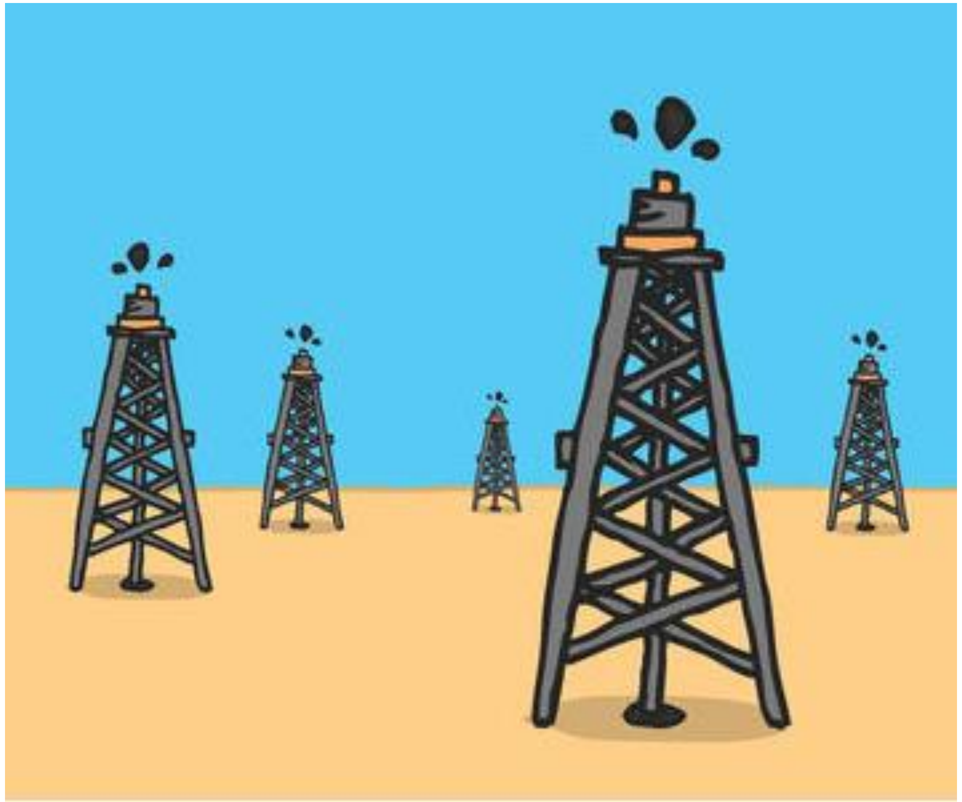
Bayesian Optimization

Minliang Lin

2019/4/26

An Example: Drilling Oil Wells

Where is the best place?

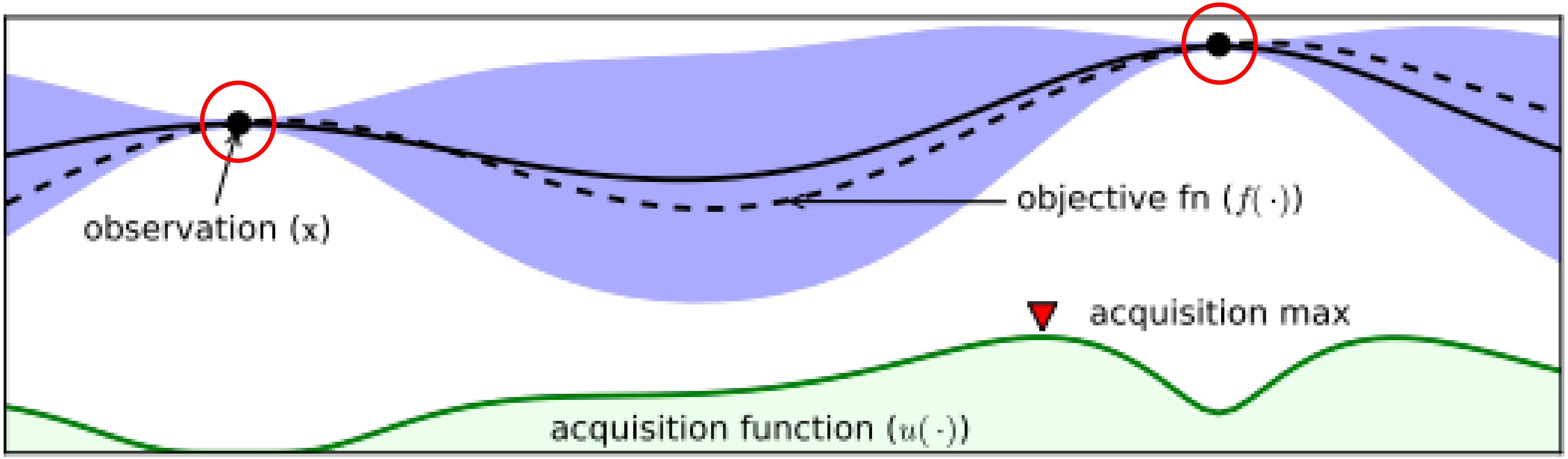


Problem Description

- Given a black box function $f(\mathbf{x})$.
- How to find the global maximum?
- Requirement: sampling from $f(\mathbf{x})$ as few as possible.

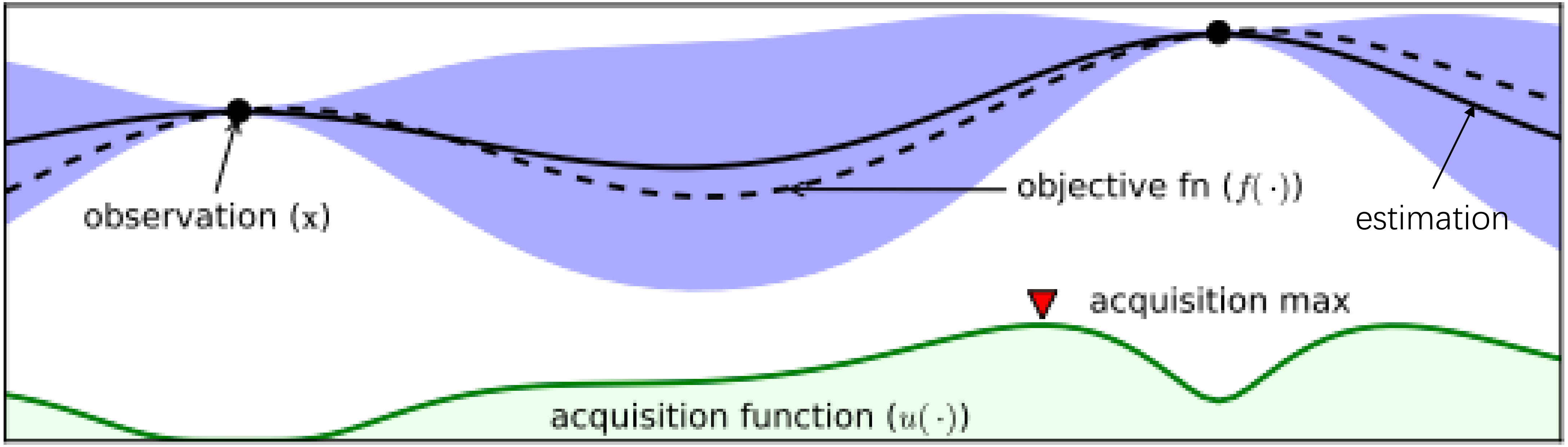
Solution Framework

- Estimation from known information
- Exploitation and Exploration



Estimation

- Interpolation of all known value, with probability.
- Weighting by distance

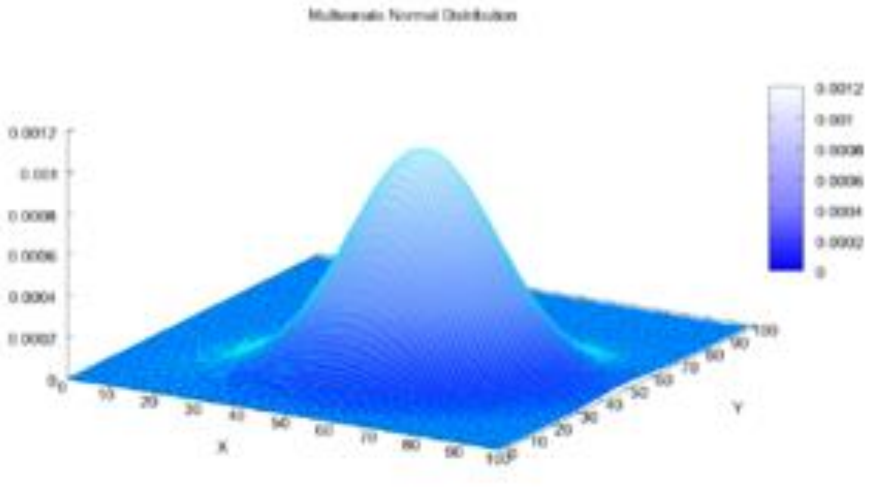


Estimation: Gaussian Process

Multivariate Gaussian distribution

$$\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}}$$



Estimation: Gaussian Process

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}) \qquad \kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

sampled data   parameters to be estimated   Entity of **K**   Distance

Assume we know $(\mathbf{x}, \mathbf{y})$, to get the new value $\mathbf{f}_*$ at a new location:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$

Estimation: Gaussian Process

Assume we know $(\mathbf{x}, \mathbf{y})$, the new value $\mathbf{f}_*$ at a new point

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^T & \mathbf{K}_{**} \end{pmatrix}\right)$$

From some algebra formulae, we have:

$$\boldsymbol{\mu}_* = \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{y}$$

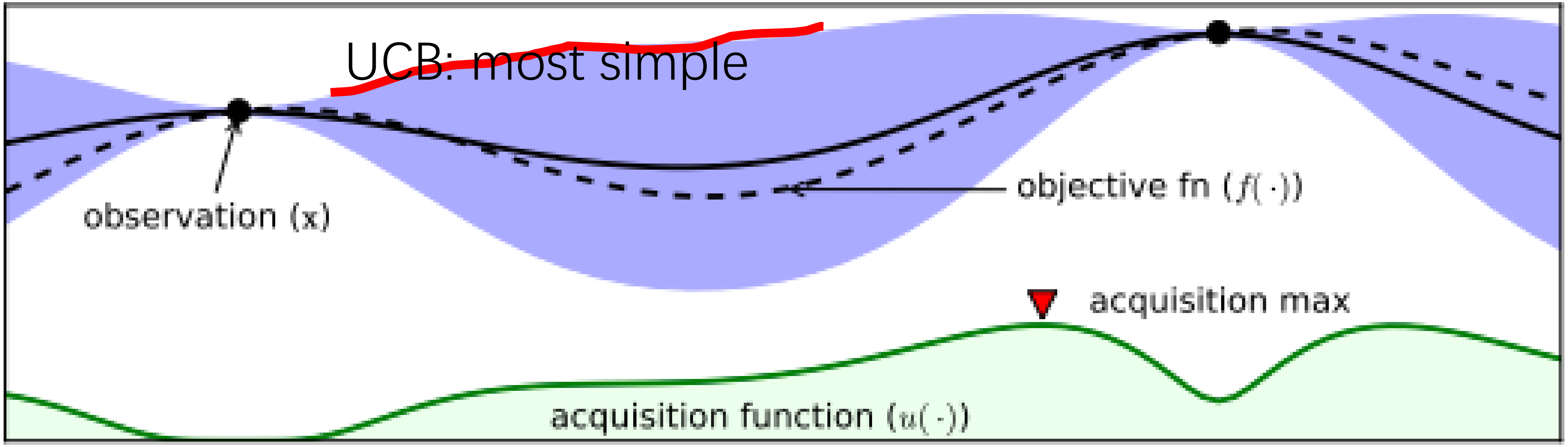
$$\boldsymbol{\Sigma}_* = \mathbf{K}_{**} - \mathbf{K}_*^T \mathbf{K}_y^{-1} \mathbf{K}_*$$

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\frac{1}{2}\|\mathbf{x}_i - \mathbf{x}_j\|^2)$$

Comments

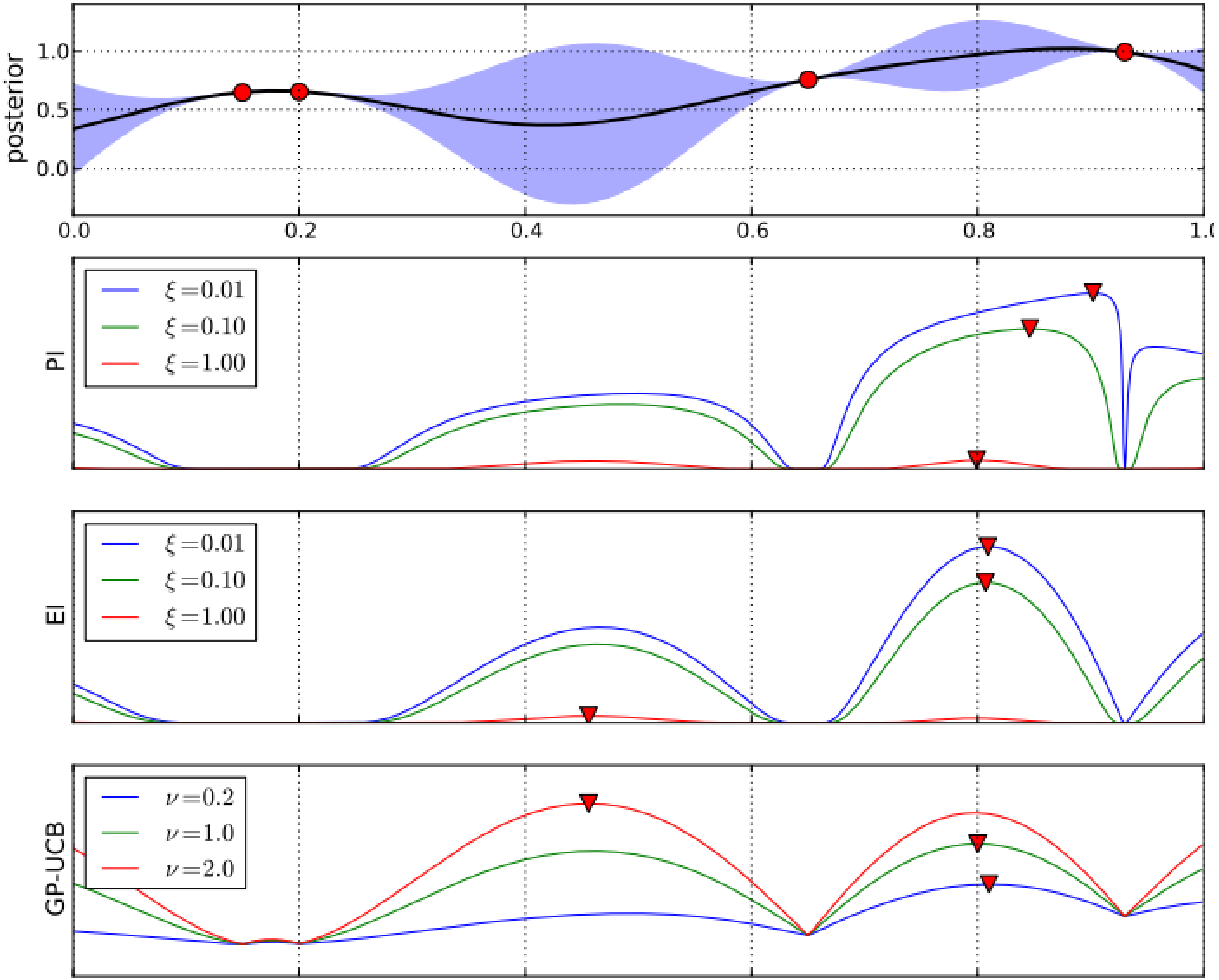
- Estimation (surrogate function) from data **instantly**.
- The choice of distance/kernel function is flexible.
- Linear combination of Gaussian kernel (squared exponential kernel), which is infinitely differentiable.
- Good probability model interpretation.

Exploitation and Exploration Trade-off (Acquisition Function [提取函数])



UCB = $\boldsymbol{\mu}_* + \boldsymbol{\Sigma}_*$

Exploitation and Exploration Trade-off (Acquisition Function)

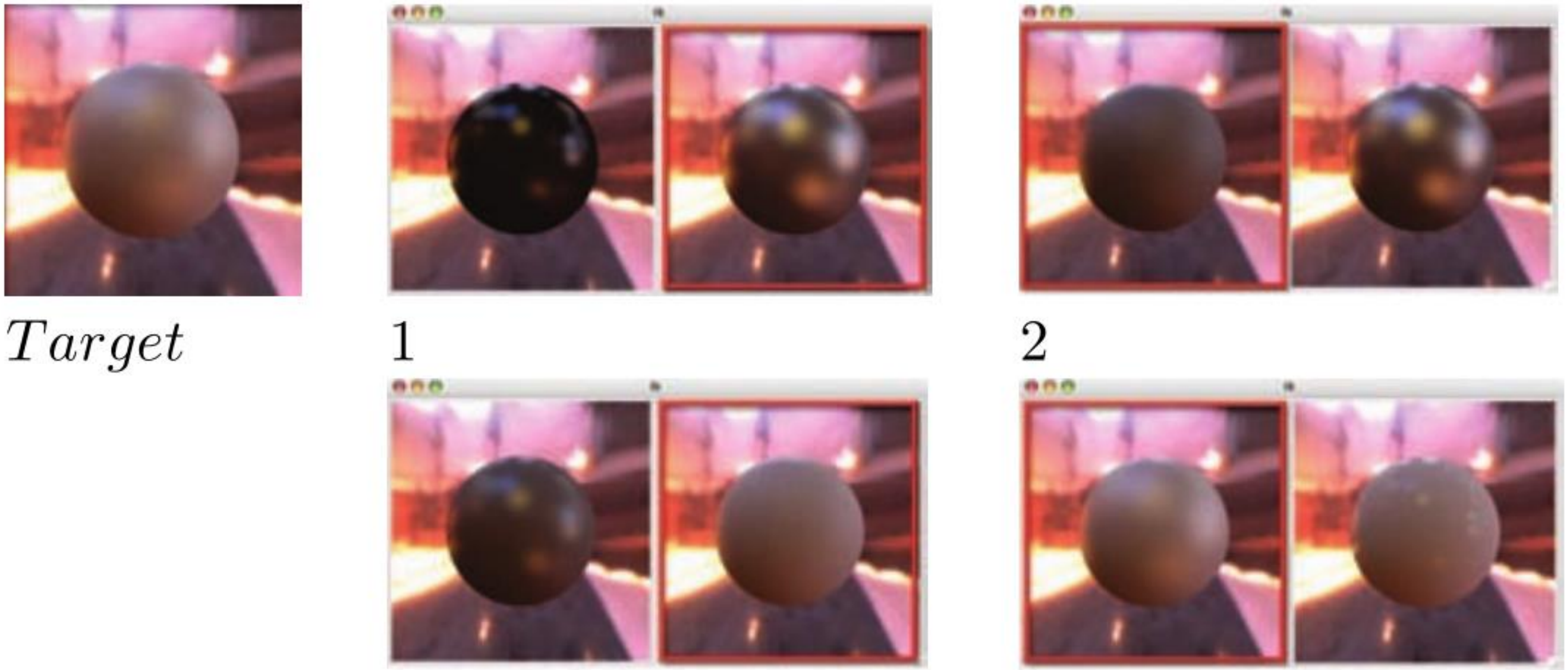


Maximization of Acquisition

- Derivative based (Newton, CG, etc.)
- Derivative free

Can we be cooler?

What about no evaluation?



Target

1

2

3

4

Psychology Model

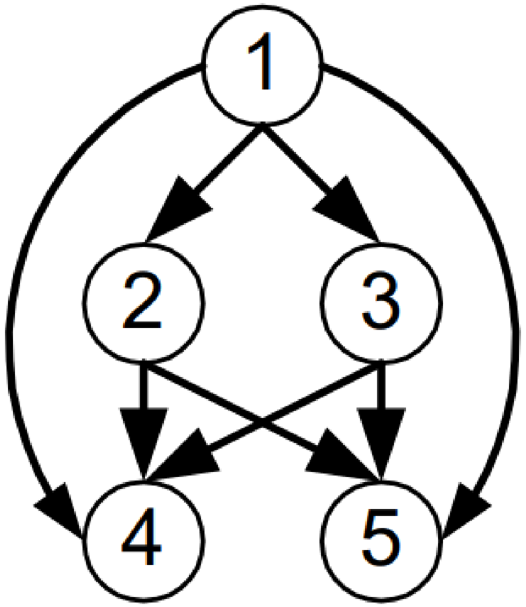
$$v(\mathbf{r}_i) = f(\mathbf{r}_i) + \varepsilon$$

User's rating

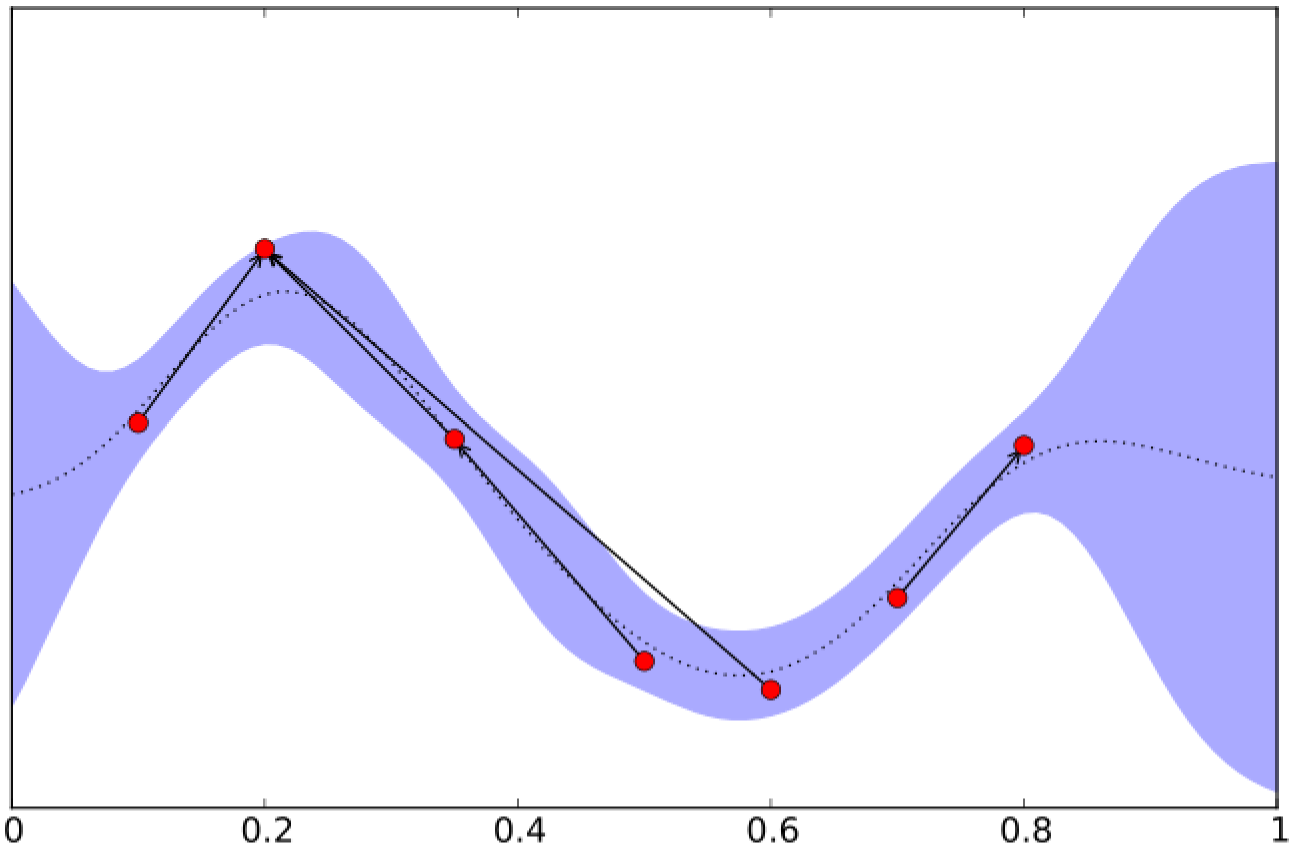
Latent function

Gaussian Noise

Bayesian framework again



Directed Graph



Probability distribution of function

Posterior = Prior * Likelihood / Evidence

f is a finite dimension vector.

$$\mathcal{P}(\boldsymbol{f}|\mathcal{D}) = \frac{\mathcal{P}(\boldsymbol{f})}{\mathcal{P}(\mathcal{D})} \prod_{k=1}^{m} \mathcal{P}\left(v_k \succ u_k | f(v_k), f(u_k)\right)$$

Evidence

$$\mathcal{P}(\boldsymbol{f}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\boldsymbol{f}^T \Sigma^{-1} \boldsymbol{f}\right)$$

Posterior = Prior * Likelihood / Evidence

f is a finite dimension vector.

$$\mathcal{P}(\boldsymbol{f}|\mathcal{D}) = \frac{\mathcal{P}(\boldsymbol{f})}{\mathcal{P}(\mathcal{D})} \prod_{k=1}^{m} \mathcal{P}\left(v_k \succ u_k | f(v_k), f(u_k)\right)$$

Evidence

$$\begin{aligned} P(\mathbf{r}_i \succ \mathbf{c}_i | f(\mathbf{r}_i), f(\mathbf{c}_i)) &= P(v(\mathbf{r}_i) > v(\mathbf{c}_i) | f(\mathbf{r}_i), f(\mathbf{c}_i)) \\ &= P(\varepsilon - \varepsilon < f(\mathbf{r}_i) - f(\mathbf{c}_i)) \\ &= \Phi(Z_i), \end{aligned}$$

Laplacian Approximation

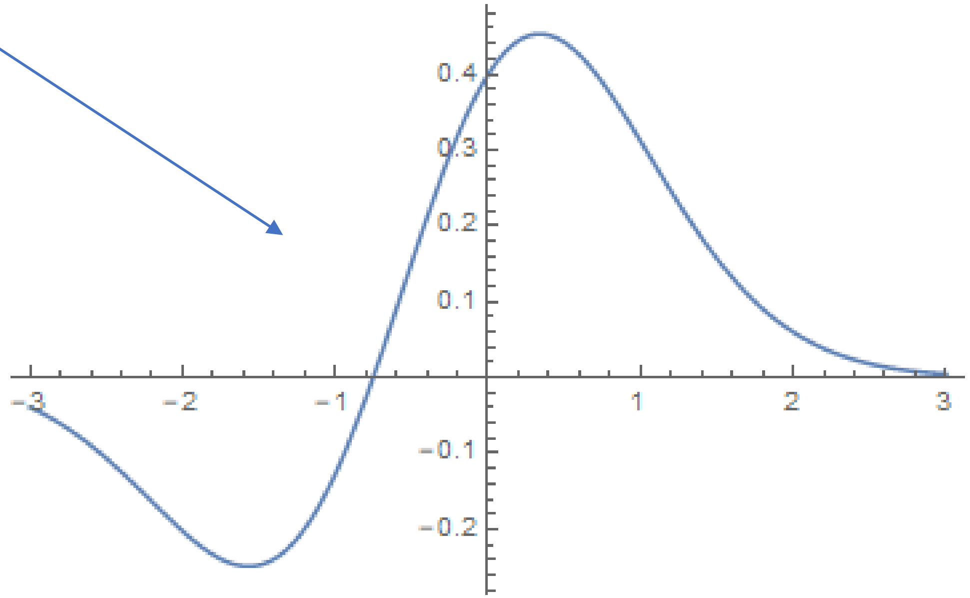
Maximize A Posteriori, **g = 0**, $\mathbf{H} = \mathbf{K}^{-1} + \mathbf{C}$

$$\log P(\mathbf{f}|\mathcal{D}) = \log P(\widehat{\mathbf{f}}|\mathcal{D}) + \mathbf{g}^T(\mathbf{f} - \widehat{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \widehat{\mathbf{f}})^T\mathbf{H}(\mathbf{f} - \widehat{\mathbf{f}})$$

f of Max a posteriori

C

$$\mathbf{C}_{m,n} = \frac{1}{2\sigma^2}\sum_{i=1}^{M} h_i(\mathbf{x}_m)h_i(\mathbf{x}_n)\left[\frac{\phi(Z_i)}{\Phi^2(Z_i)} + \frac{\phi^2(Z_i)}{\Phi(Z_i)}Z_i\right]$$



Newton Method

$$\mathbf{f}^{\text{new}} = \mathbf{f}^{\text{old}} - \mathbf{H}^{-1}\mathbf{g}\,\big|_{\mathbf{f}=\mathbf{f}^{\text{old}}}$$

Comments

- This is a general framework to many problems.
- More prior could be incorporated in.

Reference

- Preference Learning with Gaussian Processes
- A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning
- Active Preference Learning with Discrete Choice Data
- Gaussian Processes for Machine Learning

Thanks