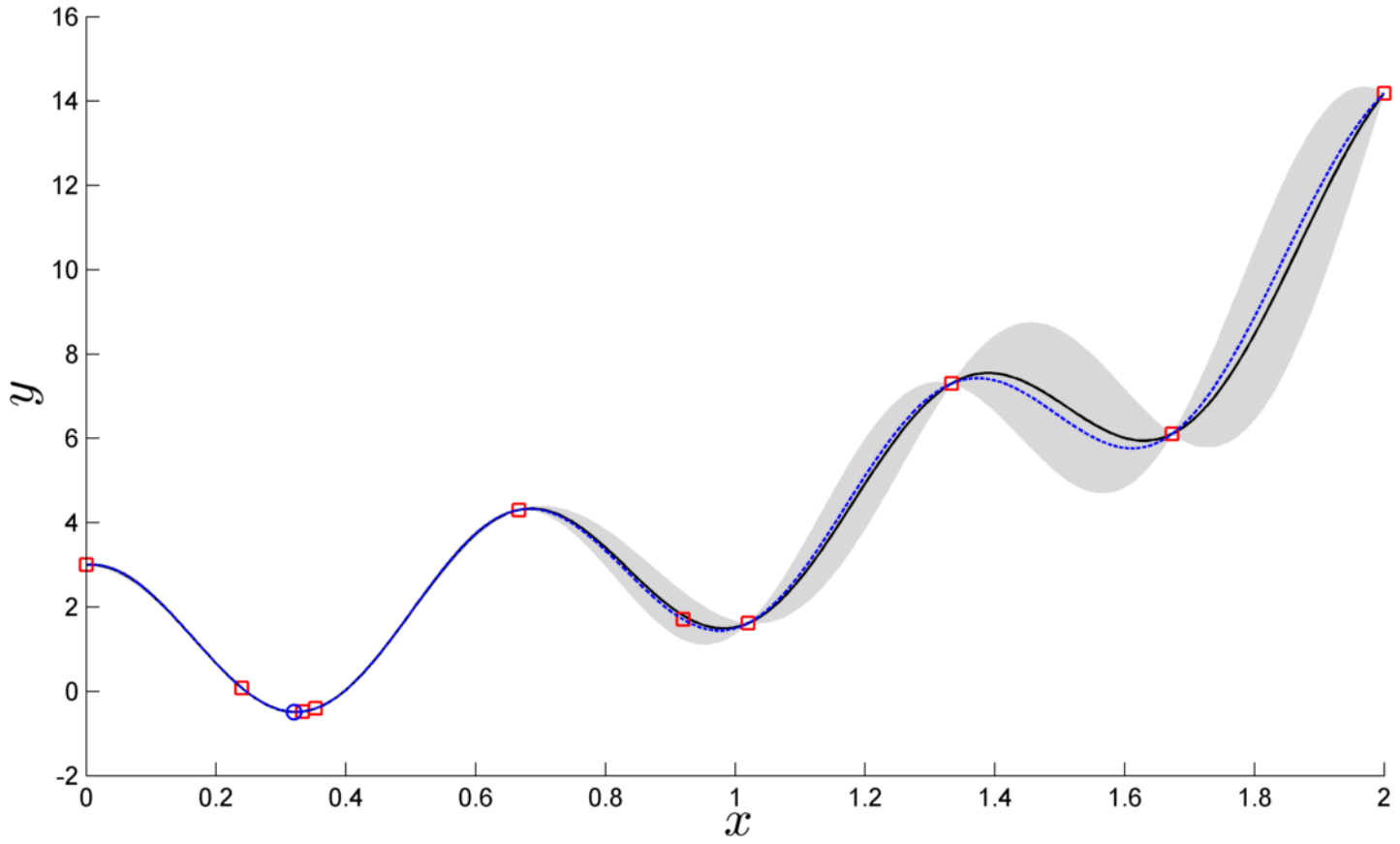
EGO Example


x 10^-3
1.2
1
0.8
0.6
0.4
0.2
0
Expected Improvement
16
14
12
10
8
6
4
2
0
-2
y
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
x

EGO Example


x 10^-17
Expected Improvement
y
x
16
14
12
10
8
6
4
2
0
-2
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2
1.4
1.2
1
0.8
0.6
0.4
0.2
0

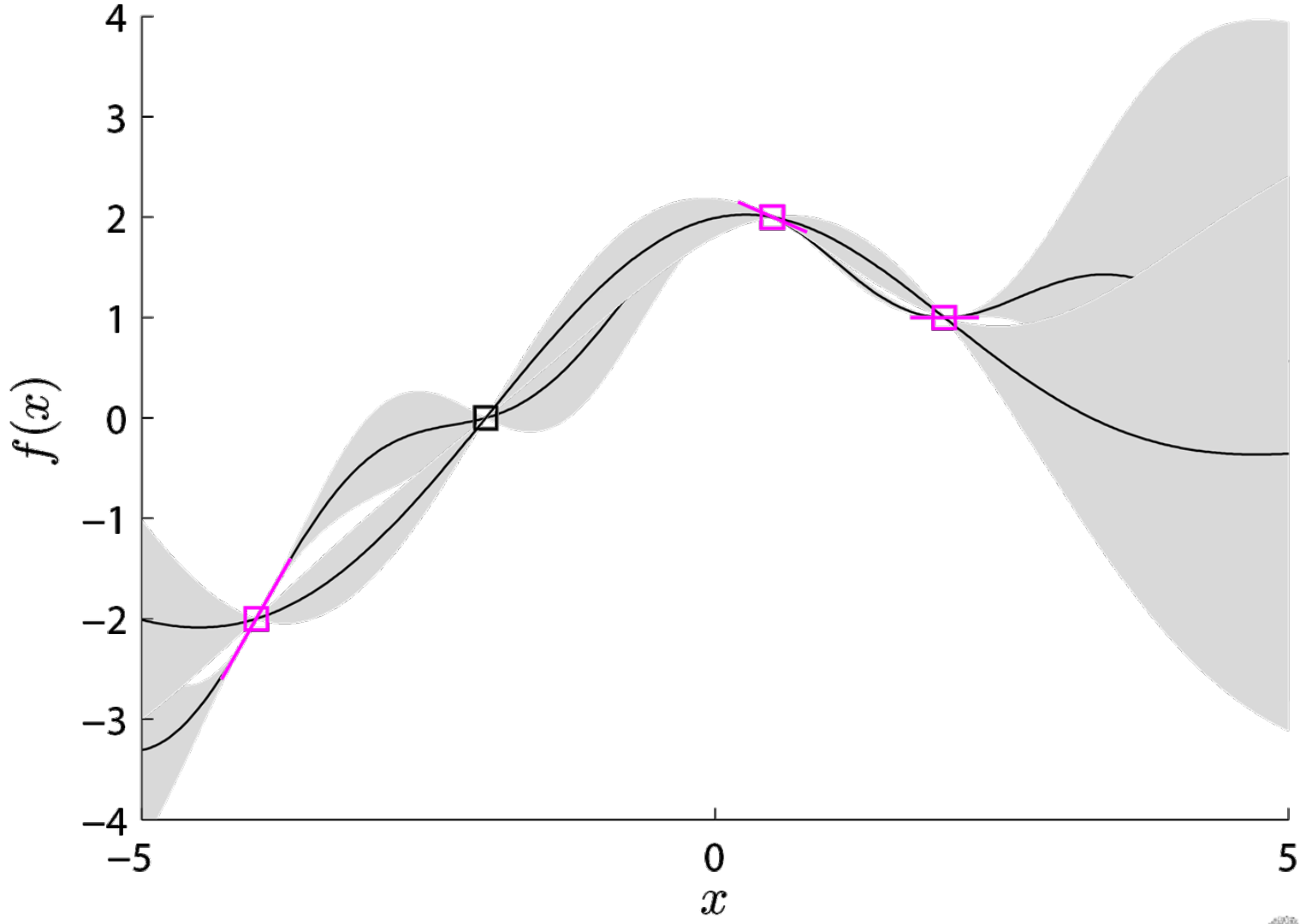
EGO Example



Other Features

- Gradient Information
  - Requires only a simple modification of the covariance function
  - Slovak et al “Derivative observations in Gaussian process models of dynamic systems” in *Conference on Neural Information Processing Systems* 2003

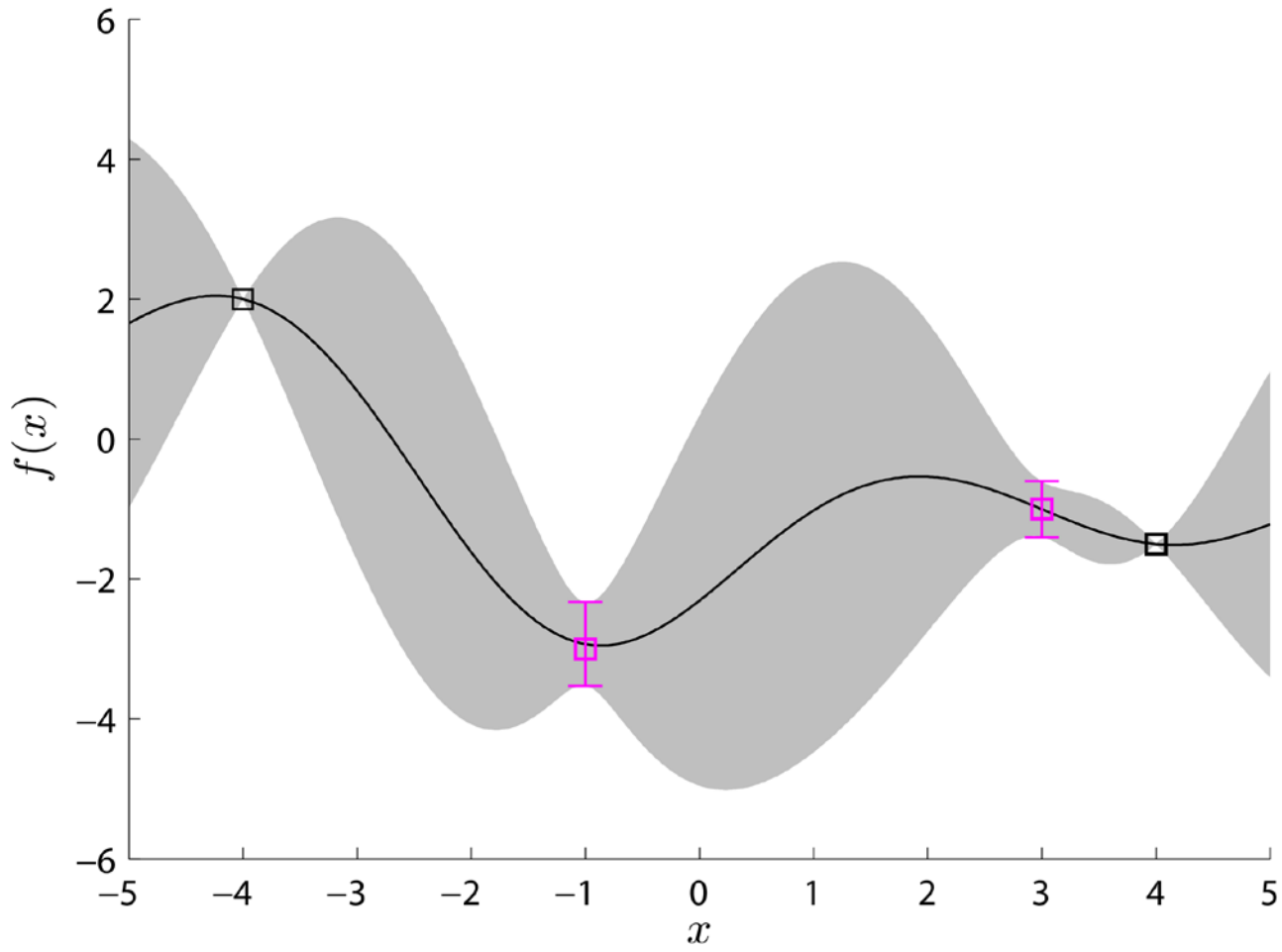
Gradient Information



Other Features

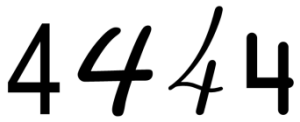
- Gradient Information
  - Requires only a simple modification of the covariance function
  - Slovak et al “Derivative observations in Gaussian process models of dynamic systems” in *Conference on Neural Information Processing Systems* 2003
- Training Data with Noise
  - Allows for the inclusion of uncertainty

Gaussian Process with Noise



Other Features

- Gradient Information
  - Requires only a simple modification of the covariance function
  - Slovak et al "Derivative observations in Gaussian process models of dynamic systems" in *Conference on Neural Information Processing Systems* 2003
- Training Data with Noise
  - Allows for the inclusion of uncertainty
- Classification Problems
  - Map inputs into discrete outputs
  - Example: Text recognition software



Keep in mind…

- A model is only as good as the assumptions behind it
  - Be careful when choosing a covariance function
  - Stationary and differentiable may not be good assumptions for your model
- Gaussian process models are useful for applications other than surrogate modeling and optimization
  - Geostatistics
  - Robotics
  - Meteorology
  - Artificial Intelligence
  - Many more…