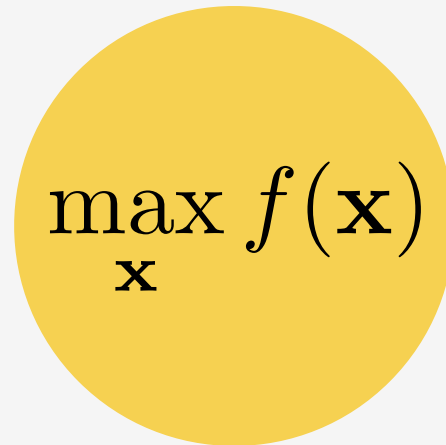


# Bayesian Optimization

Matt J. Kusner

# Bayesian Optimization


$$\max_{\mathbf{x}} f(\mathbf{x})$$

# Bayesian Optimization


$$\max_{\mathbf{x}} f(\mathbf{x})$$

properties of  $f(\mathbf{x})$ :

$f(\mathbf{x})$  is **very expensive** to compute

$f(\mathbf{x})$  is **nonconvex**

# Bayesian Optimization


$$\max_{\mathbf{x}} f(\mathbf{x})$$

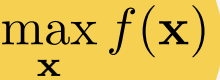
properties of  $f(\mathbf{x})$ :

$f(\mathbf{x})$  is **very expensive** to compute

$f(\mathbf{x})$  is **nonconvex**

**idea:** model  $f(\mathbf{x})$  with an easy-to-evaluate surrogate

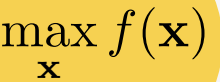
# Bayes. Opt. Application


$$\max_{\mathbf{x}} f(\mathbf{x})$$

## Hyperparameter tuning

[Bergstra et al. 2011, Hutter et al. 2011, Snoek et al. 2012]

# Bayes. Opt. Application


$$\max_{\mathbf{x}} f(\mathbf{x})$$

## Hyperparameter tuning

[Bergstra et al. 2011, Hutter et al. 2011, Snoek et al. 2012]

RBF Kernel SVM. Hyperparameters:  $(\lambda, \sigma)$

$1 - f(\lambda, \sigma)$  is **validation error**

# Bayes. Opt. Application

$$\max_{\mathbf{x}} f(\mathbf{x})$$

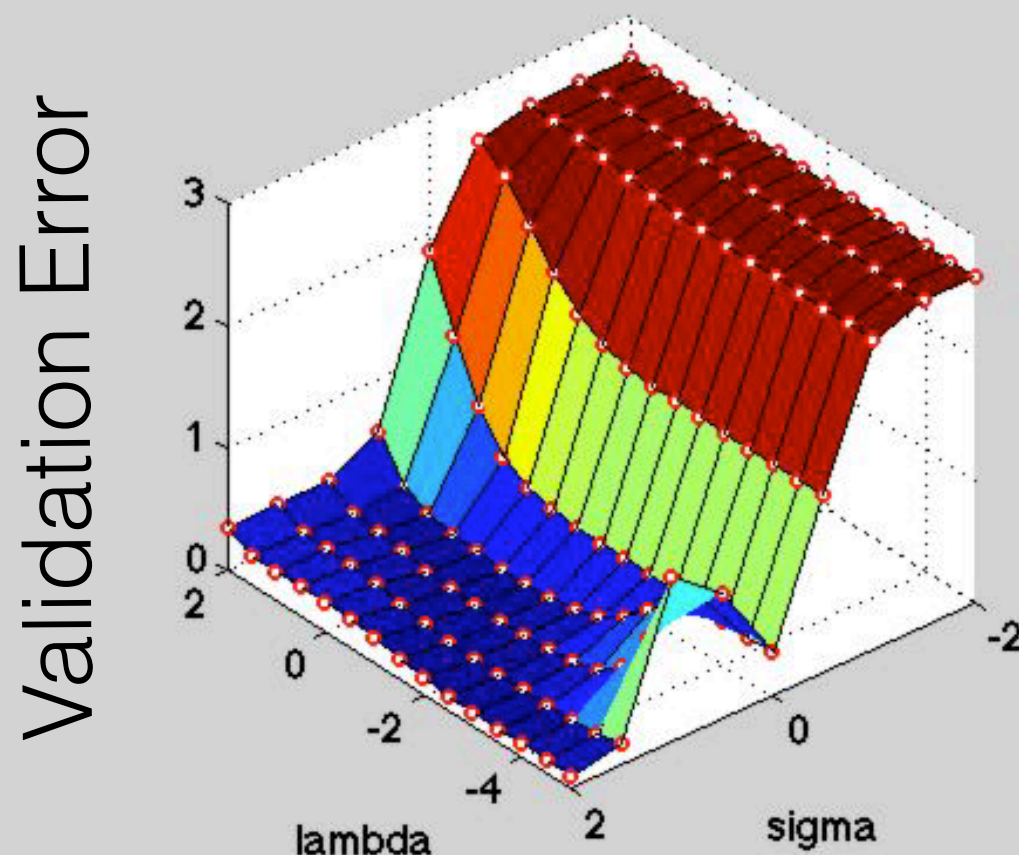
## Hyperparameter tuning

[Bergstra et al. 2011, Hutter et al. 2011, Snoek et al. 2012]

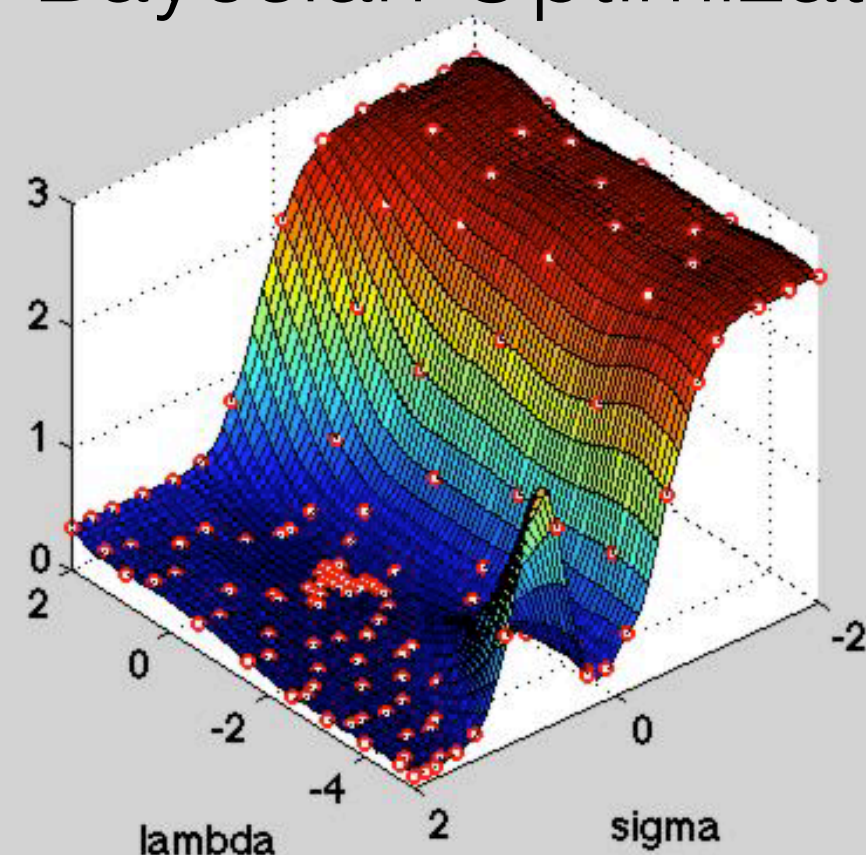
RBF Kernel SVM. Hyperparameters:  $(\lambda, \sigma)$

$1 - f(\lambda, \sigma)$  is **validation error**


Grid Search



Bayesian Optimization



# Bayesian Optimization

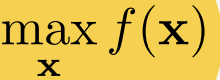

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



# Bayesian Optimization


$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

a prior distribution  
over functions

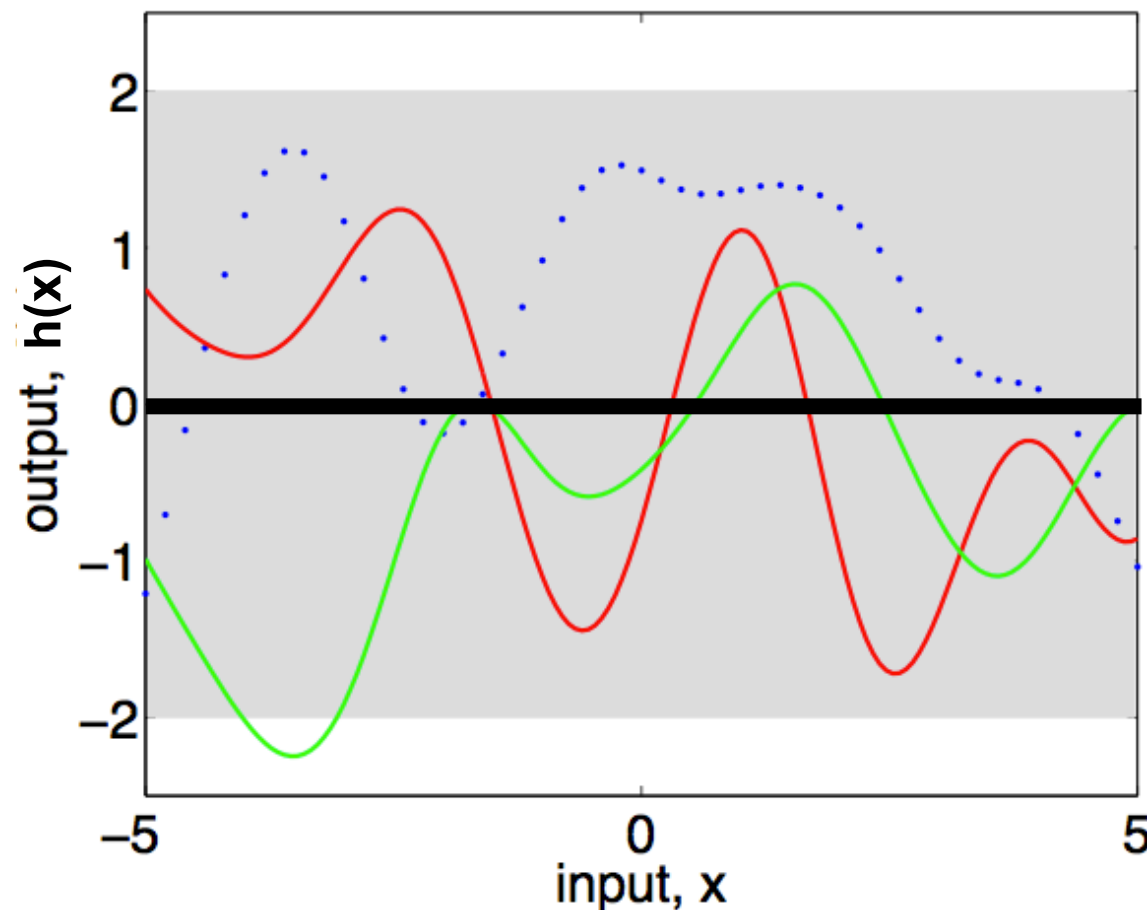
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

a prior distribution  
over functions



(a), prior

figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

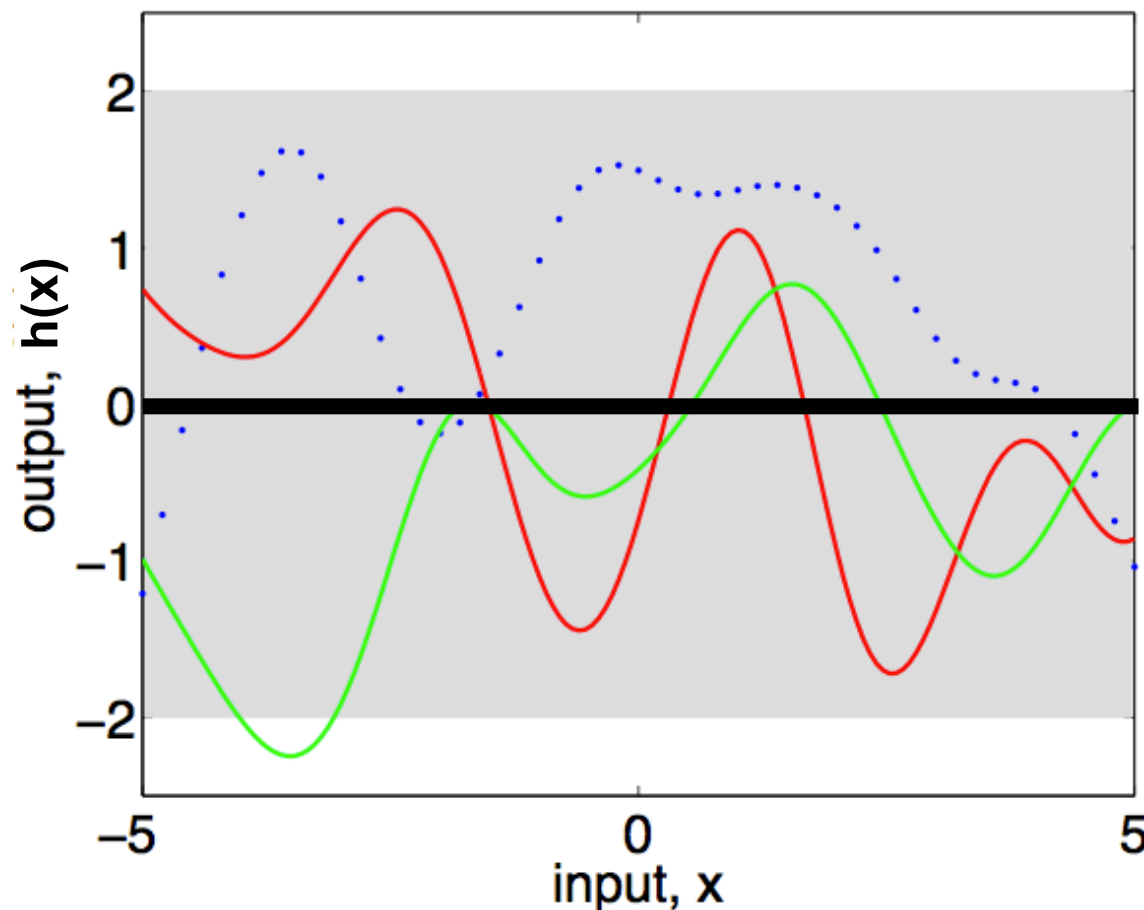
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

kernel function  
 $k(\mathbf{x}, \mathbf{x}')$

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

a prior distribution  
over functions



(a), prior

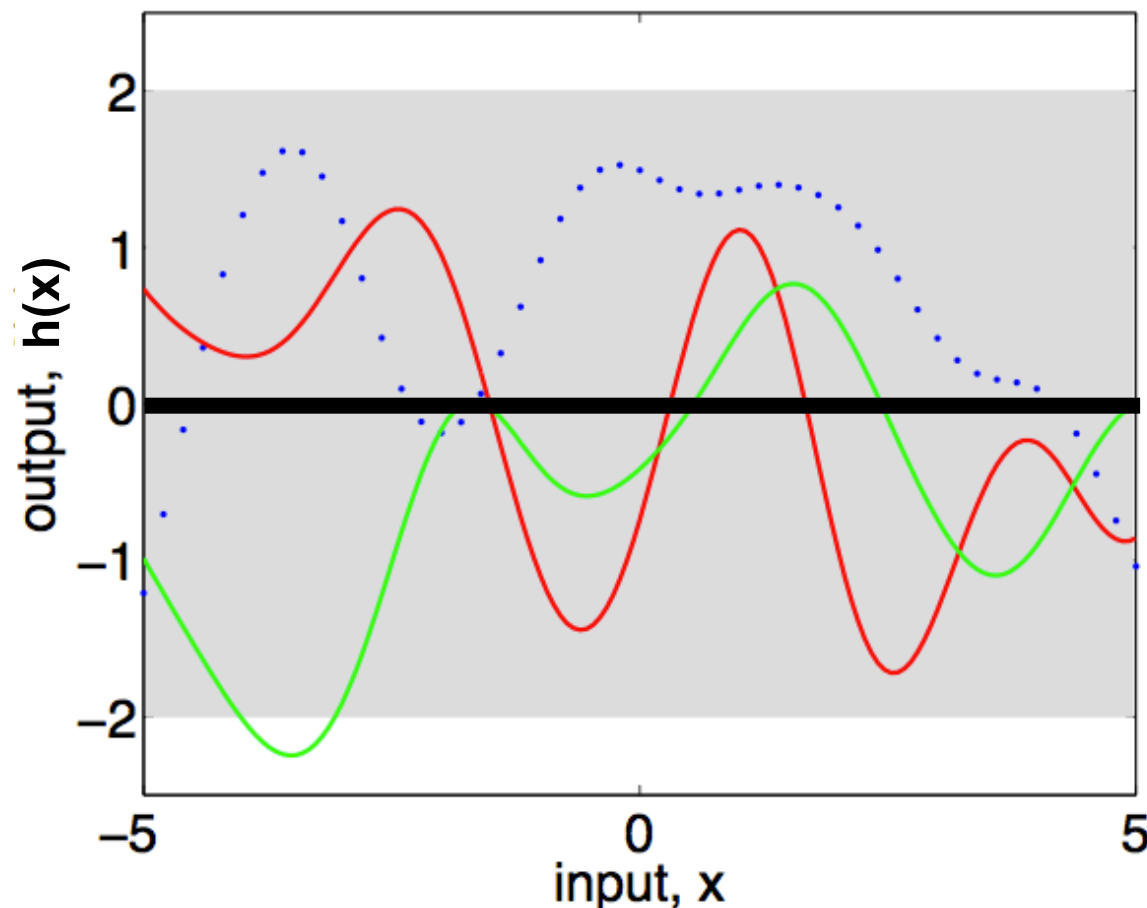
figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

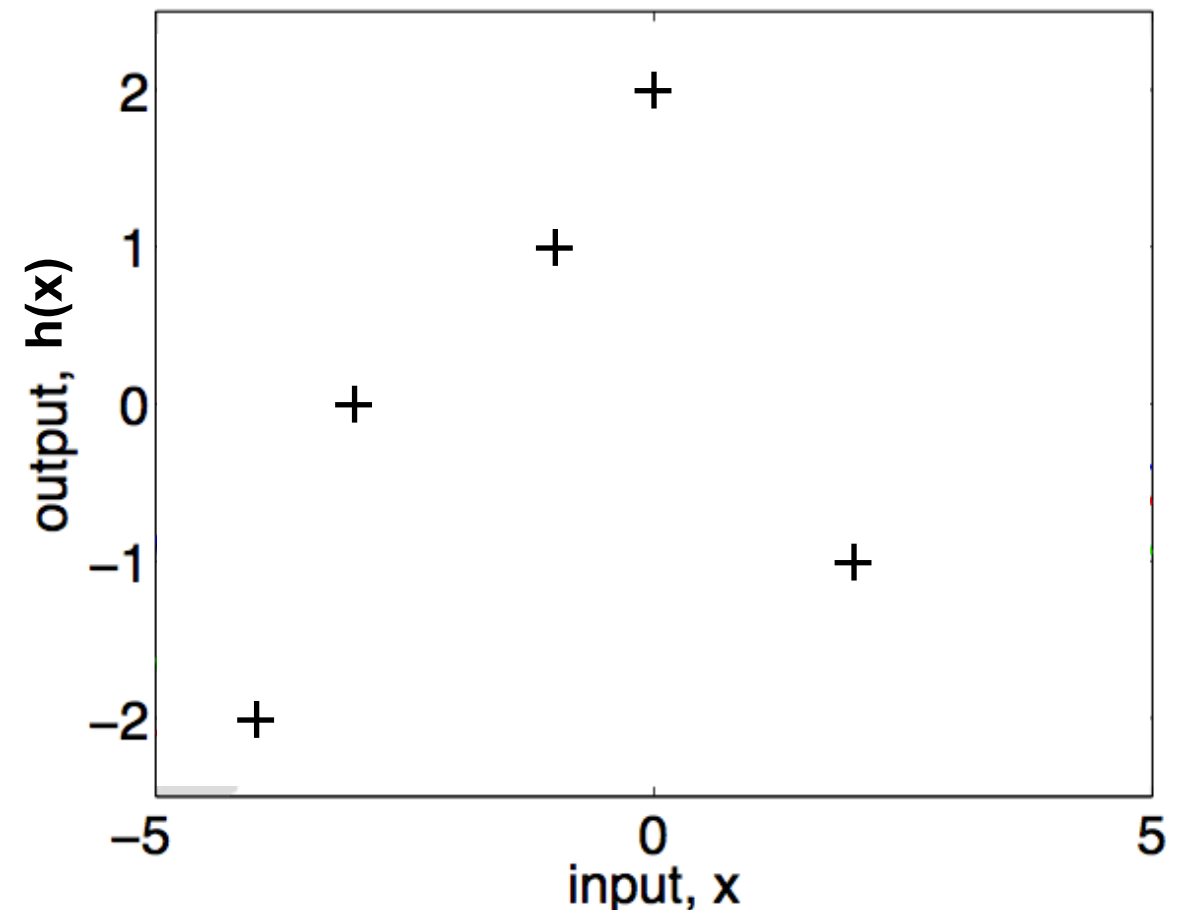
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



(a), prior

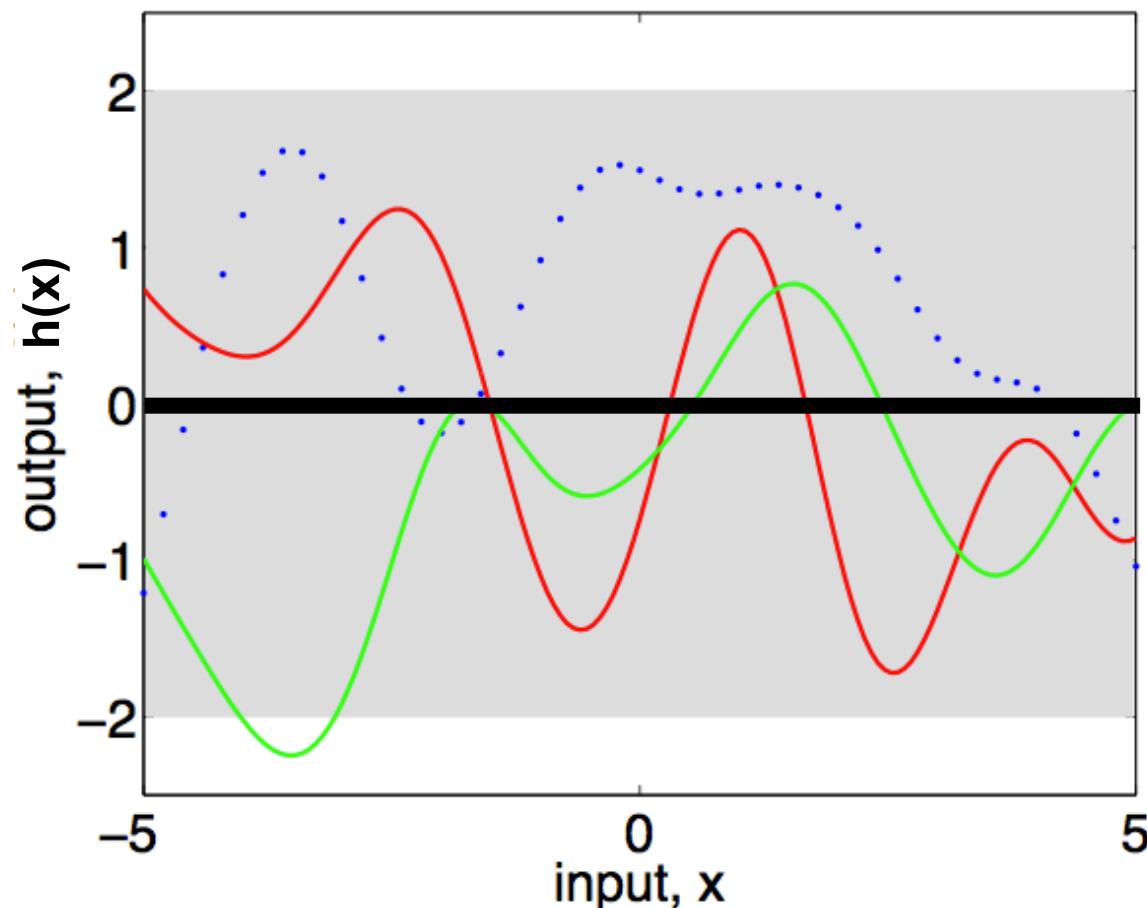


# Bayesian Optimization

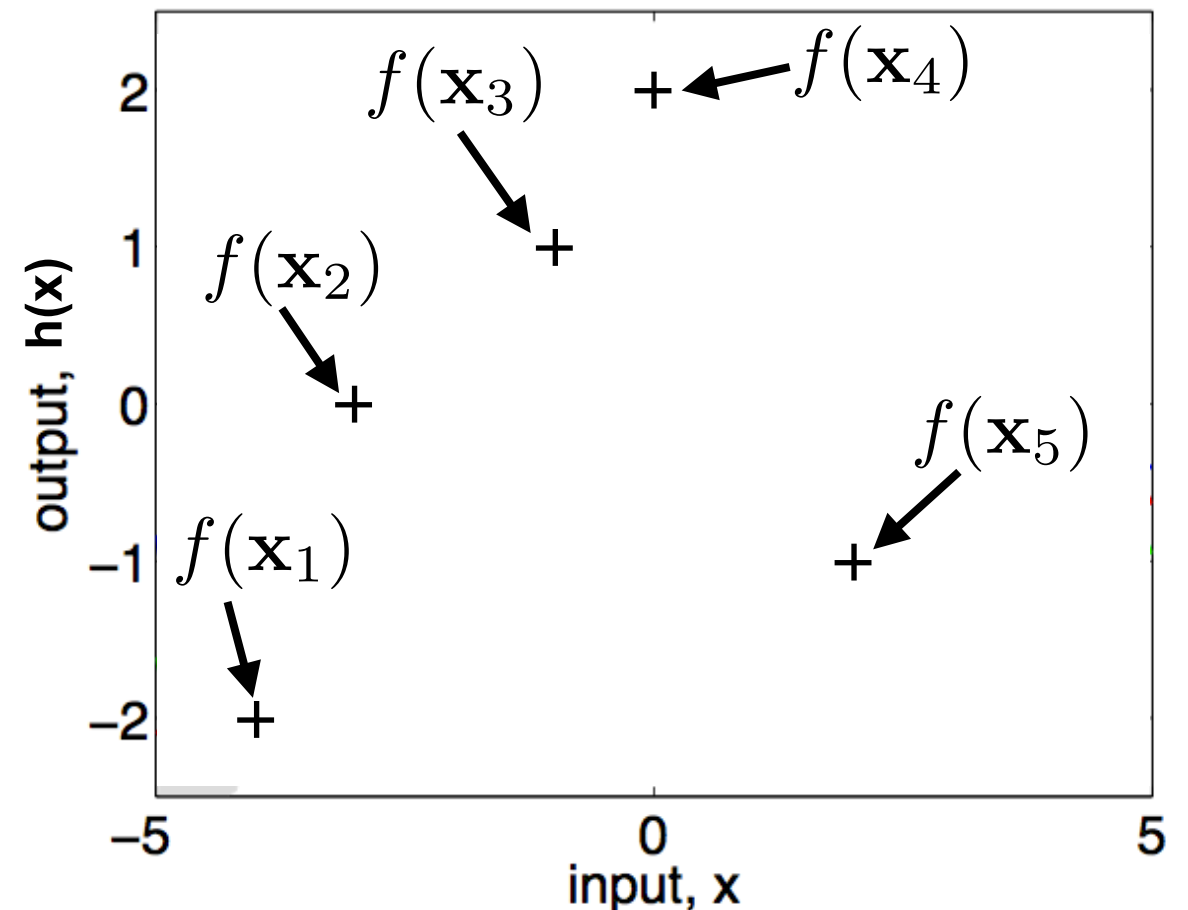
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



(a), prior

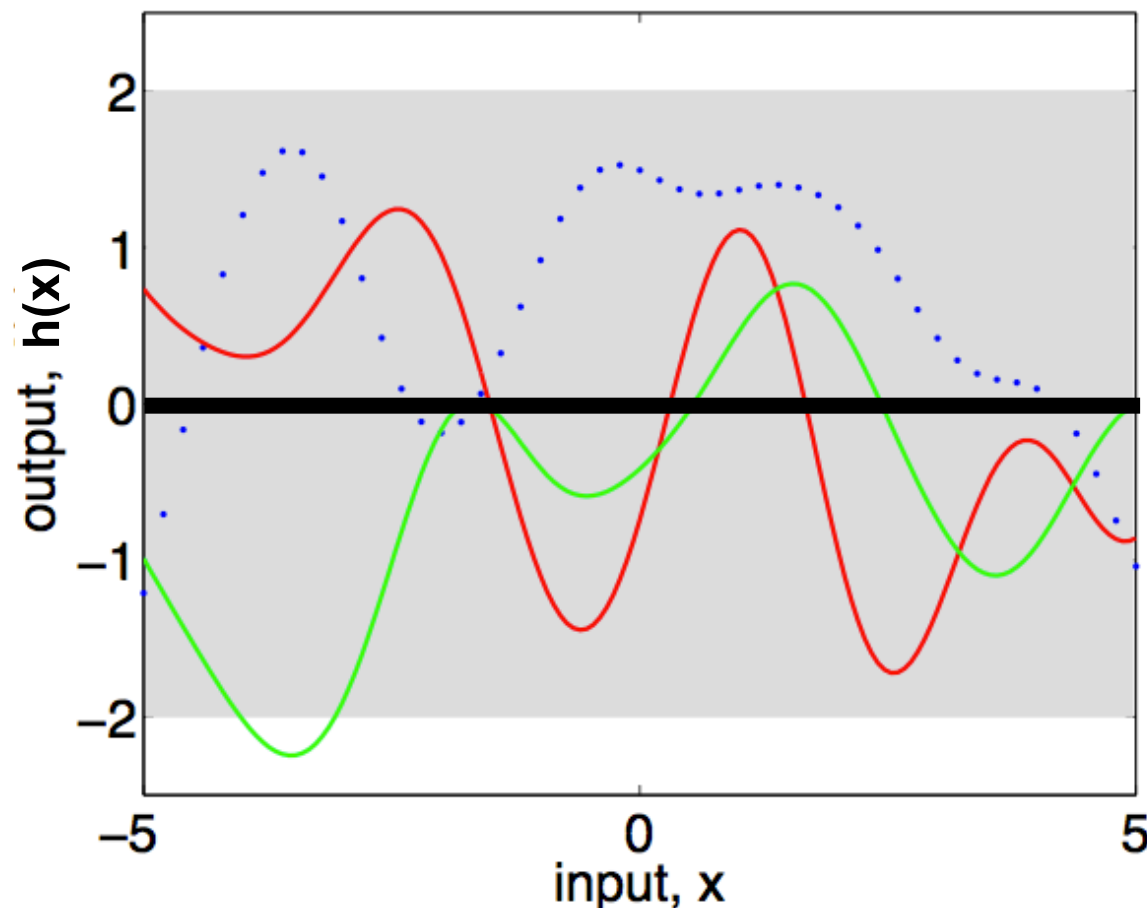


# Bayesian Optimization

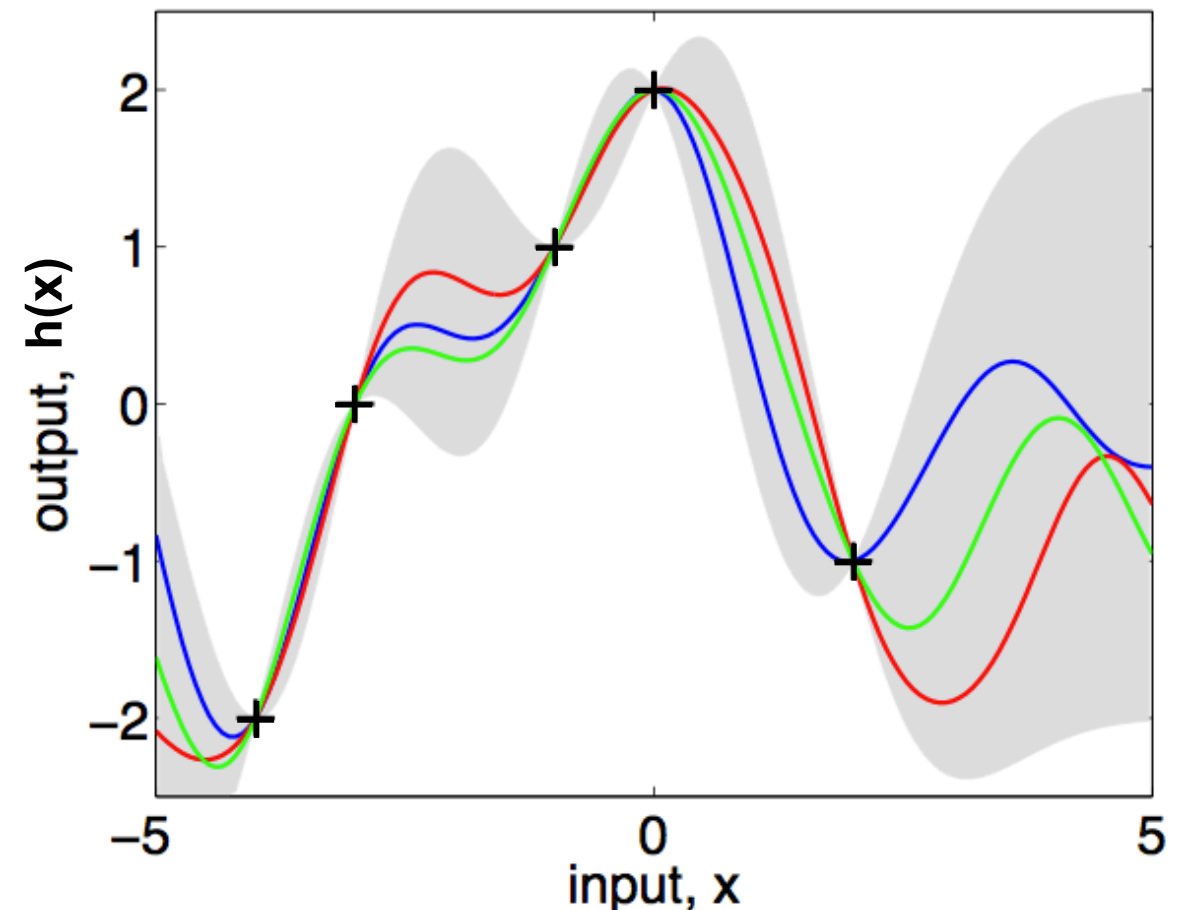
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



(a), prior



(b), posterior

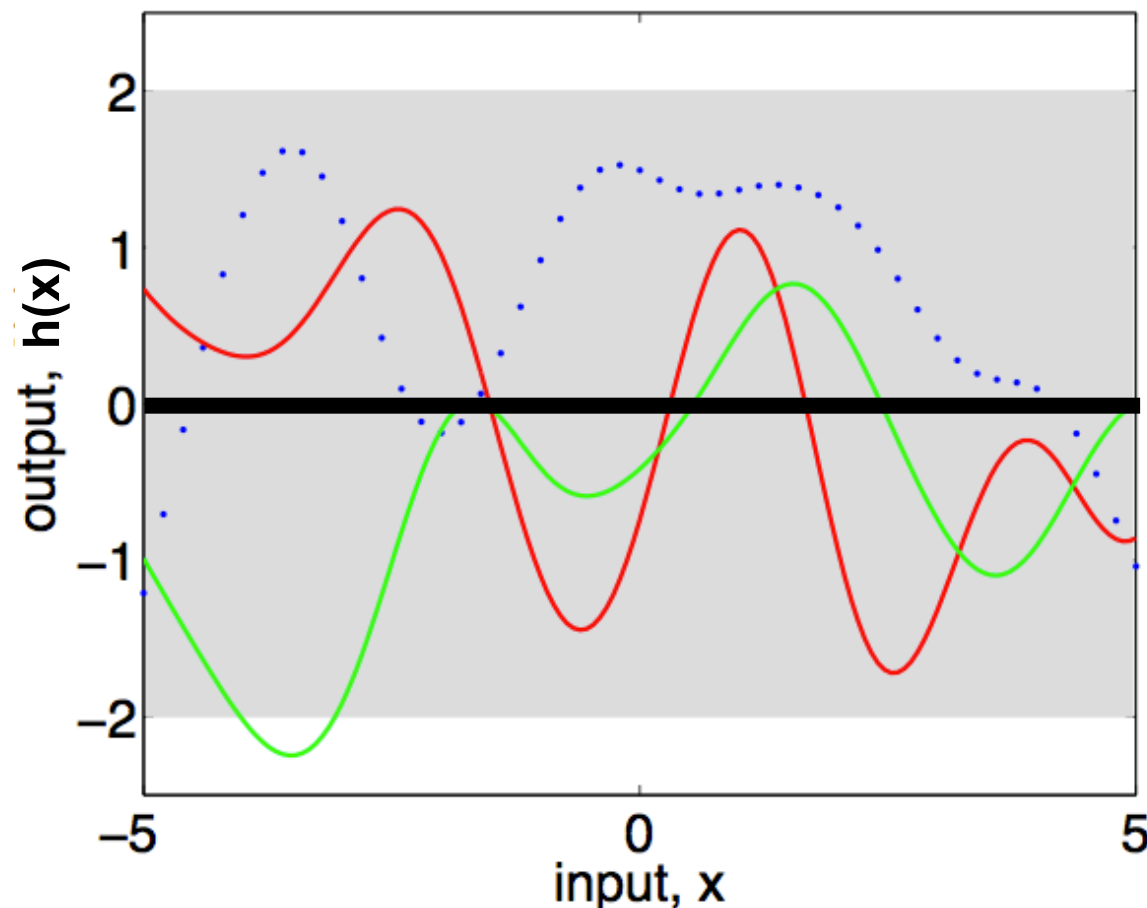
figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

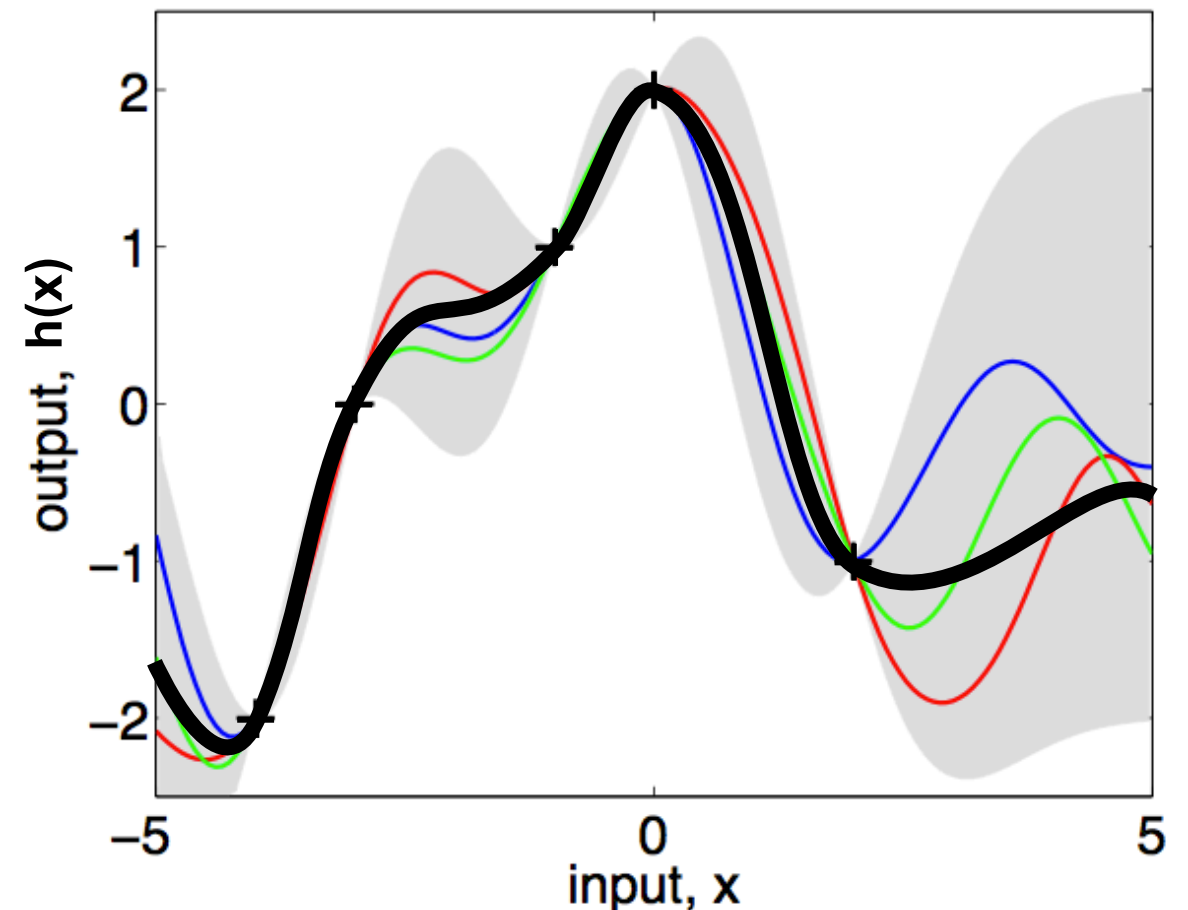
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



(a), prior



(b), posterior

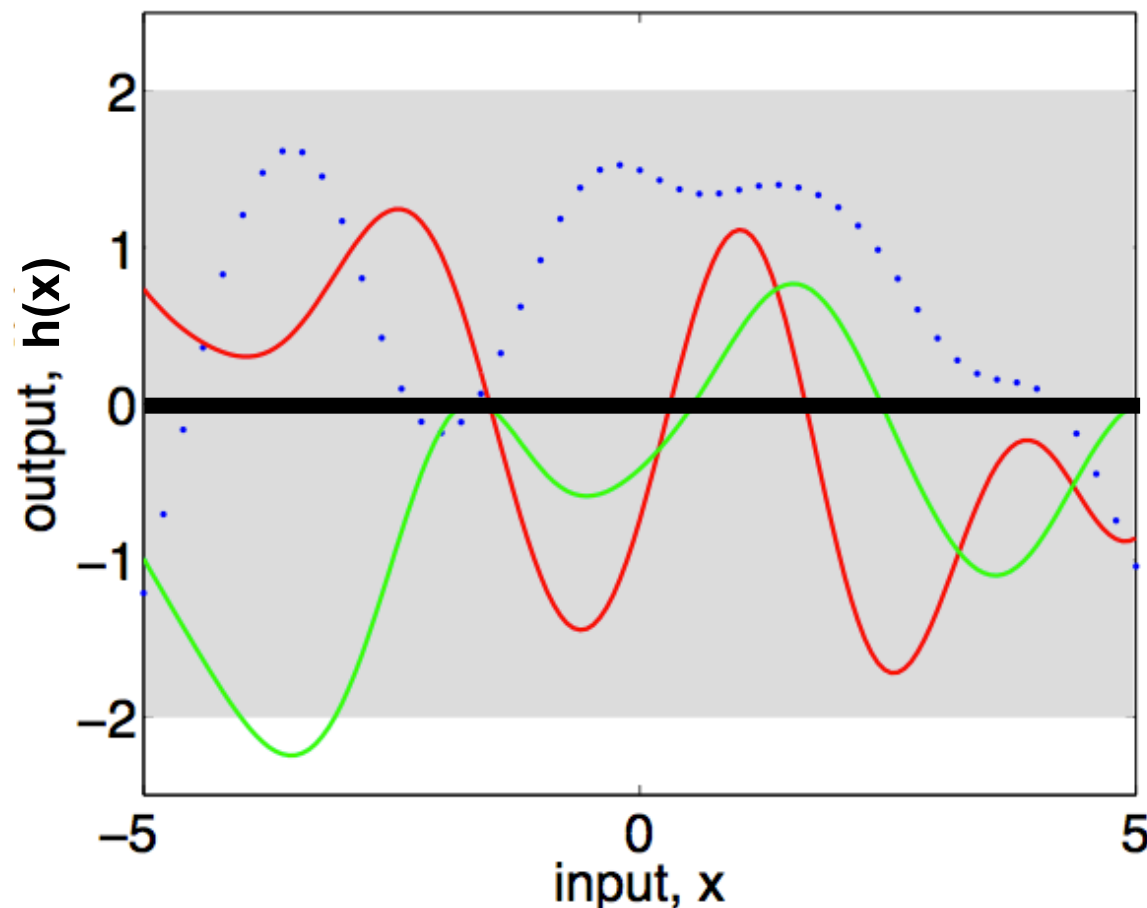
figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

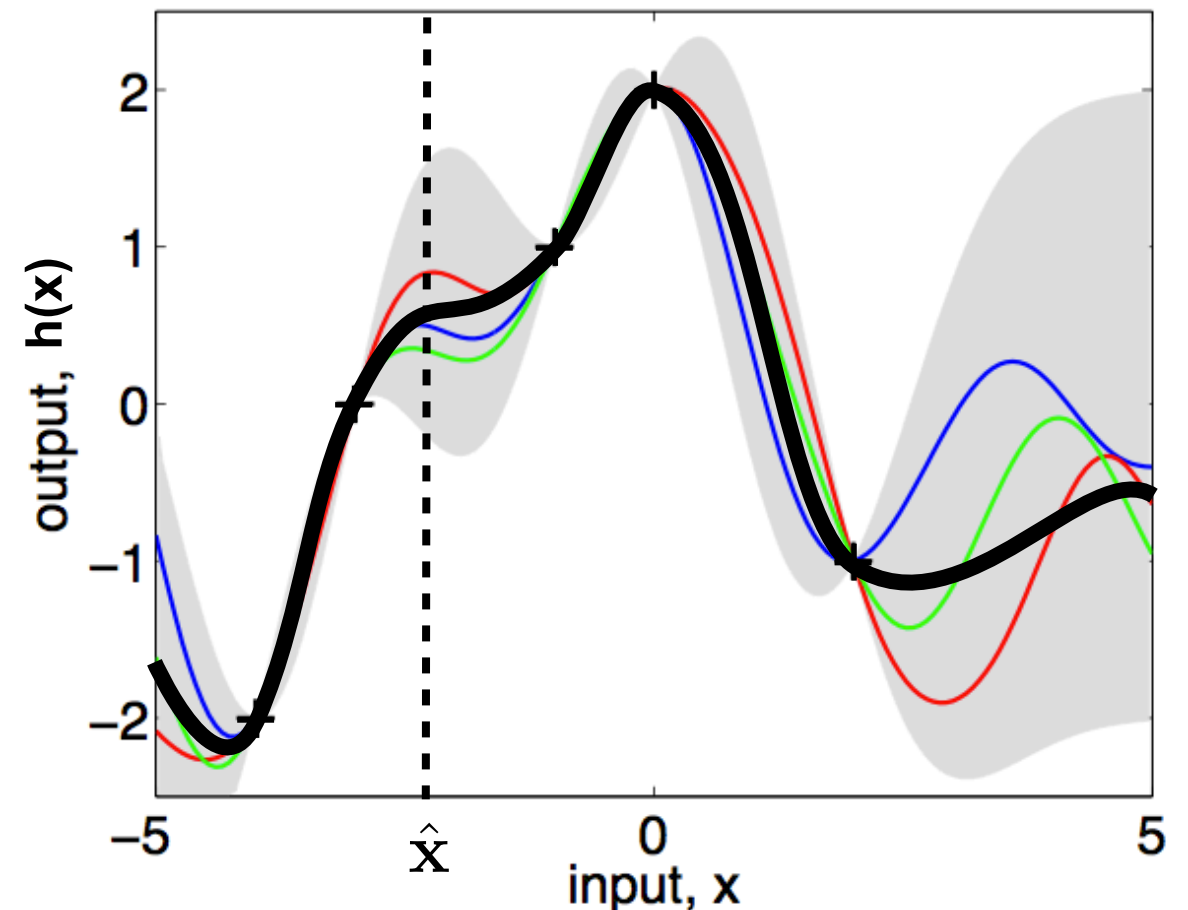
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



(a), prior



(b), posterior

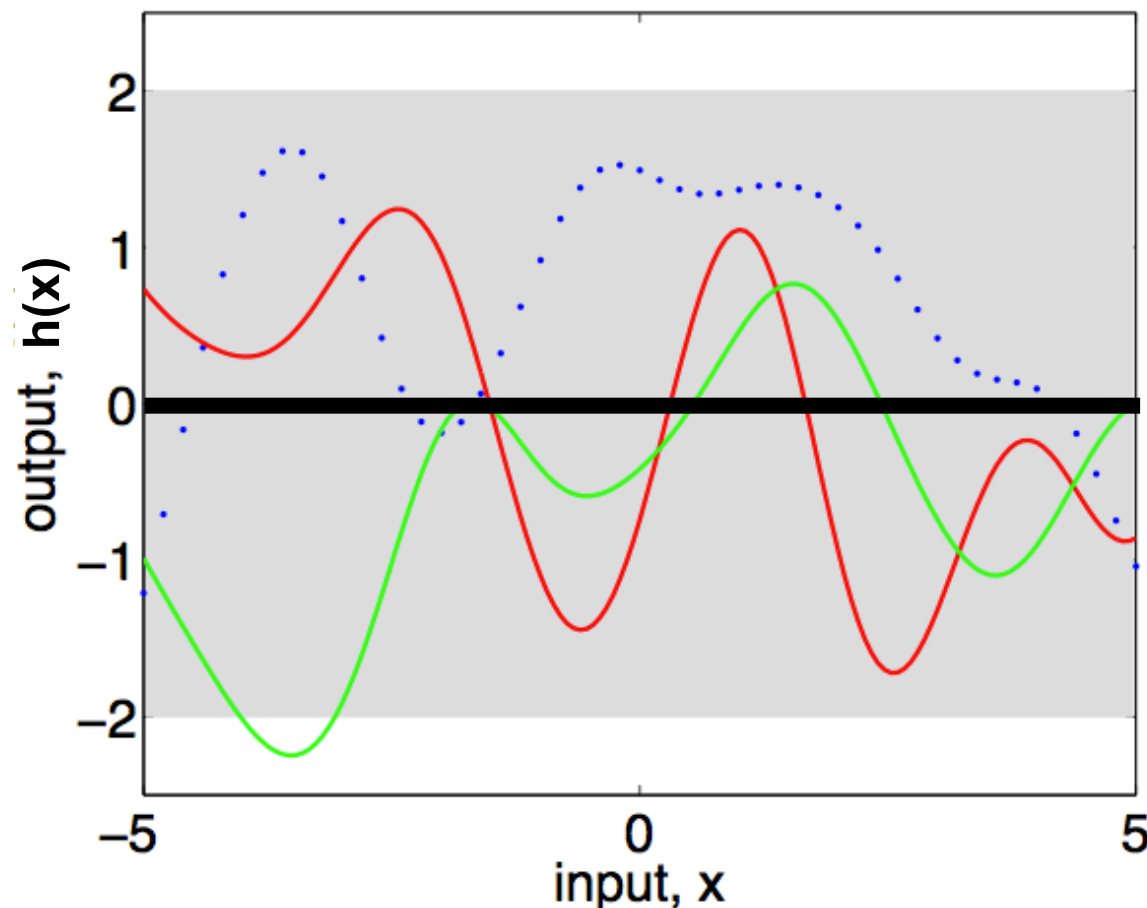


# Bayesian Optimization

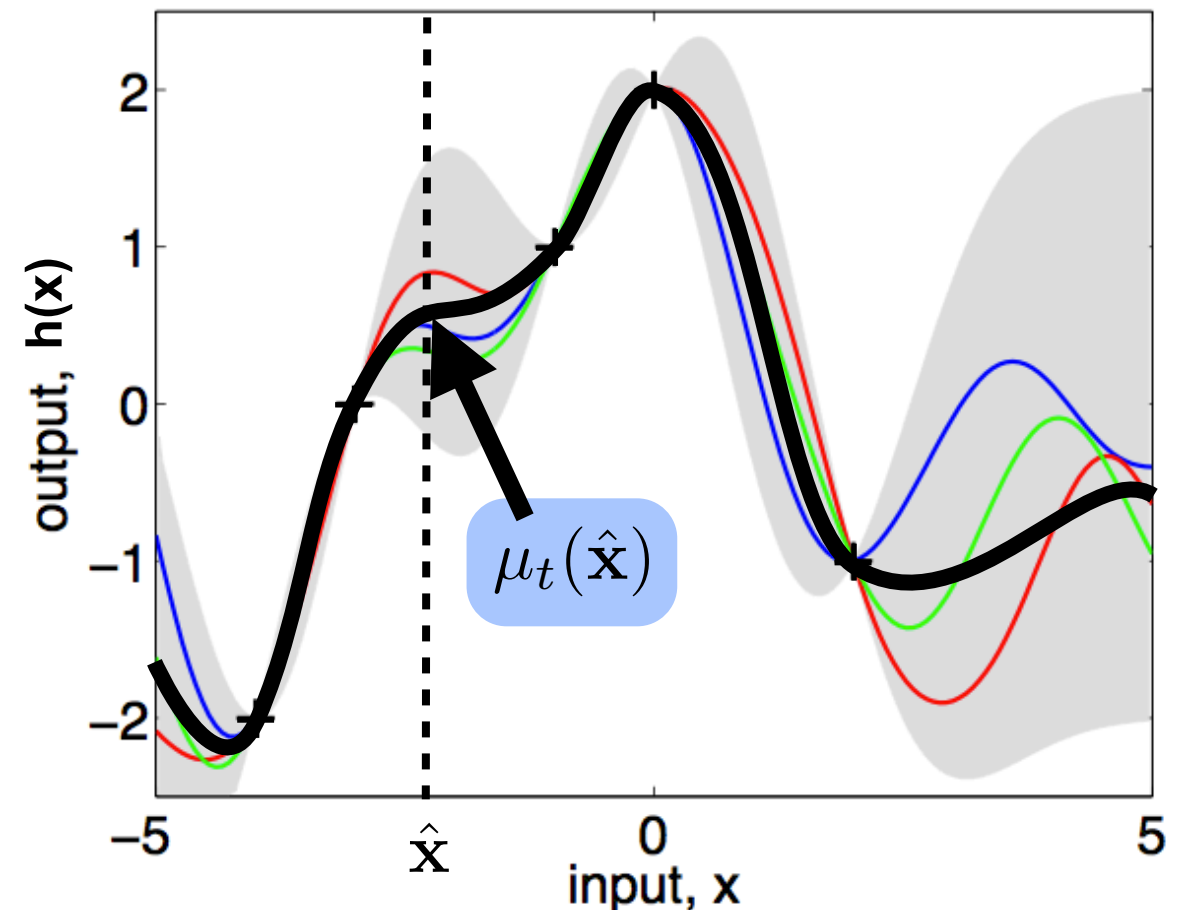
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



(a), prior



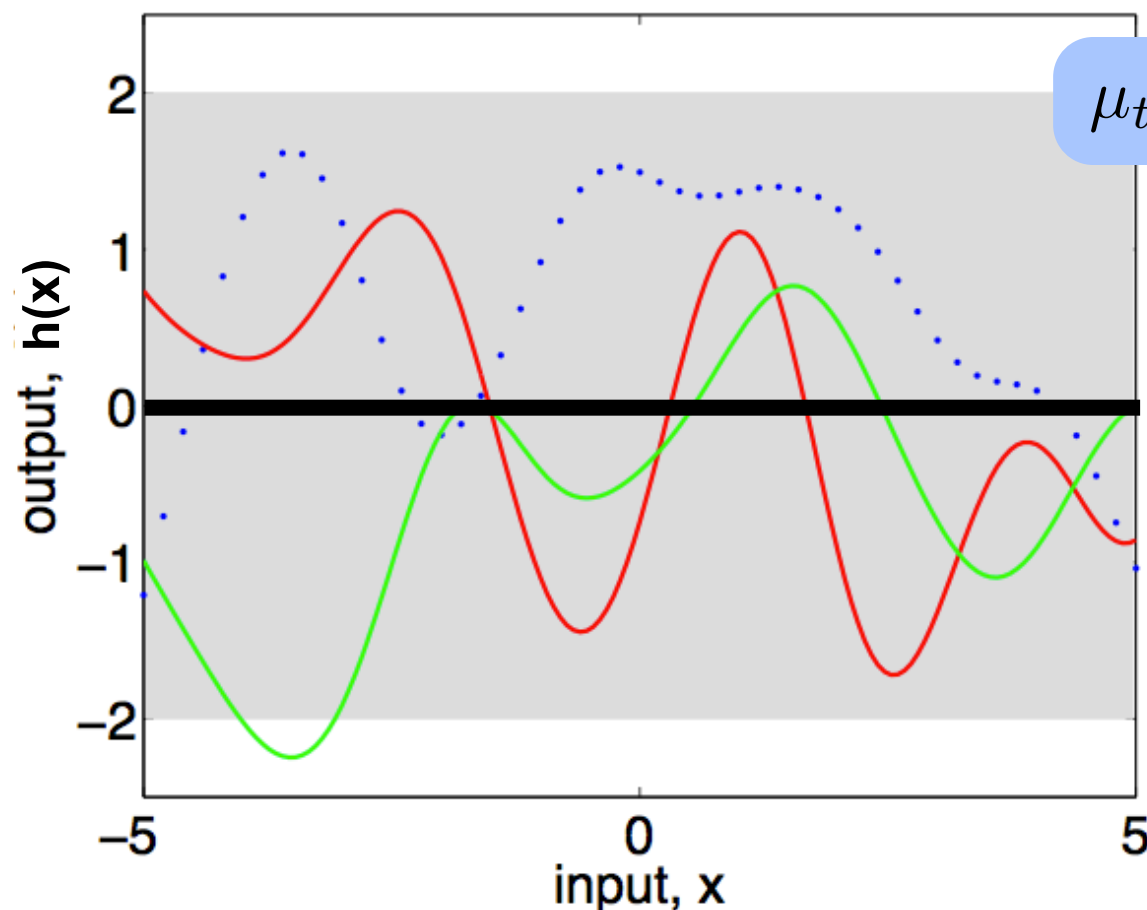
(b), posterior

# Bayesian Optimization

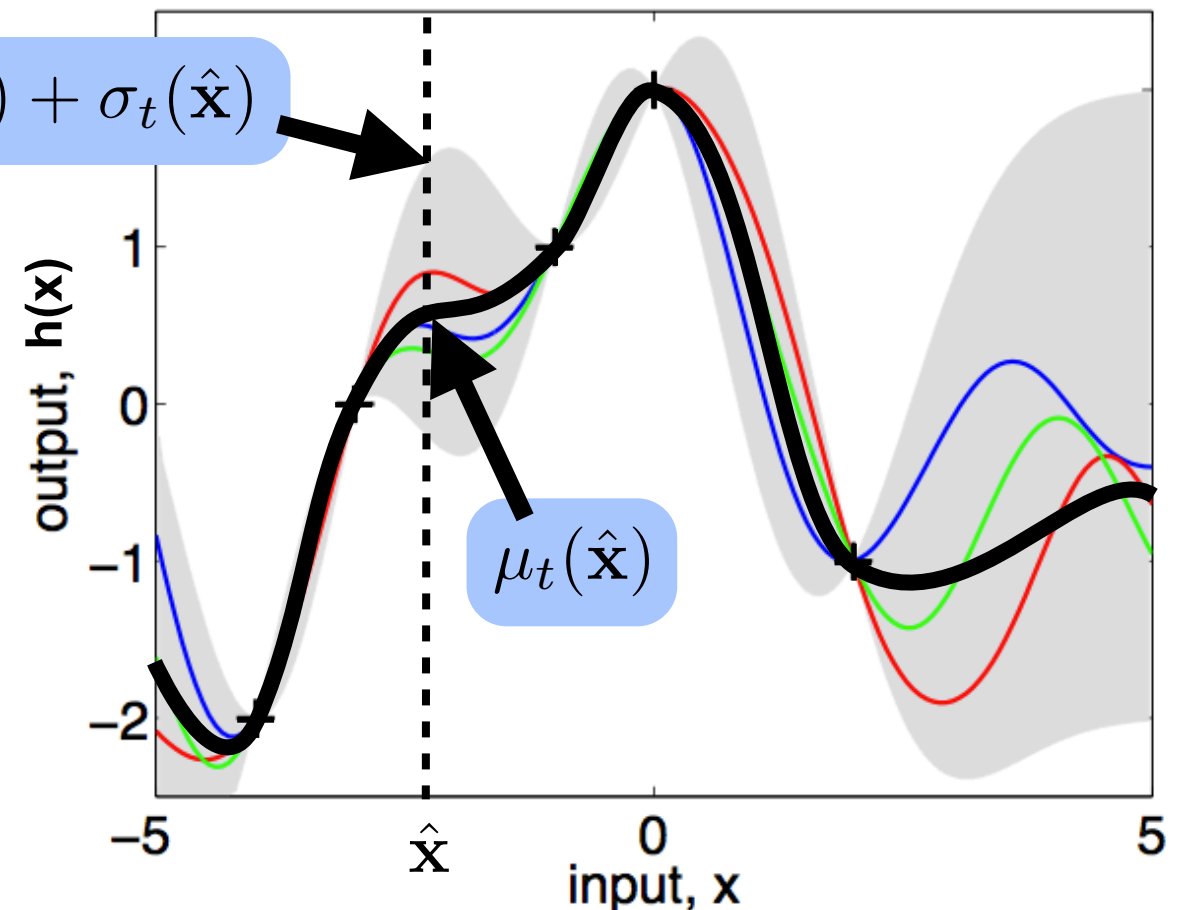
$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$



(a), prior



(b), posterior

figure credit: [Rasmussen & Williams, 2006]

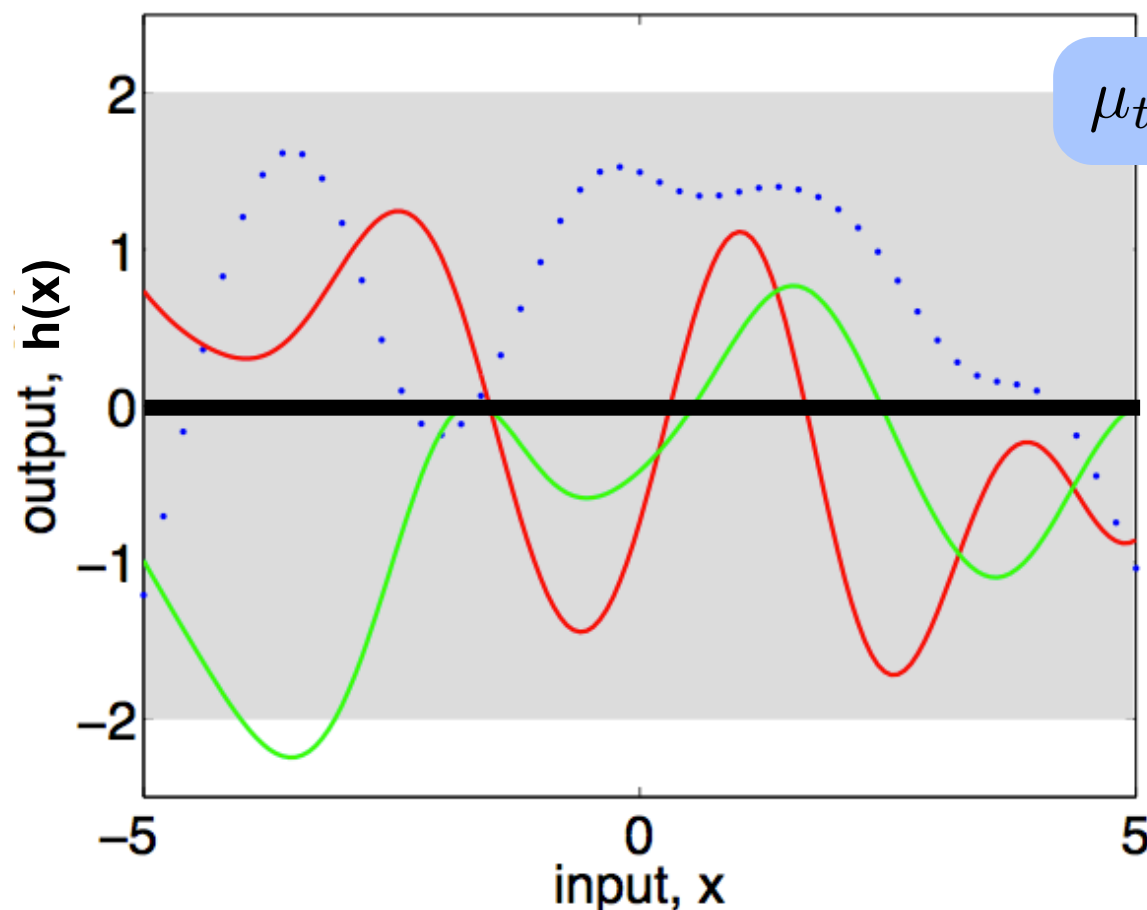
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

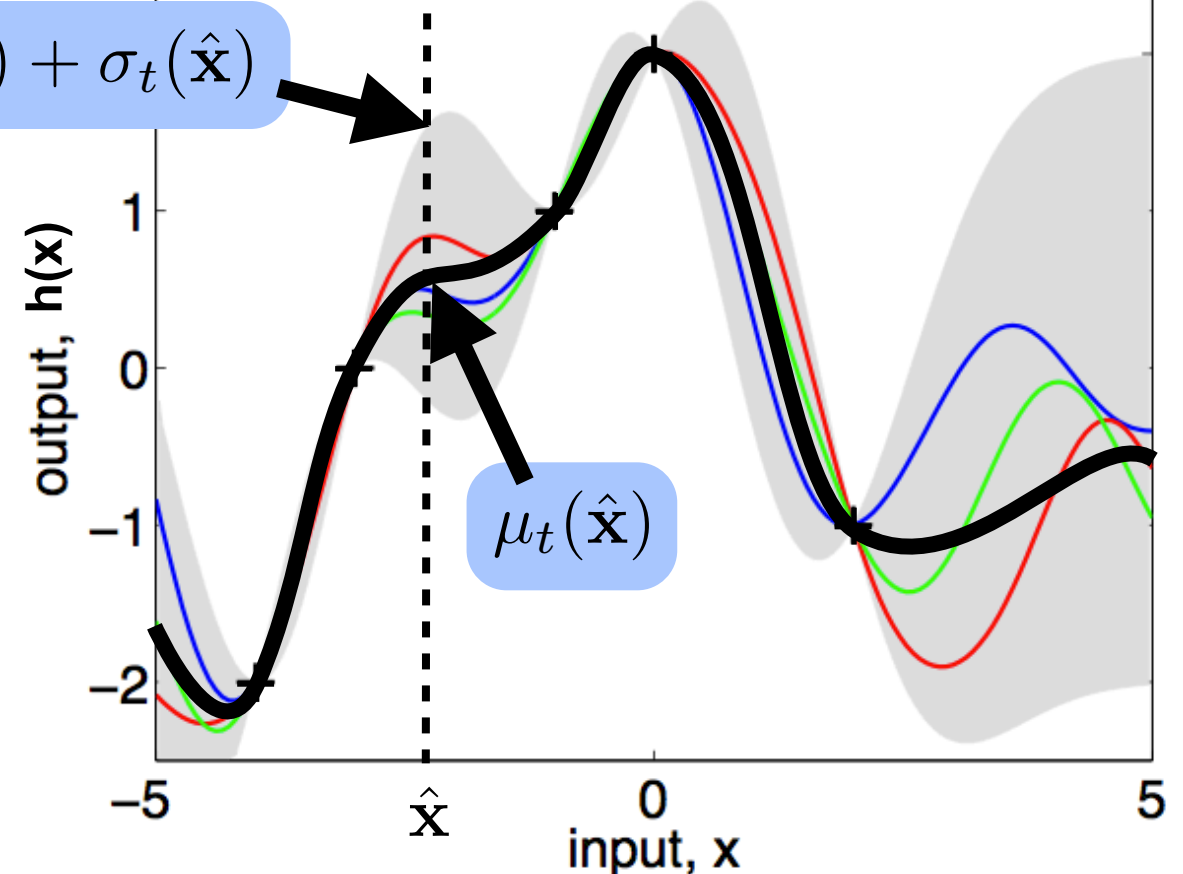
Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

$$h(\hat{\mathbf{x}}) \sim \mathcal{N}(\mu_t(\hat{\mathbf{x}}), \sigma_t^2(\hat{\mathbf{x}}))$$



(a), prior



(b), posterior

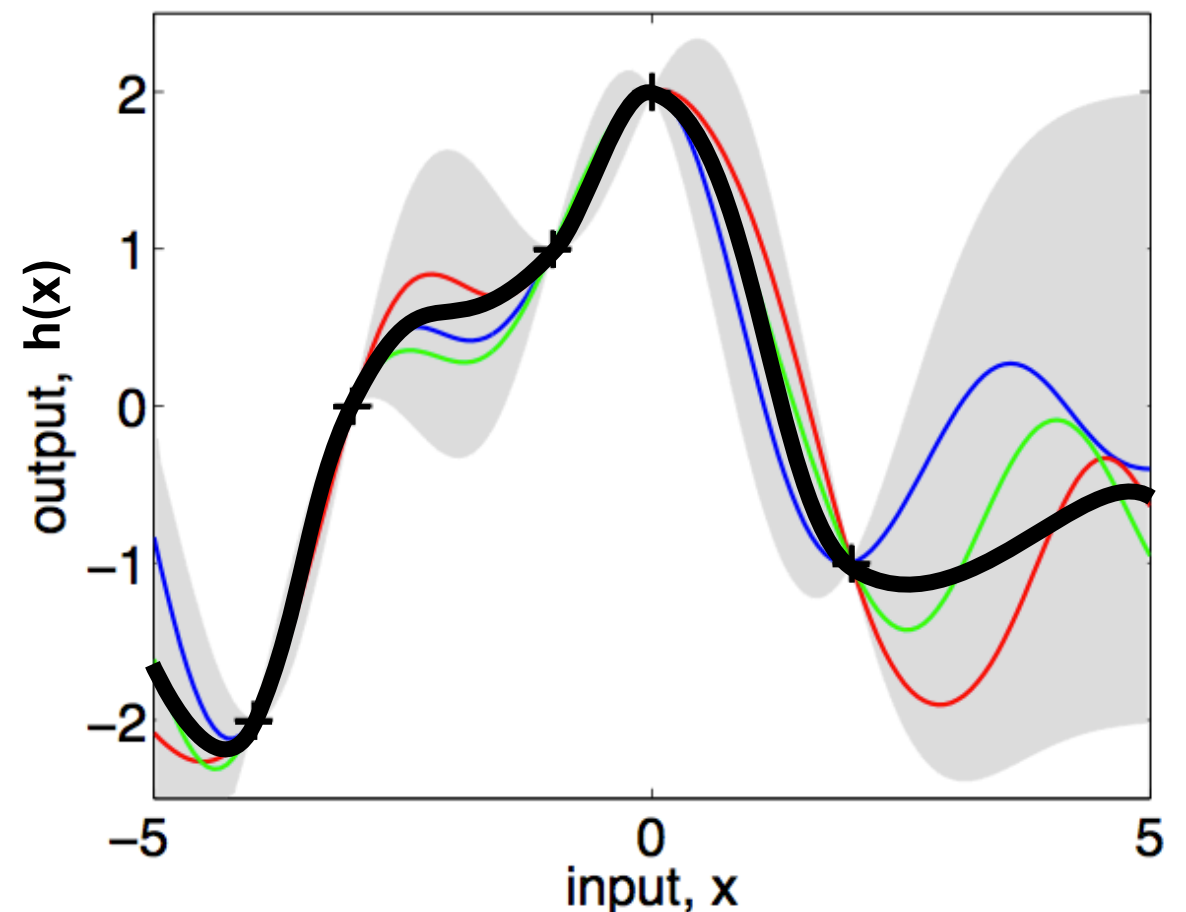
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

where to sample to maximize  $f(\mathbf{x})$ ?



(b), posterior

figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

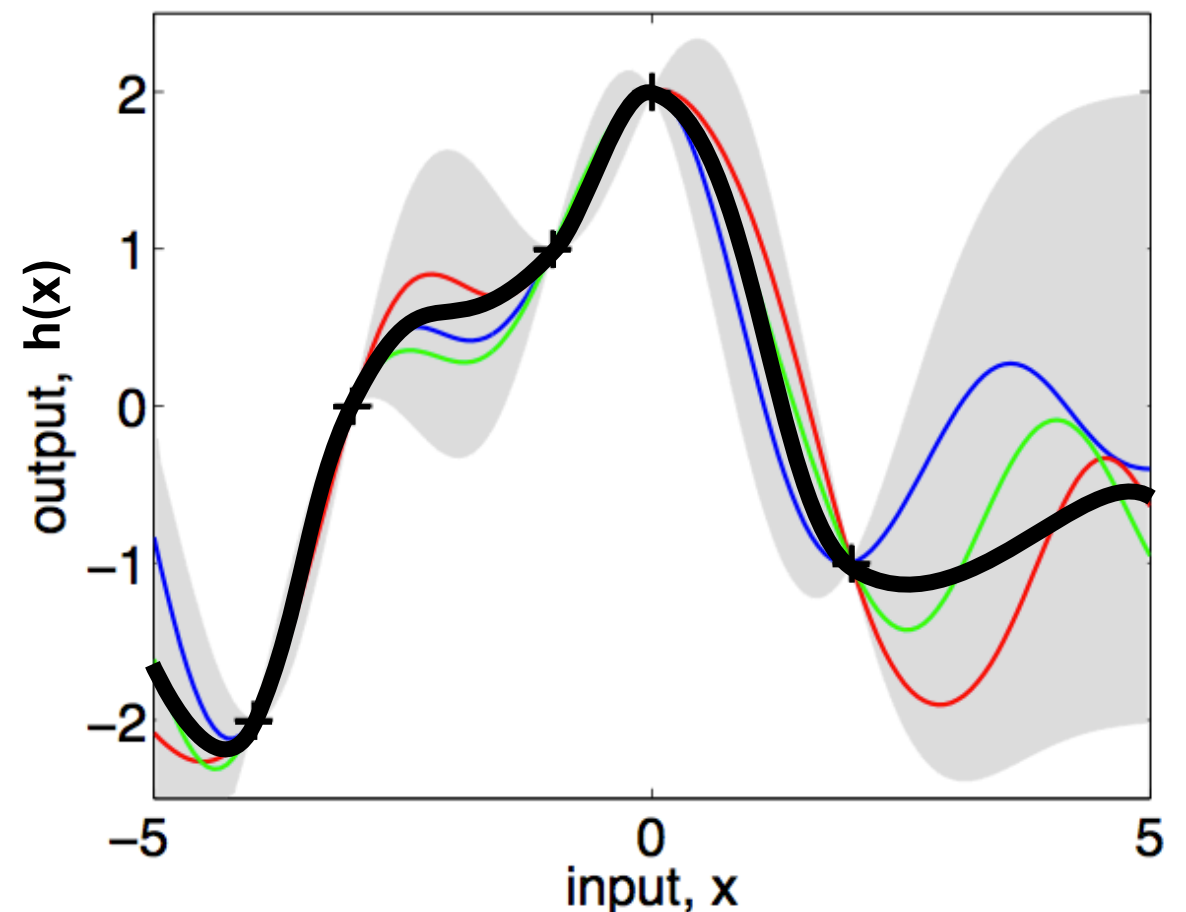
$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

where to sample to maximize  $f(\mathbf{x})$ ?

[Srinivas et al., 2010]

Upper Confidence Bound (GP-UCB)

$$\mu_t(\mathbf{x}) + \sqrt{\beta} \sigma_t(\mathbf{x})$$



(b), posterior

figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

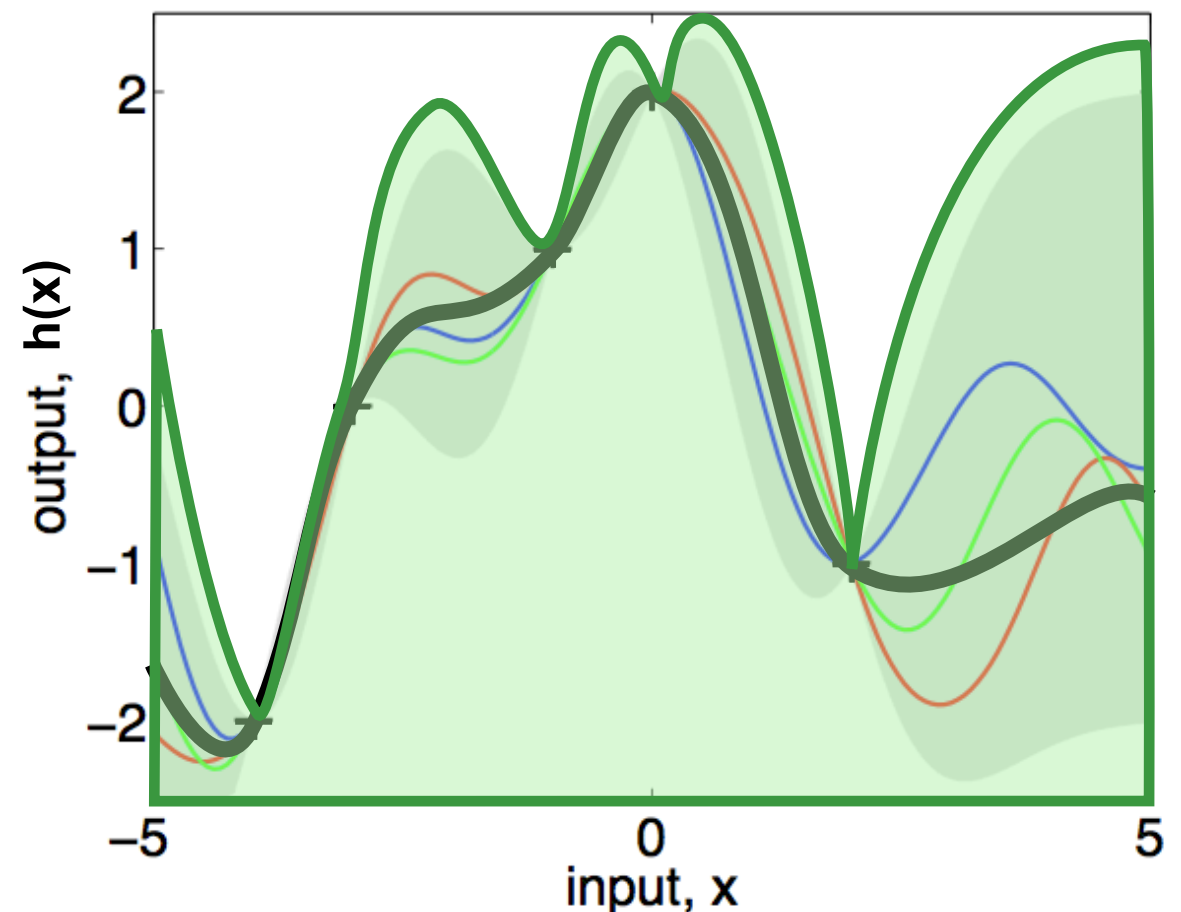
$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

where to sample to maximize  $f(\mathbf{x})$ ?

[Srinivas et al., 2010]

Upper Confidence Bound (GP-UCB)

$$\mu_t(\mathbf{x}) + \sqrt{\beta} \sigma_t(\mathbf{x})$$



(b), posterior

figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

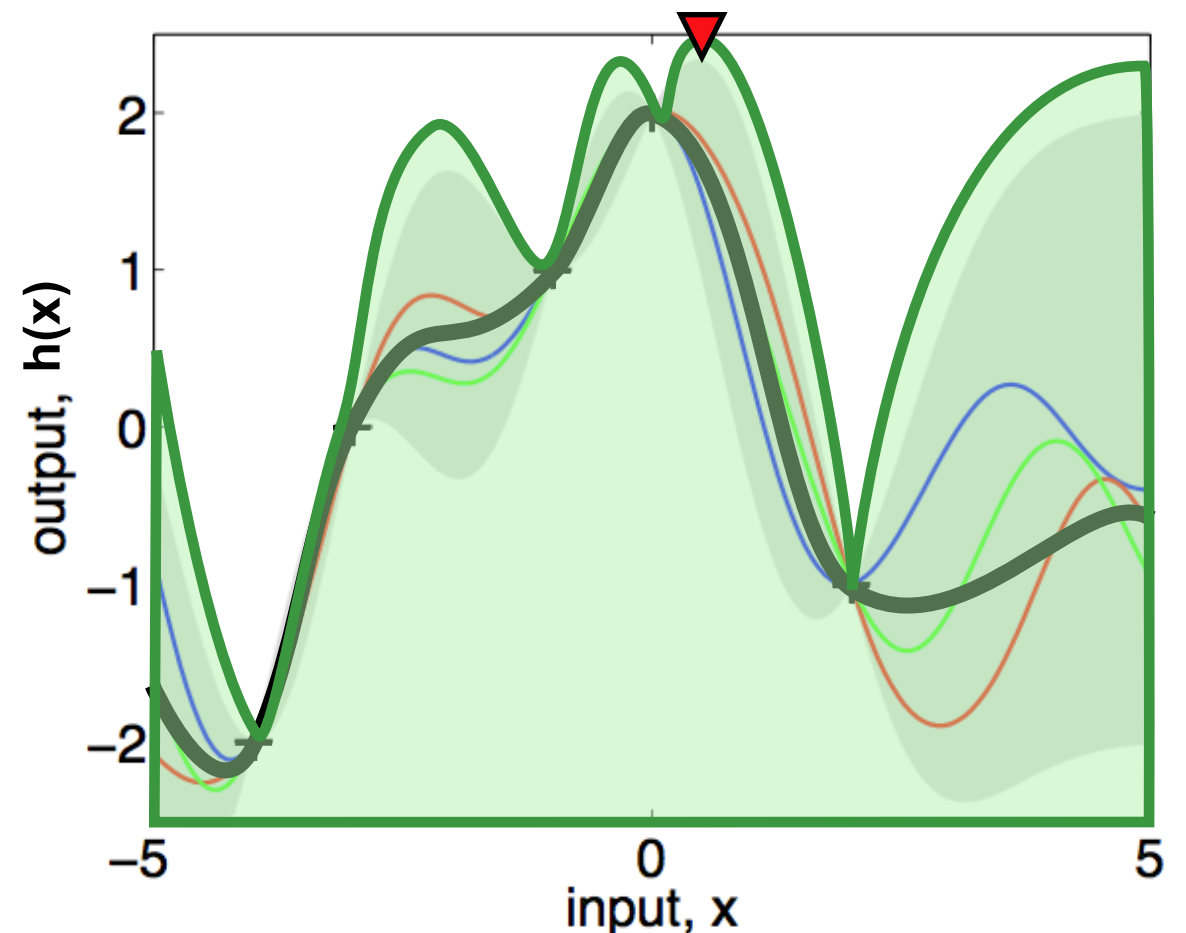
$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

where to sample to maximize  $f(\mathbf{x})$ ?

[Srinivas et al., 2010]

Upper Confidence Bound (GP-UCB)

$$\mu_t(\mathbf{x}) + \sqrt{\beta} \sigma_t(\mathbf{x})$$



(b), posterior

figure credit: [Rasmussen & Williams, 2006]

# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

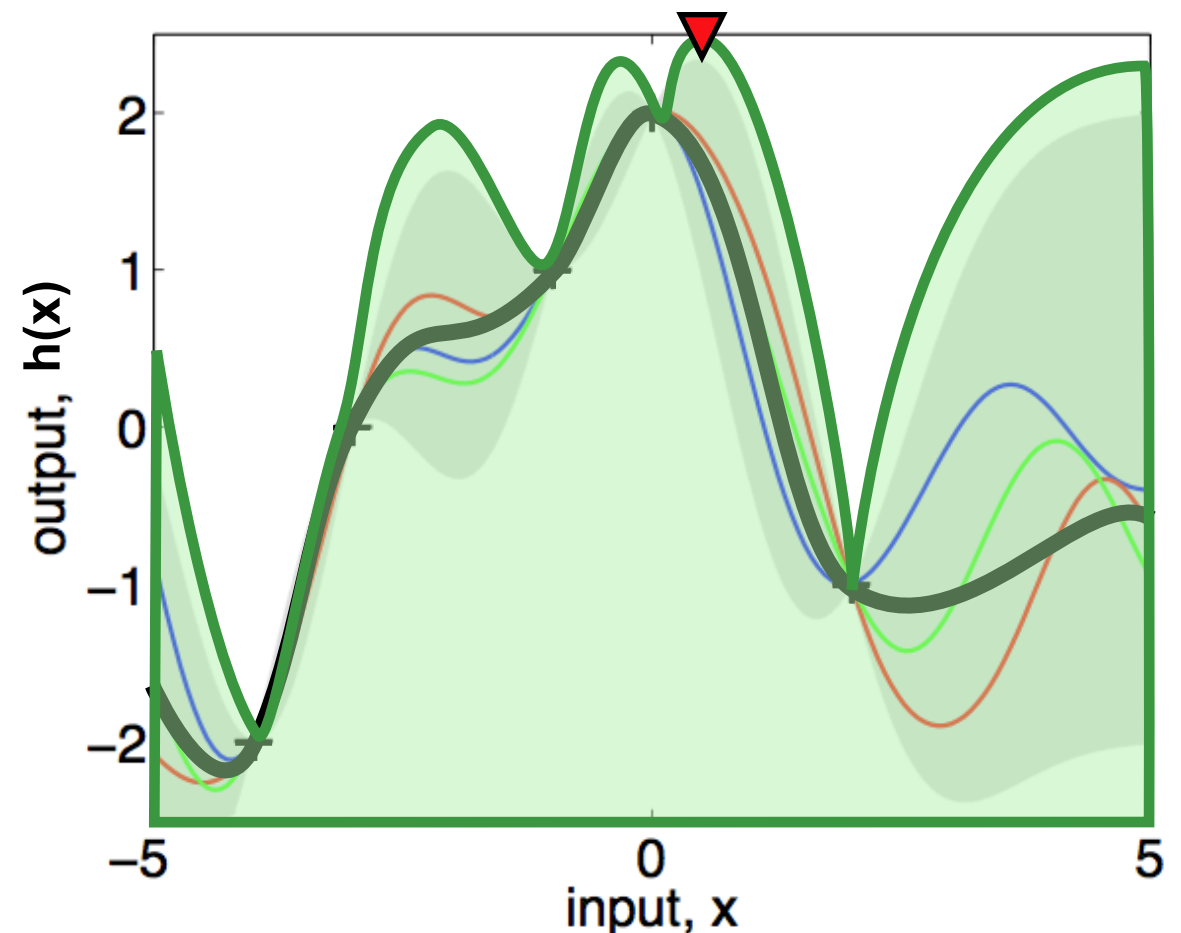
where to sample to maximize  $f(\mathbf{x})$ ?

[Srinivas et al., 2010]  
Upper Confidence Bound (GP-UCB)

$$\mu_t(\mathbf{x}) + \sqrt{\beta} \sigma_t(\mathbf{x})$$

exploitation

exploration



(b), posterior

figure credit: [Rasmussen & Williams, 2006]



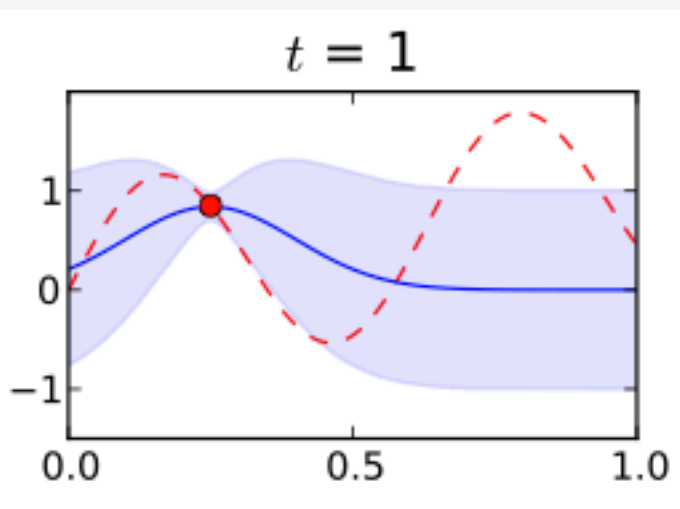
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

figure credit:  
[Brochu et al., 2010]



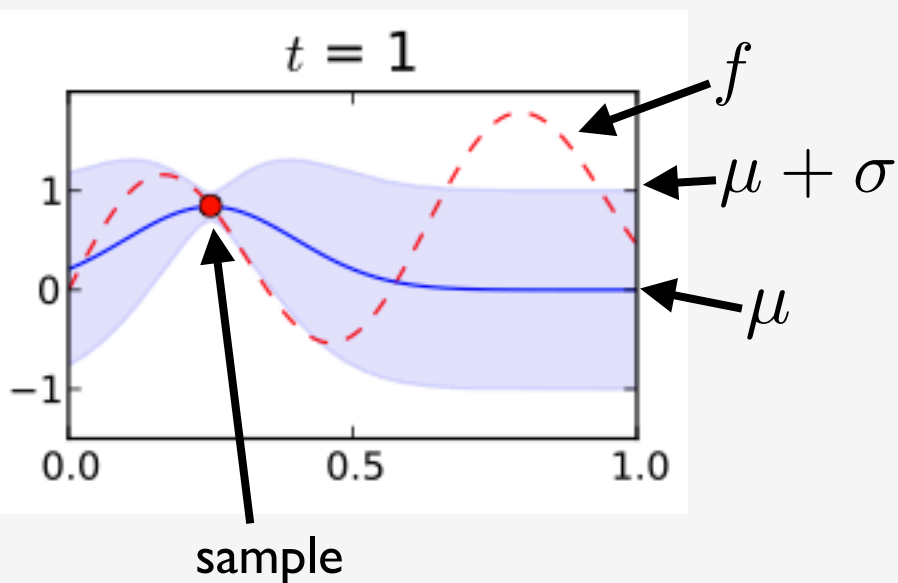
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

figure credit:  
[Brochu et al., 2010]



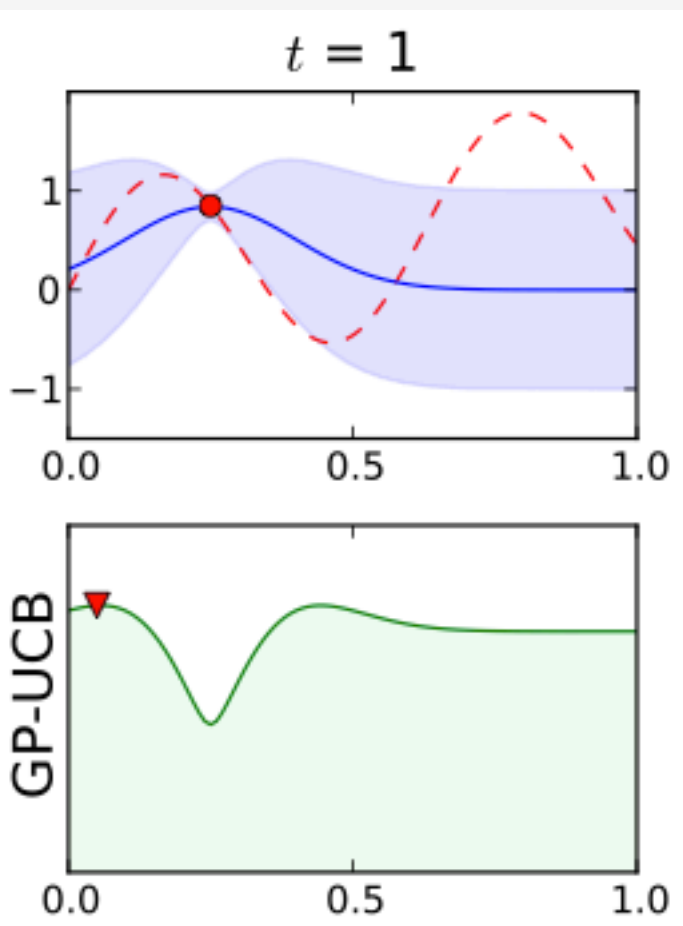
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

figure credit:  
[Brochu et al., 2010]



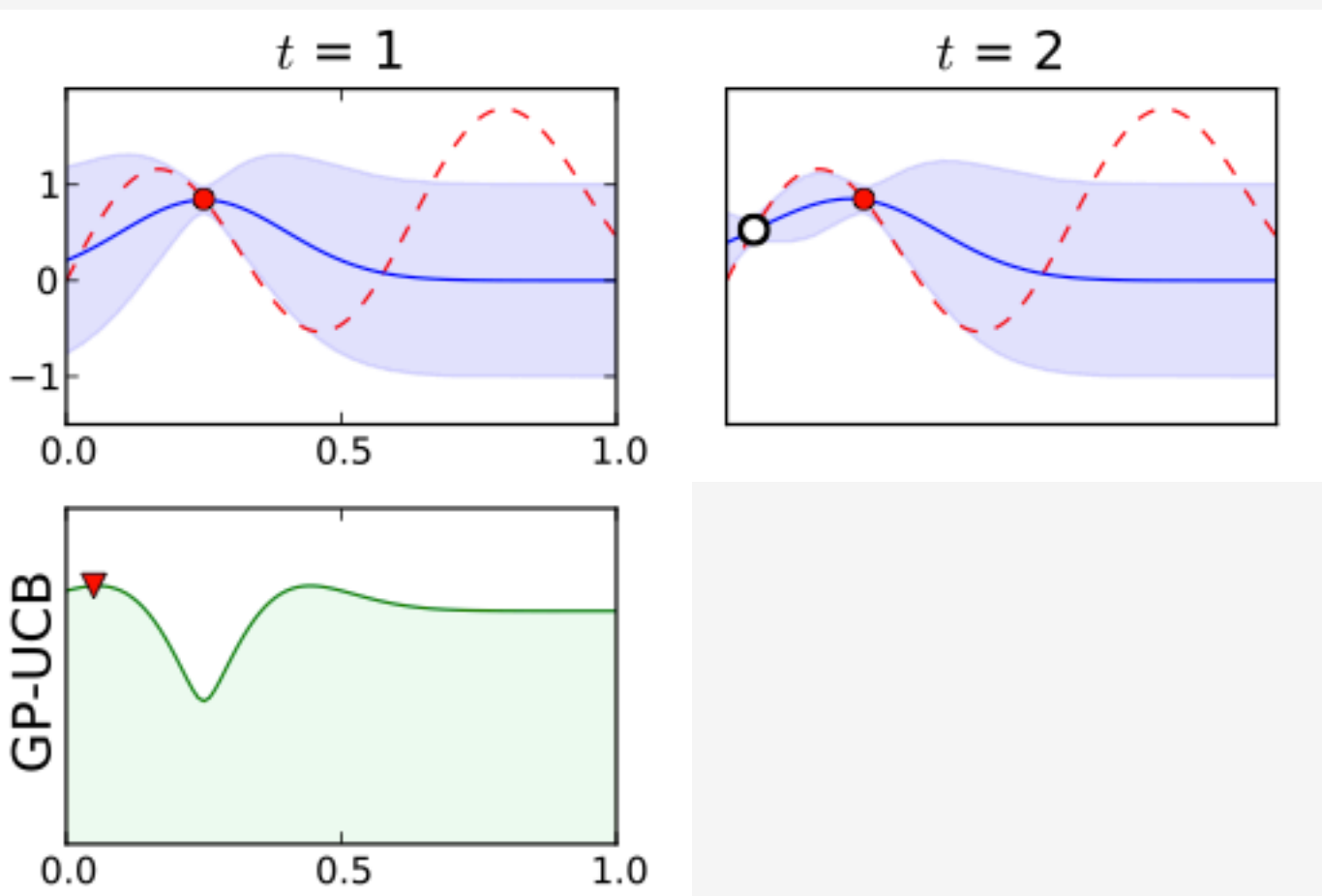
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

figure credit:  
[Brochu et al., 2010]



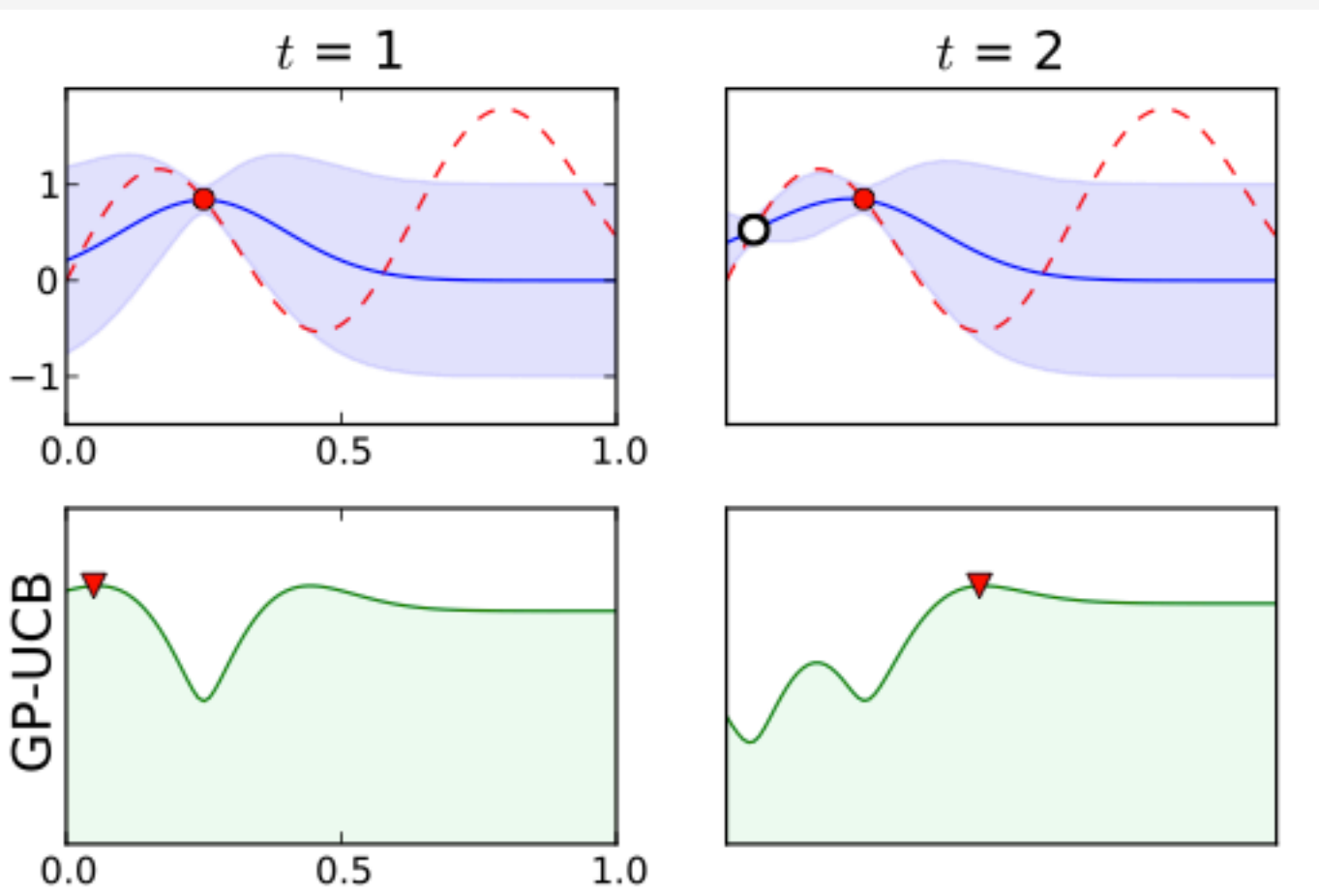
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

figure credit:  
[Brochu et al., 2010]



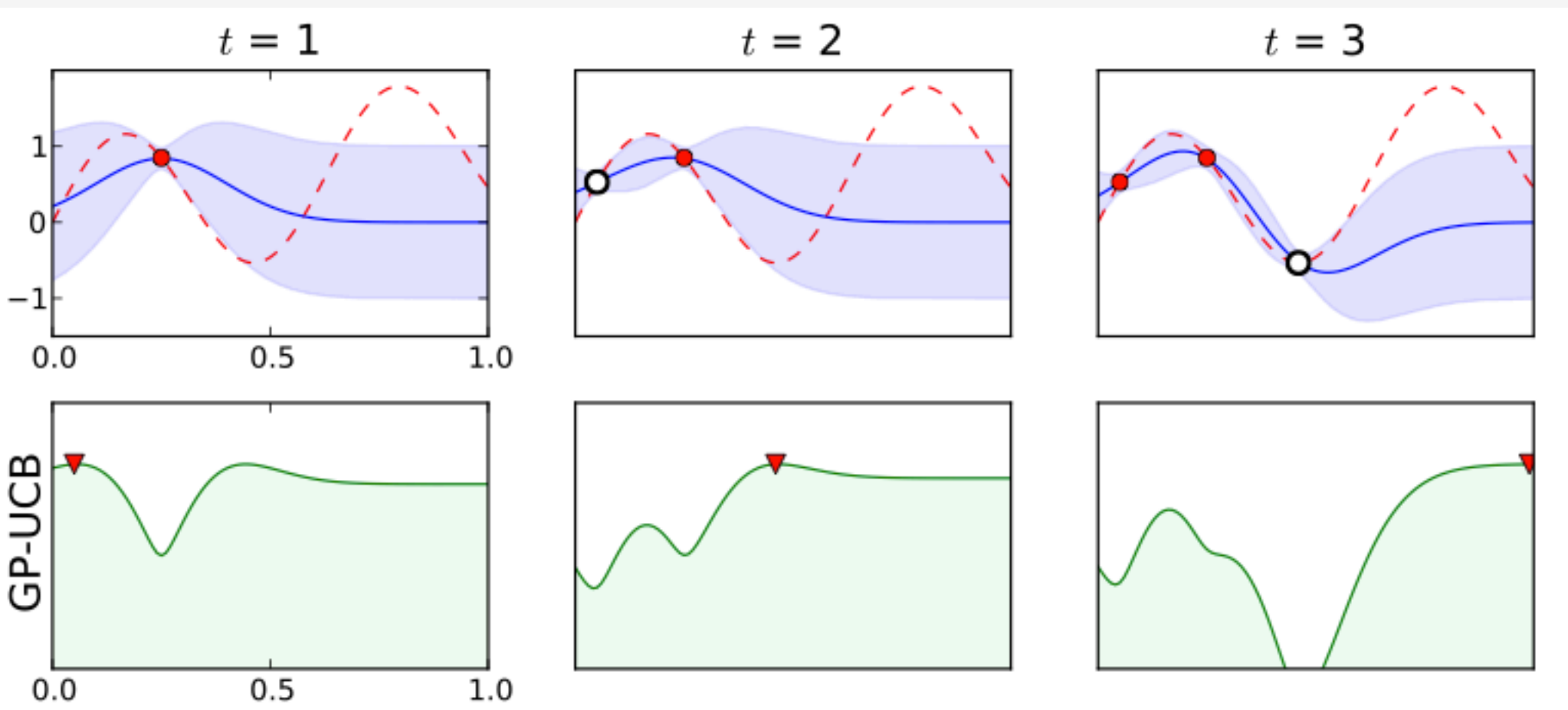
# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

figure credit:  
[Brochu et al., 2010]



# Bayesian Optimization

$$\max_{\mathbf{x}} f(\mathbf{x})$$

Gaussian Processes as surrogates

$$h \sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}'))$$

figure credit:  
[Brochu et al., 2010]

