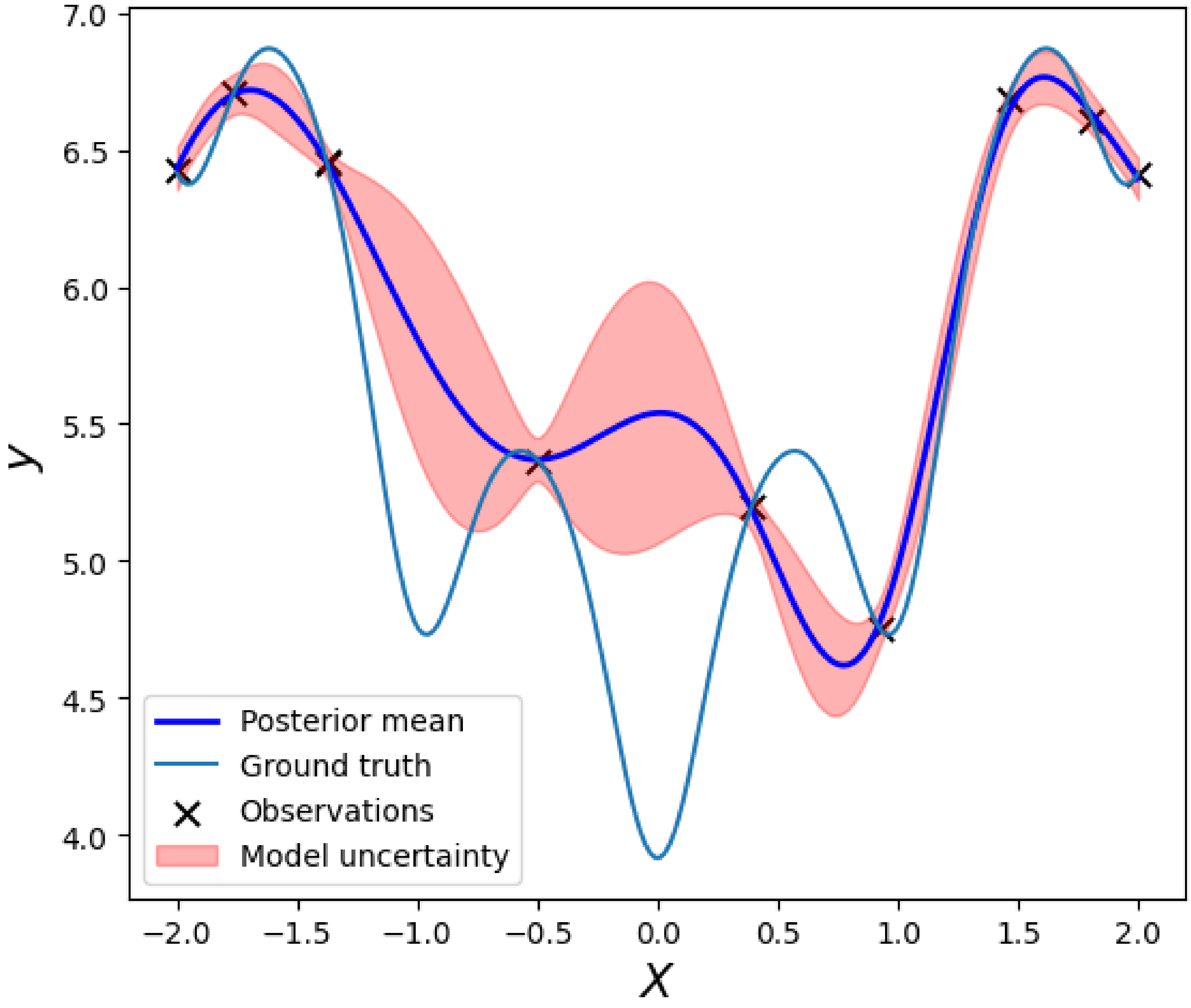
What have we learned from lecture 1:

- Gaussian Process
- Kernel and kernel parameters
- Kernel Priors
- Noise Priors
- Posteriors
- (Acquisition function)

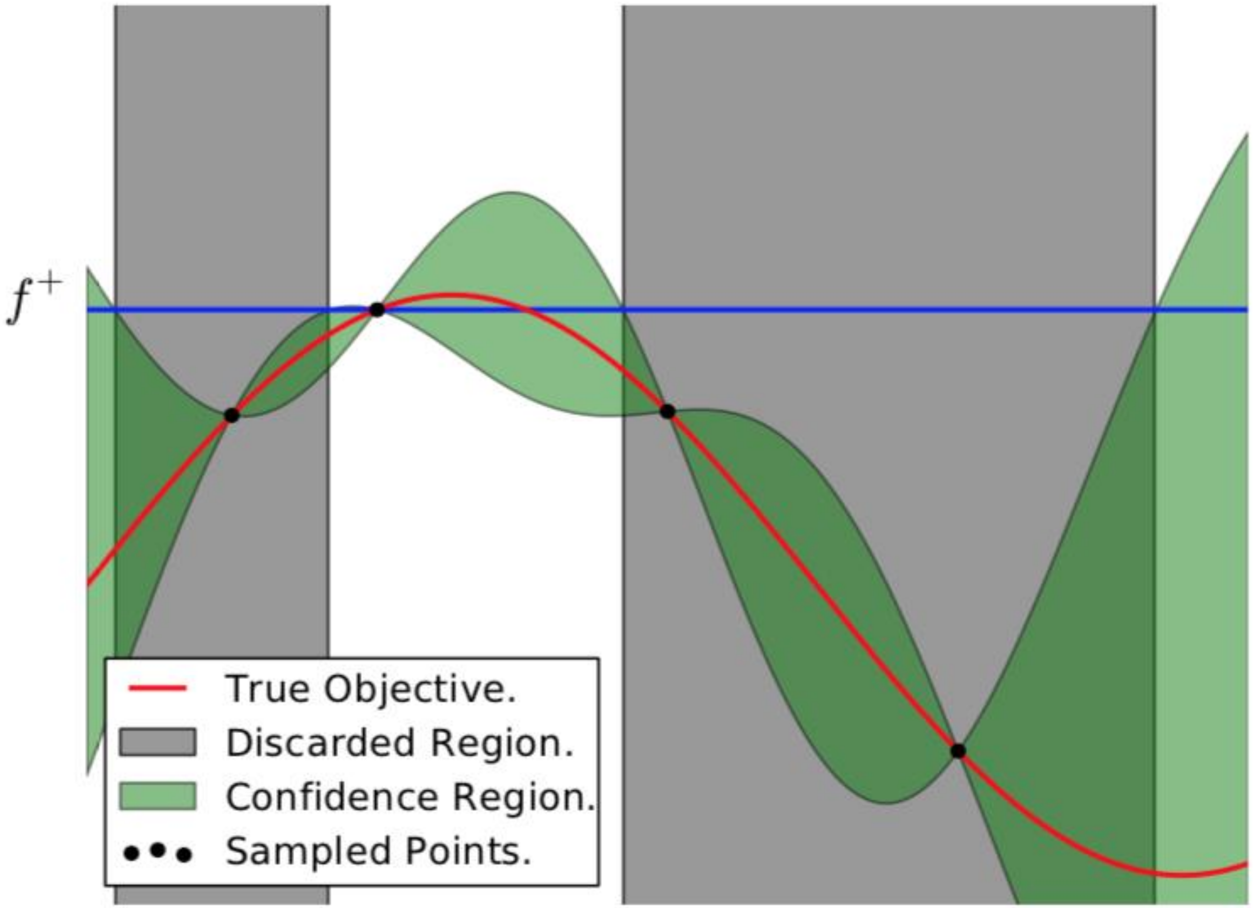
What will we learn:

- Bayesian Optimization
- Bayesian Optimization based on Gaussian Process
- Acquisition Functions



7.0
6.5
6.0
5.5
5.0
4.5
4.0
y
−2.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
X
Posterior mean
Ground truth
Observations
Model uncertainty

Bayesian Optimization

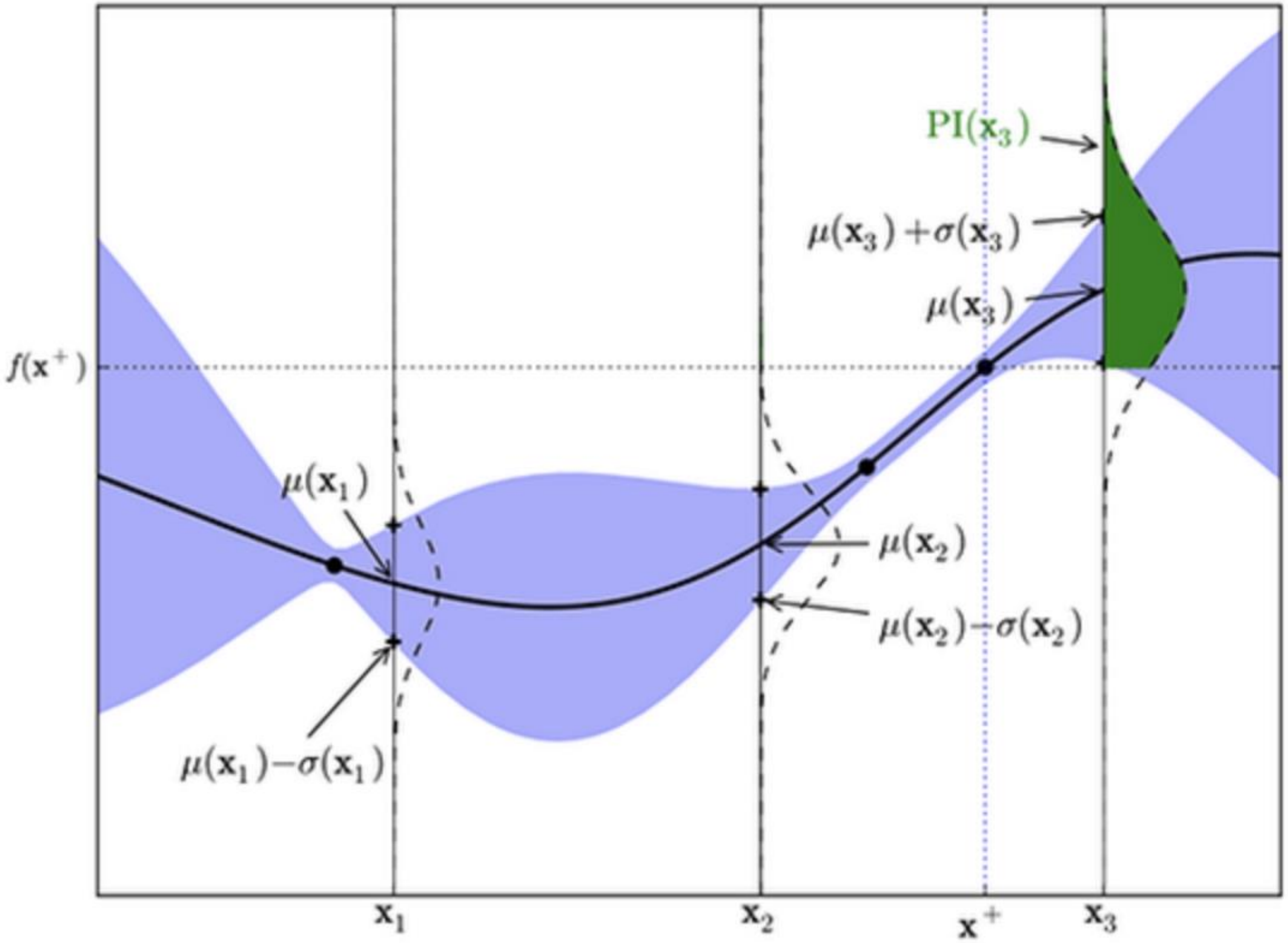


N. de Freitas et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization , *Proceedings of the IEEE* **104**, 148 (2015)

- We have some measurements in space X, and we want to maximize some property f(X).
- How can we decide what point to measure next to best maximize f?
- We need to balance the exploration of the space with exploitation of regions near we have already know

Acquisition Functions

Probability of Improvement
Acquisition Function



1. **Upper confidence bound:** simplest possible - just take the upper confidence bound from the prediction

2. **Probability of Improvement:** Integral from current functional maximum to upper limit of distribution as test point

3. **Expected Improvement:** Instead of probability of improvement, we want to maximize the expected increase in the function value

4. **There are (always) more…**

Expected improvement is defined as

$$\mathrm{EI}(\mathbf{x}) = \mathbb{E}\max(f(\mathbf{x}) - f(\mathbf{x}^+), 0) \qquad (1)$$

where $f(\mathbf{x}^+)$ is the value of the best sample so far and $\mathbf{x}^+$ is the location of that sample i.e. $\mathbf{x}^+ = \mathrm{argmax}_{\mathbf{x}_i \in \mathbf{x}_{1:t}} f(\mathbf{x}_i)$. The expected improvement can be evaluated analytically under the GP model[3]:

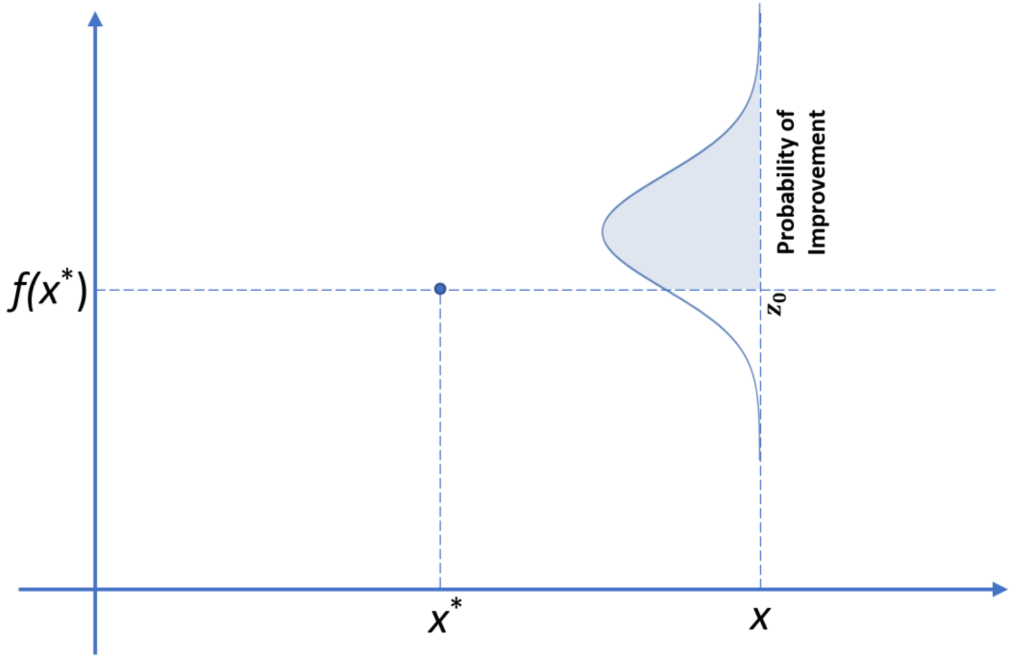
$$\mathrm{EI}(\mathbf{x}) = \begin{cases} (\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi)\Phi(Z) + \sigma(\mathbf{x})\phi(Z) & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases} \qquad (2)$$

where

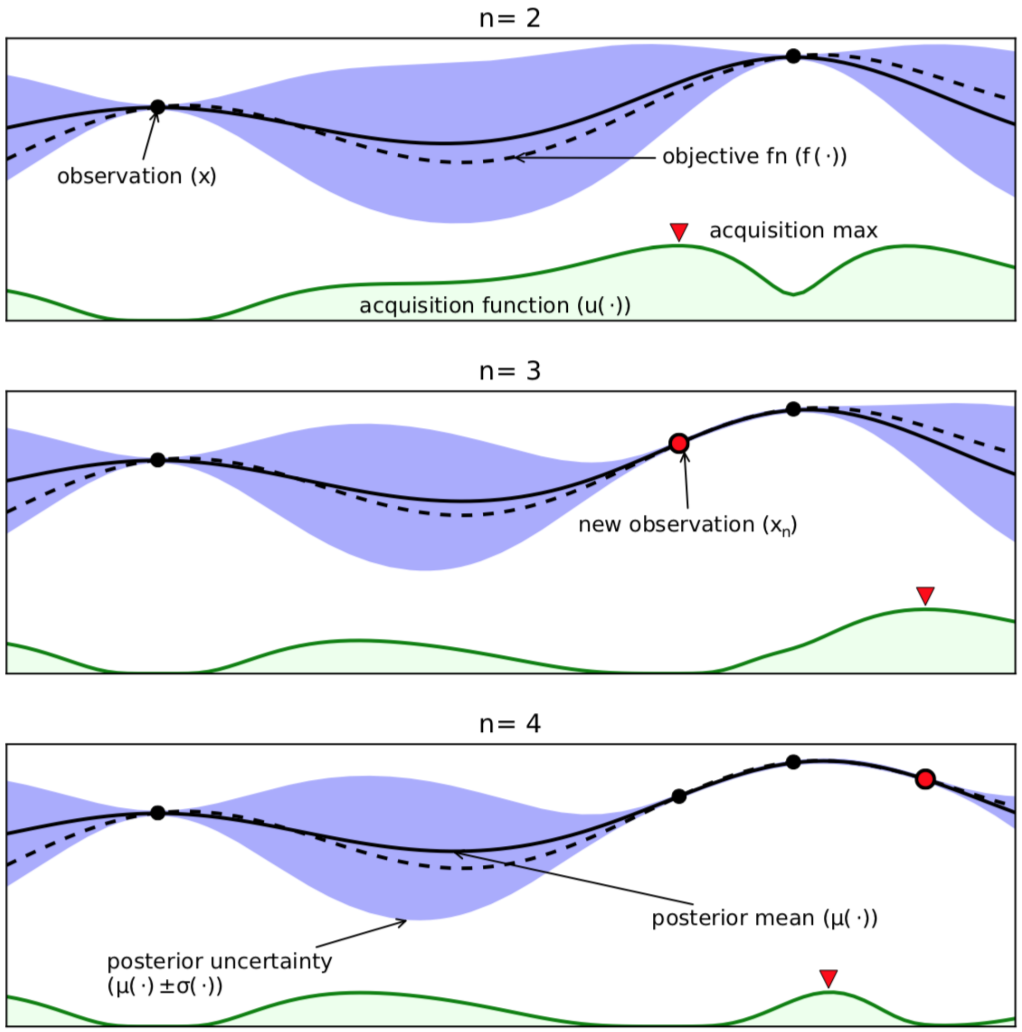
$$Z = \begin{cases} \frac{\mu(\mathbf{x}) - f(\mathbf{x}^+) - \xi}{\sigma(\mathbf{x})} & \text{if } \sigma(\mathbf{x}) > 0 \\ 0 & \text{if } \sigma(\mathbf{x}) = 0 \end{cases}$$

where $\mu(\mathbf{x})$ and $\sigma(\mathbf{x})$ are the mean and the standard deviation of the GP posterior predictive at $\mathbf{x}$, respectively. Φ and ϕ are the CDF and PDF of the standard normal distribution, respectively. The first summation term in Equation (2) is the exploitation term and second summation term is the exploration term.

Parameter ξ in Equation (2) determines the amount of exploration during optimization and higher ξ values lead to more exploration. In other words, with increasing ξ values, the importance of improvements predicted by the GP posterior mean $\mu(\mathbf{x})$ decreases relative to the importance of potential improvements in regions of high prediction uncertainty, represented by large $\sigma(\mathbf{x})$ values. A recommended default value for ξ is 0.01.



The basics: Bayesian Optimization



X, y: (sparse) Training data
X_*: New (not yet evaluated) points

Obtain/update training data *D*

Train surrogate model $X \rightarrow y$

Make prediction $X_* \rightarrow \bar{f}_*, \mathbb{V}[f_*]$

Acquisition function $acq(\bar{f}_*, \mathbb{V}[f_*])$

Evaluate new point $X_*^i = \mathrm{argmax}(acq)$

N. de Freitas et al., Taking the Human Out of the Loop: A Review of Bayesian Optimization , *Proceedings of the IEEE* **104**, 148 (2015)

Bayesian Optimization for physical discovery

Discovering regions where heat capacityis maximized in NNN Ising model