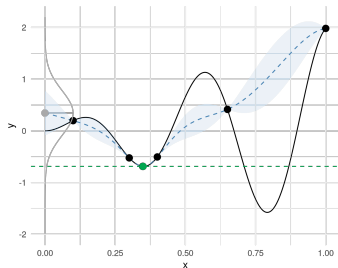
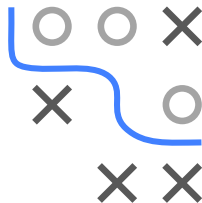


# Optimization in Machine Learning

## Bayesian Optimization Posterior Uncertainty and Acquisition Functions II



### Learning goals

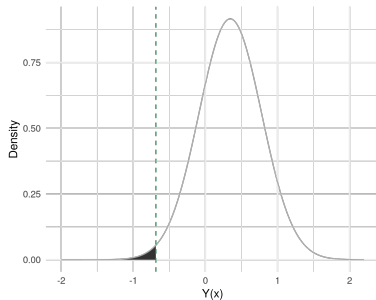
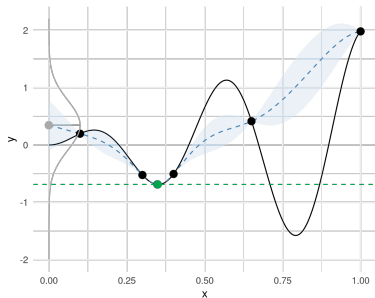
- Probability of improvement
- Expected improvement

# PROBABILITY OF IMPROVEMENT

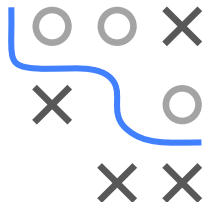
**Goal:** Find  $\mathbf{x}^{[t+1]}$  that maximizes the **Probability of Improvement (PI)**:

$$a_{\text{PI}}(\mathbf{x}) = \mathbb{P}(Y(\mathbf{x}) < f_{\min}) = \Phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right)$$

where  $\Phi(\cdot)$  is the standard normal cdf (assuming Gaussian posterior)



**Left:** The green vertical line represents  $f_{\min}$ . **Right:**  $a_{\text{PI}}(\mathbf{x})$  is given by the black area.



# PROBABILITY OF IMPROVEMENT

**Goal:** Find  $\mathbf{x}^{[t+1]}$  that maximizes the **Probability of Improvement** (PI):

$$a_{\text{PI}}(\mathbf{x}) = \mathbb{P}(Y(\mathbf{x}) < f_{\min}) = \Phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right)$$

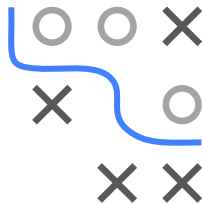
where  $\Phi(\cdot)$  is the standard normal cdf (assuming Gaussian posterior)

**Note:**  $a_{\text{PI}}(\mathbf{x}) = 0$  for design points  $\mathbf{x}$ , since

- $\hat{s}(\mathbf{x}) = 0$ ,
- $\hat{f}(\mathbf{x}) = f(\mathbf{x}) \geq f_{\min} \Leftrightarrow f_{\min} - \hat{f}(\mathbf{x}) \leq 0$ .

Therefore:

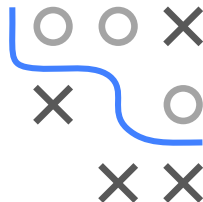
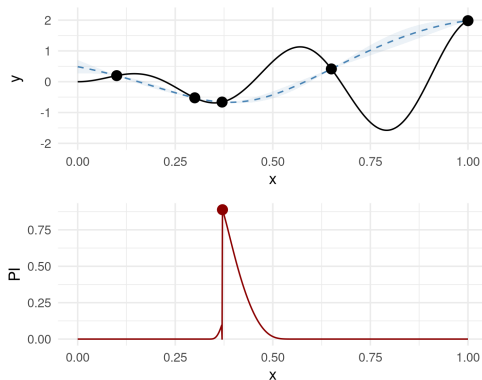
$$\Phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) = \Phi(-\infty) = 0$$



# PROBABILITY OF IMPROVEMENT

The PI does not take the size of the improvement into account  
Often it will propose points close to the current  $f_{\min}$

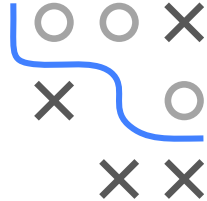
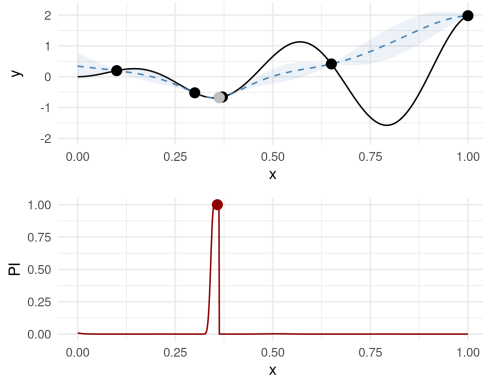
We use the PI (red line) to propose the next point ...



The red point depicts  $\arg \max_{\mathbf{x} \in \mathcal{S}} a_{\text{PI}}(\mathbf{x})$

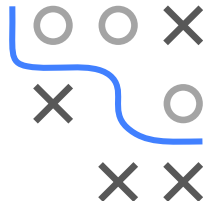
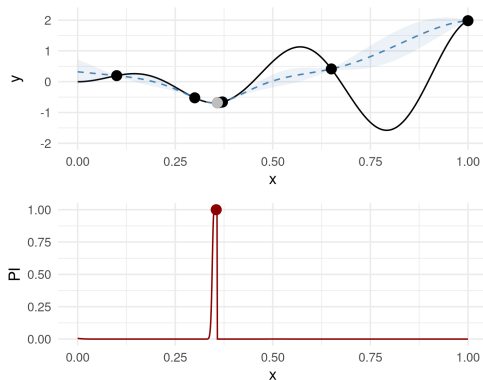
# PROBABILITY OF IMPROVEMENT

...



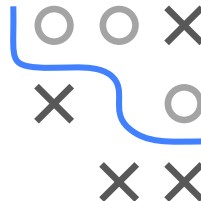
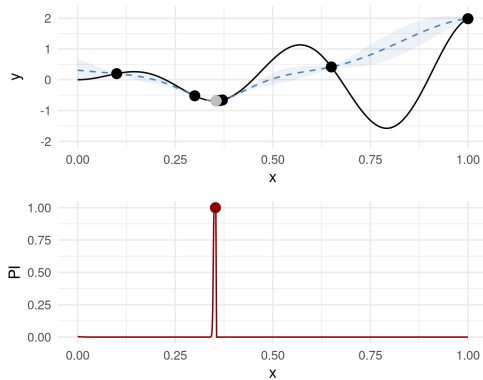
# PROBABILITY OF IMPROVEMENT

In our example, using the PI results in spending plenty of time optimizing the local optimum ...



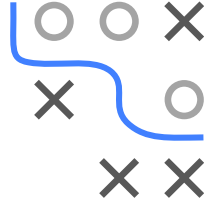
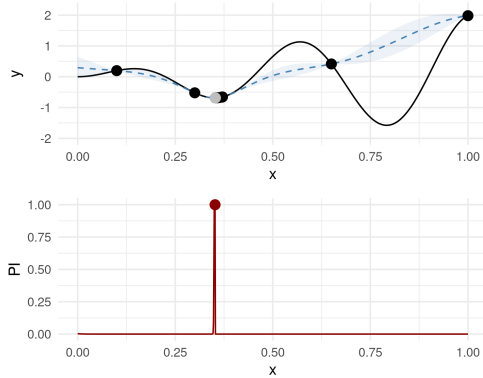
# PROBABILITY OF IMPROVEMENT

...



# PROBABILITY OF IMPROVEMENT

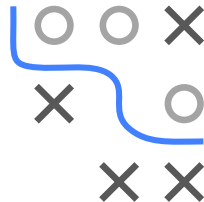
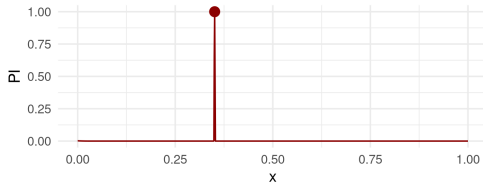
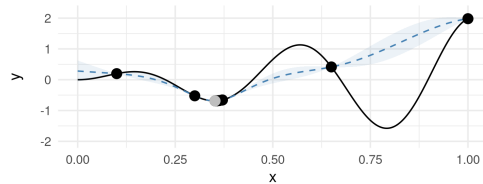
...





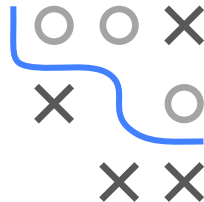
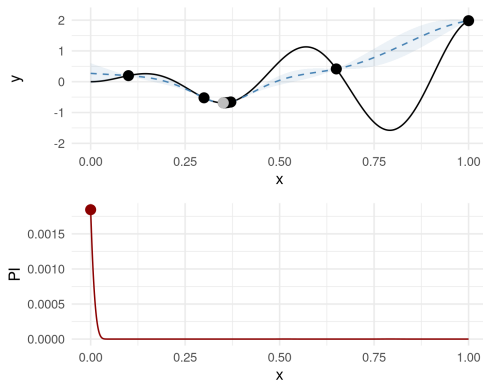
# PROBABILITY OF IMPROVEMENT

...



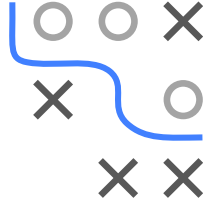
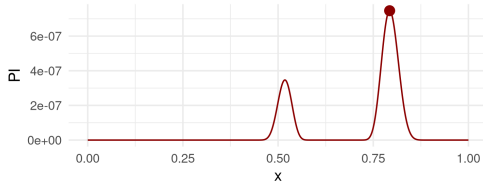
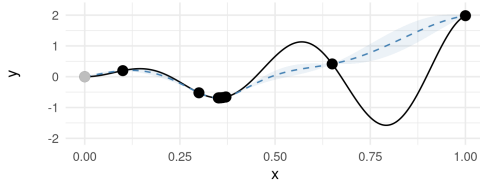
# PROBABILITY OF IMPROVEMENT

... eventually, we explore other regions ...



# PROBABILITY OF IMPROVEMENT

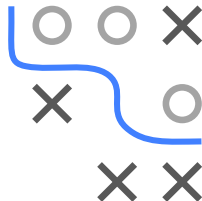
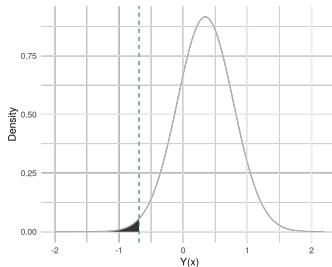
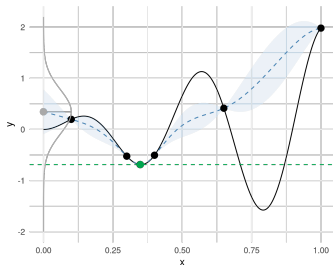
...



# EXPECTED IMPROVEMENT

**Goal:** Propose  $\mathbf{x}^{[t+1]}$  that maximizes the **Expected Improvement (EI)**:

$$a_{\text{EI}}(\mathbf{x}) = \mathbb{E}(\max\{f_{\min} - Y(\mathbf{x}), 0\})$$

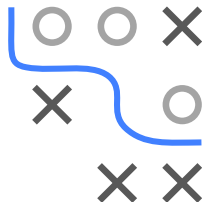
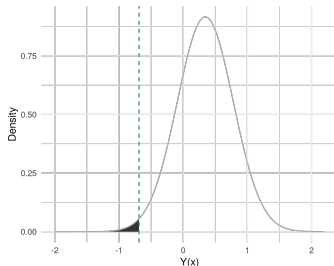
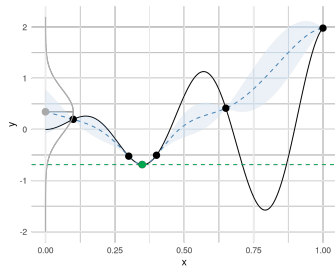


- We now take the expectation in the tail, instead of the prob as in PI.
- Improvement is always assumed  $\geq 0$ .

# EXPECTED IMPROVEMENT

**Goal:** Propose  $\mathbf{x}^{[t+1]}$  that maximizes the **Expected Improvement** (EI):

$$a_{\text{EI}}(\mathbf{x}) = \mathbb{E}(\max\{f_{\min} - Y(\mathbf{x}), 0\})$$



If  $Y(\mathbf{x}) \sim \mathcal{N}(\hat{f}(\mathbf{x}), \hat{s}^2(\mathbf{x}))$ , we can express the EI in closed-form as:

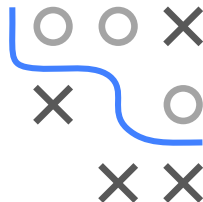
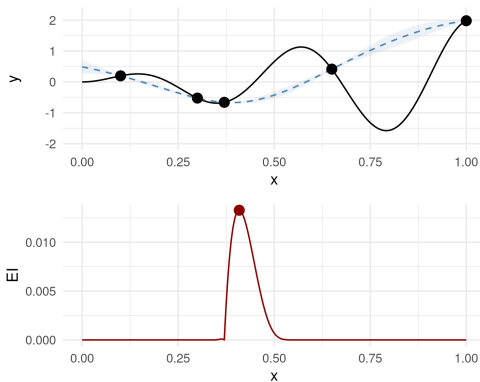
$$a_{\text{EI}}(\mathbf{x}) = (f_{\min} - \hat{f}(\mathbf{x}))\Phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) + \hat{s}(\mathbf{x})\phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right),$$

- $a_{\text{EI}}(\mathbf{x}) = 0$  at design points  $\mathbf{x}$ :

$$a_{\text{EI}}(\mathbf{x}) = \underbrace{(f_{\min} - \hat{f}(\mathbf{x}))\Phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right)}_{=0, \text{ see PI}} + \underbrace{\hat{s}(\mathbf{x})\phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right)}_{=0}$$

# EXPECTED IMPROVEMENT

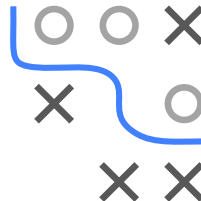
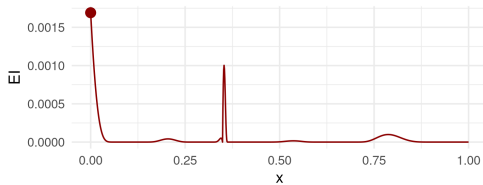
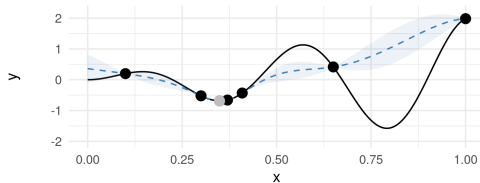
We use the EI (red line) to propose the next point ...



The red point depicts  $\arg \max_{\mathbf{x} \in \mathcal{S}} a_{\text{EI}}(\mathbf{x})$

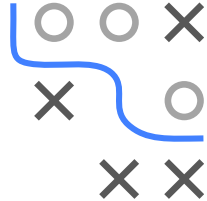
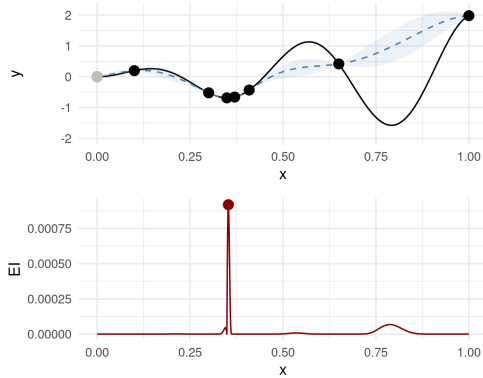
# EXPECTED IMPROVEMENT

...



# EXPECTED IMPROVEMENT

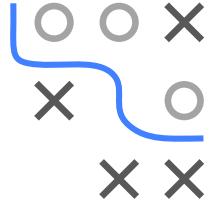
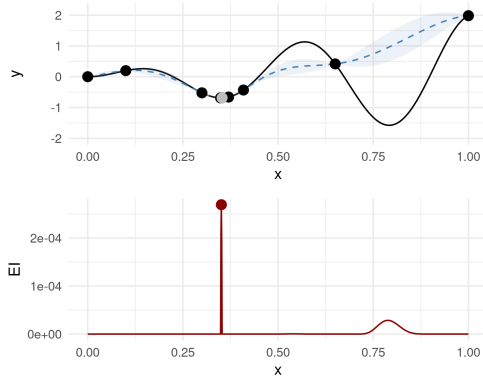
...





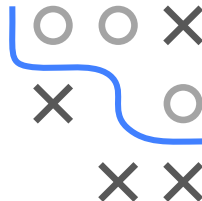
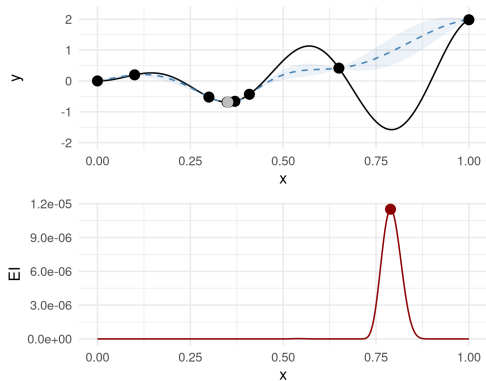
# EXPECTED IMPROVEMENT

...



# EXPECTED IMPROVEMENT

The EI is capable of exploration and quickly proposes promising points in areas we have not visited yet



Here, also a result of well-calibrated uncertainty  $\hat{s}(\mathbf{x})$  of our GP.

# DISCUSSION

- Under some mild conditions: BO with a GP as SM and EI is a **global optimizer**, i.e., convergence to the **global** (!) optimum is guaranteed given unlimited budget (Bull 2011)
- Cannot be proven for the PI or the vanilla LCB
- LCB can be proven to converge in a similar manner if the mean-variance trade-off parameter is chosen adaptively and “correctly” (Srinivas et al. 2010)
- In practice, both LCB and EI work quite well

## Other ACQFs:

- Entropy based: Entropy search, predictive entropy search, max value entropy search
- Knowledge Gradient
- Thompson Sampling
- ...

