ECE382N.23: Embedded System Design and Modeling

Lecture 12 – Bayesian Optimization

Sources:
R. Calandra, Y. Chu, M. Deisenroth
(Hyperparameter tuning in AutoML)

Andreas Gerstlauer
Electrical and Computer Engineering
University of Texas at Austin
gerstl@ece.utexas.edu

The University of Texas at Austin
Chandra Department of Electrical and Computer Engineering
Cockrell School of Engineering

Lecture 12: Outline

- **Bayesian optimization**
  - Surrogate function
  - Gaussian processes
  - Bayesian regression
  - Acquisition function
- **Applications to system mapping & exploration**
  - Multi-objective optimization

Optimization Problem

- **Find decisions x that minimize cost function $f(x)$**

$$x^* = \underset{x}{\operatorname{argmin}}\, f(x)$$

  - $f(x)$ is unknown (black box) & complex (non-convex)
  - Can evaluate (sample) $y_i = f(x_i)$ for given decisions x_i
  - But expensive to evaluate (e.g. simulation)

➢ **Learn and optimize a surrogate function $\tilde{f}(x) \sim f(x)$**

$$x^* = \underset{x}{\operatorname{argmin}}\, \tilde{f}(x)$$

  - Surrogate $\tilde{f}(x)$ that captures behavior but is optimizable
  - Learn $\tilde{f}(x)|_D$ from observations $D = \{ (f(x_i), y_i) \}$
  - Form of regression, but with uncertainty
  - ➢ Bayesian (probabilistic) regression

Bayesian Regression

- **Bayes' theorem (distribution case)**

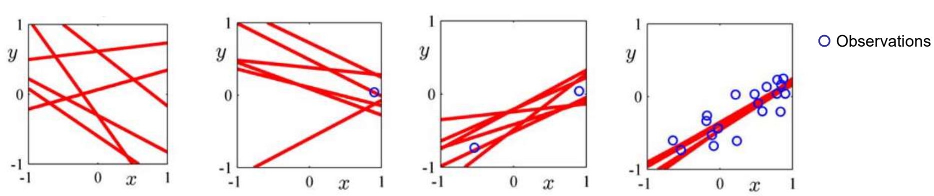
$$p(A|B) = \frac{p(B|A) \cdot p(A)}{p(B)}$$

Posterior distribution given B → $p(A|B)$; Likelihood of B given A → $p(B|A)$; Prior belief (distribution of A) → $p(A)$; Marginal likelihood of evidence B (normalizing constant) → $p(B)$

- **In regression (distributions over functions)**

$$p(\tilde{f}|D) = \frac{p(D|\tilde{f}) \cdot p(\tilde{f})}{p(D)} \propto p(D|\tilde{f}) \cdot p(\tilde{f})$$

Posterior function distribution with a', b' → $p(\tilde{f}|D)$; Prior knowledge/guess, e.g. linear $y = ax+b$, with a,b normally distributed → $p(\tilde{f})$



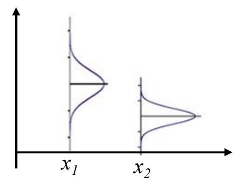
Source: C. M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.

Gaussian Process

- **Multi-variate Gaussian distribution**

$$\begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad \boldsymbol{\mu} = \begin{bmatrix} \mu_{X_1} \\ \mu_{X_2} \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \sigma_{X_1}^2 & \sigma_{X_1}\sigma_{X_2} \\ \sigma_{X_2}\sigma_{X_1} & \sigma_{X_2}^2 \end{bmatrix}$$

(Co-)variance matrix



- **Gaussian process *GP***
  - Extension to infinite number of variables
  - Generalized function distribution
    - Normal distribution at each x with mean $m(x)$ and variance $k(x,x)$
    - "Kernel" $k(x,x')$ controls possible function (interpolation) shapes

$$p(\tilde{f}(x)) \sim GP(m(x), k(x,x'))$$

Mean function — Co-variance function

- ➢ Bayesian regression

$$p(\tilde{f}|D) = GP(m'(\cdot), k'(\cdot)) \qquad m'(\cdot), k'(\cdot) \leftarrow m(\cdot), k(\cdot), \mathrm{D}$$

https://distill.pub/2019/visual-exploration-gaussian-processes/

Bayesian Optimization

1. **Initialize $D = \{\}$**
   Initialize GP prior with selected $m(x)$ and $k(x,x')$
   - E.g. $m(x)=0$ and $k(x,x')$ as exponential square
     - GP hyperparameters for $m(x)$ and $k(x,x')$

2. **Repeat until stop criteria**
   a. Select point x_t to sample and observe its value $f(x_t)$
      - ➢ Maximize acquisition function $x_t = \mathrm{argmax}_x\, a(m(x), k(x,x))$
   b. Add to dataset $D = D \cup \{ (x_t, f(x_t)) \}$
   c. Update $m(x)$ and $k(x,x')$ using Bayes' rule

3. **Return best input in data set $x^* = \underset{t}{\mathrm{argmin}}\, f(x_t)$**

Acquisition Function

- **Balance between exploration and exploitation**
  - Exploration: choose point with high uncertainty (variance)
  - Exploitation: choose point with low expected goal (mean)
- **Common examples**
  - Probability of improvement
  - Expected improvement
  - Upper confidence bound $\mathrm{m}(x) + \kappa \cdot k(x, x)$
- **Optimizing the acquisition function** $\mathrm{argmax}_x\, a(m(x), k(x, x))$
  - Finding the global maximum can by itself be challenging
  - E.g. using some form of gradient descent

Example (Prior)

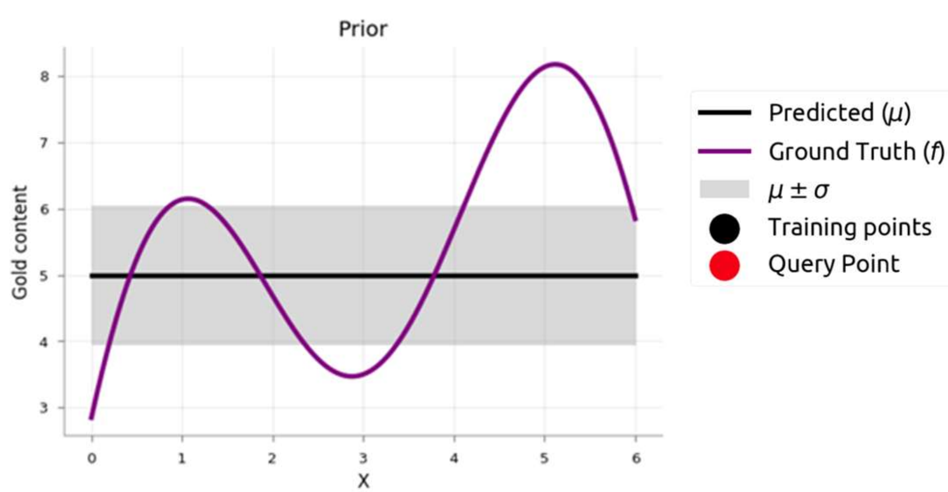
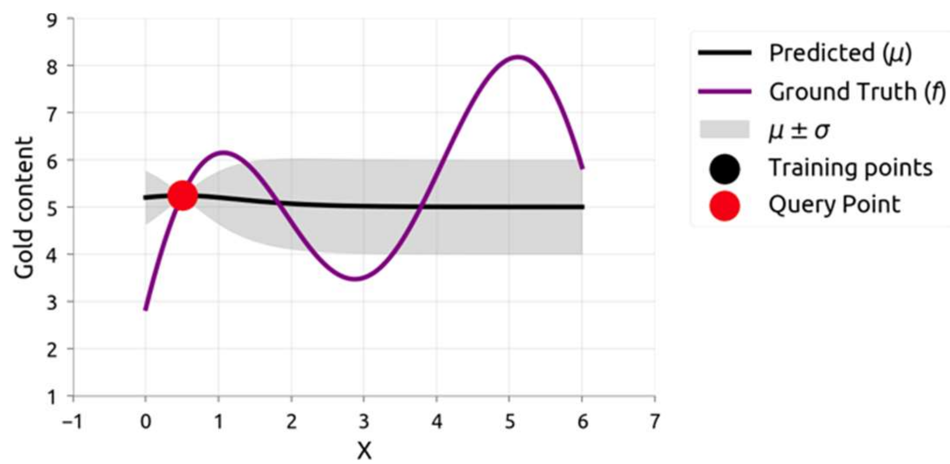
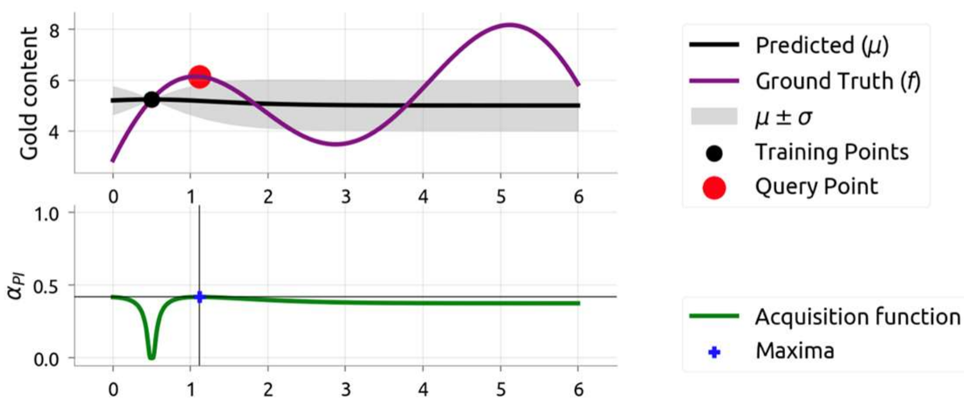


Image sources: https://distill.pub/2020/bayesian-optimization/

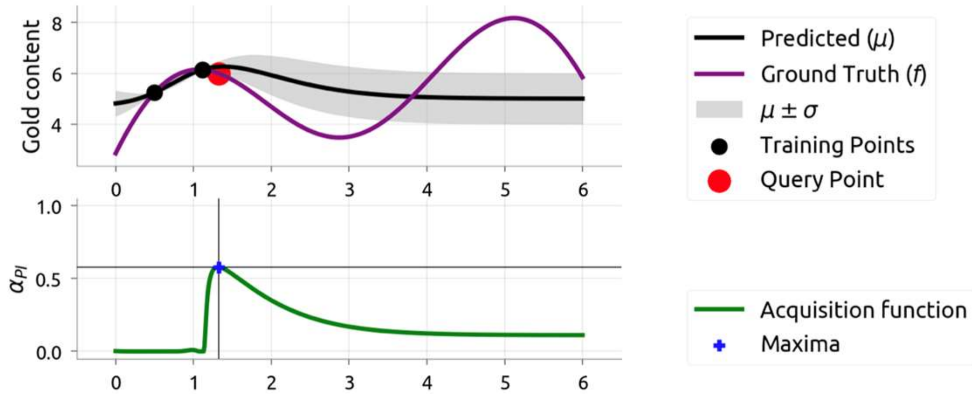
Example (Iteration 0)



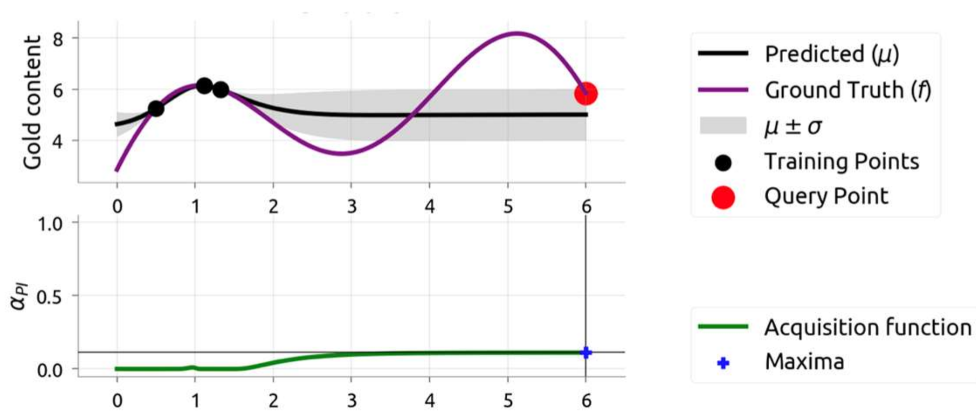
Example (Iteration 1)



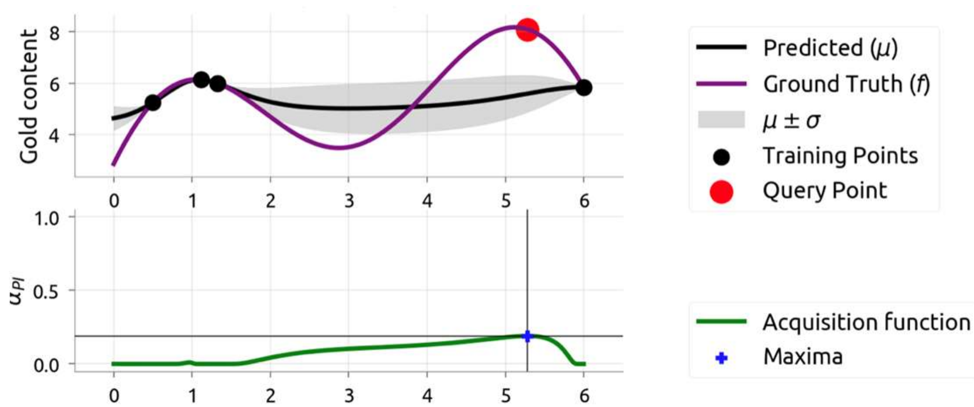
Example (Iteration 2)



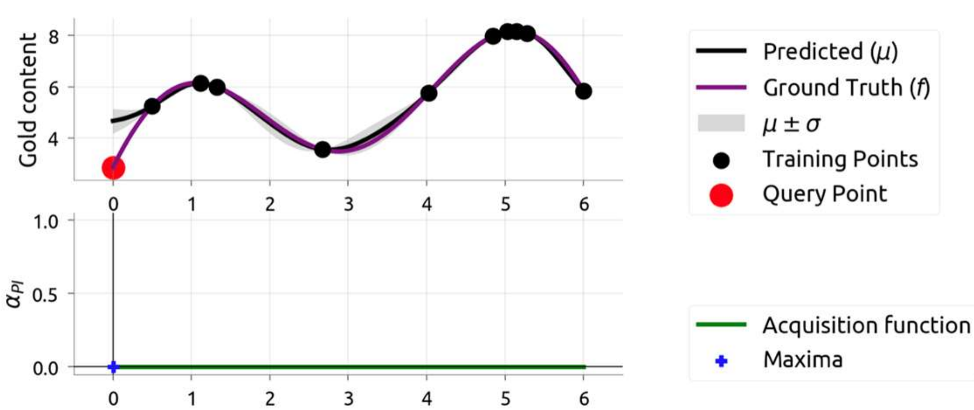
Example (Iteration 3)



Example (Iteration 4)



Example (Iteration *n*)



Multi-Objective Bayesian Optimization

- **Common way to scalarize multiple objectives into one**

$$x^* = \underset{x}{\operatorname{argmin}}\, \tilde{g}(f_1(x), \dots, f_n(x))$$

  - E.g., Chebyshev scalarization

$$g(x) = \max_i \left(\lambda_i f_i(x)\right) + \rho \sum_i \lambda_i f_i(x)$$

- **Alternatively, fold into acquisition function**
  - Goal is to explore Pareto front

https://botorch.org/