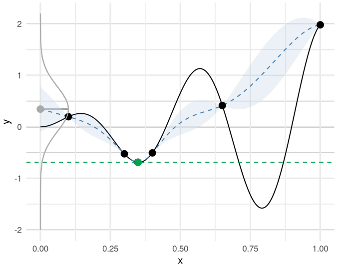
Optimization in Machine Learning

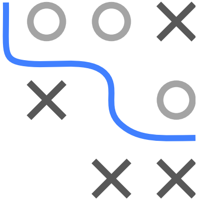
Bayesian Optimization
Posterior Uncertainty and Acquisition Functions II



Learning goals

- Probability of improvement
- Expected improvement

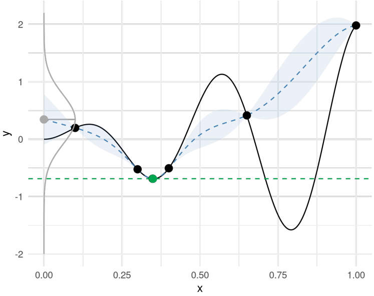
PROBABILITY OF IMPROVEMENT

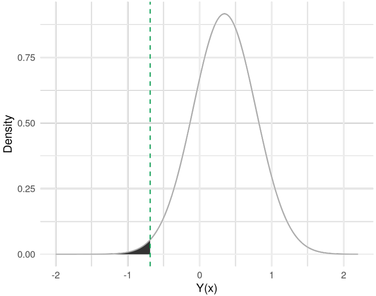


Goal: Find $\mathbf{x}^{[t+1]}$ that maximizes the **Probability of Improvement** (PI):

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

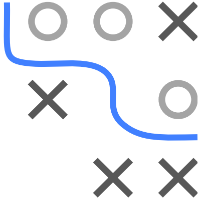
where $\Phi(\cdot)$ is the standard normal cdf (assuming Gaussian posterior)





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

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

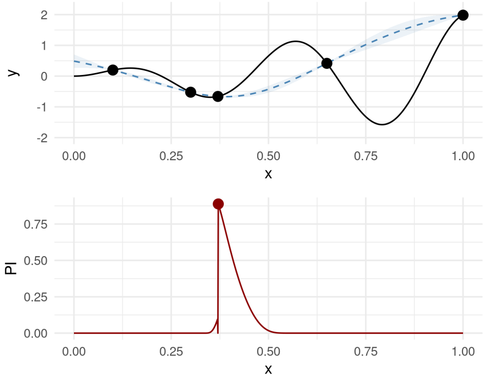
Therefore:

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

PROBABILITY OF IMPROVEMENT

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

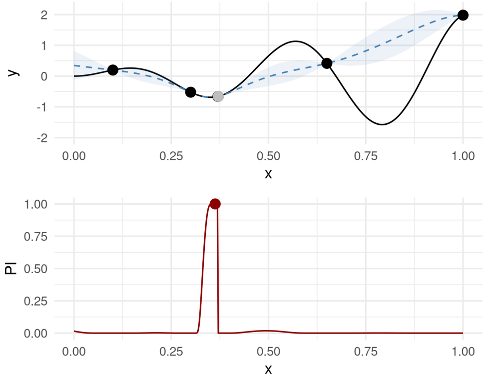
We use the PI (red line) to propose the next point ...



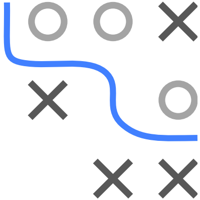
The red point depicts $\arg\max_{\mathbf{x}\in\mathcal{S}} a_{\mathrm{PI}}(\mathbf{x})$

PROBABILITY OF IMPROVEMENT / 2

... evaluate that point, refit the SM and propose the next point

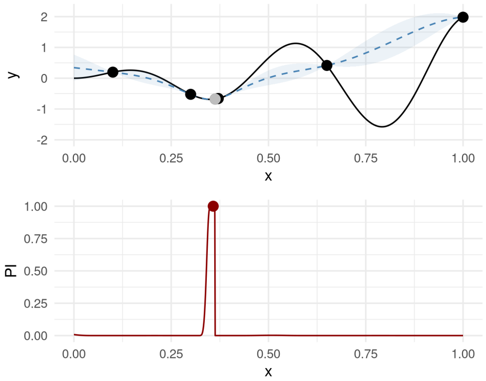


(grey point = prev point from last iter)



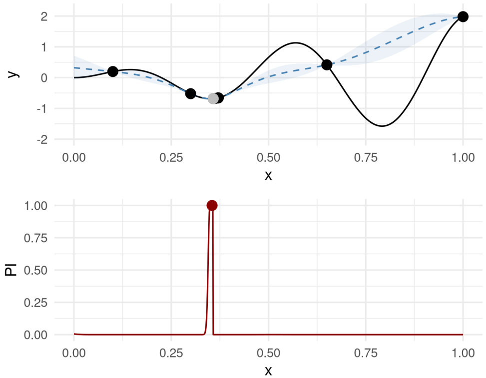
PROBABILITY OF IMPROVEMENT

...



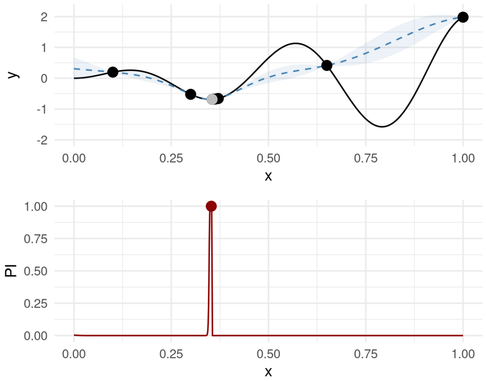
PROBABILITY OF IMPROVEMENT

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



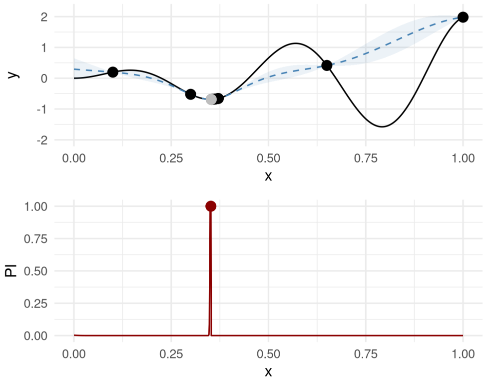
PROBABILITY OF IMPROVEMENT

...



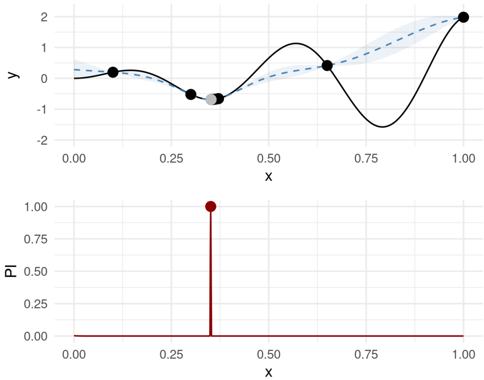
PROBABILITY OF IMPROVEMENT

...



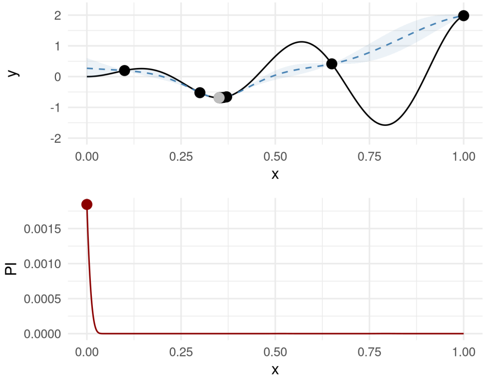
PROBABILITY OF IMPROVEMENT

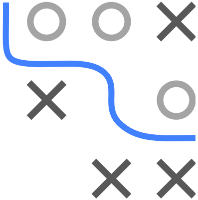
...



PROBABILITY OF IMPROVEMENT

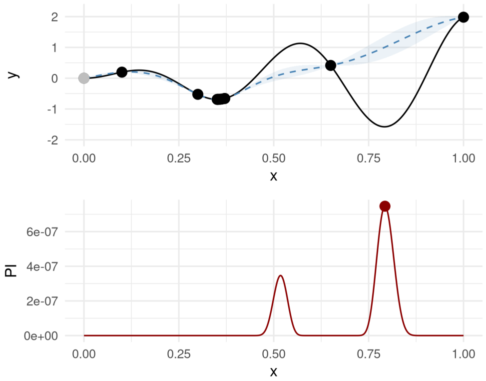
... eventually, we explore other regions ...



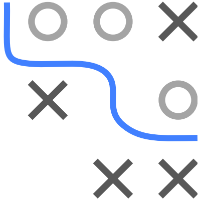


PROBABILITY OF IMPROVEMENT

...

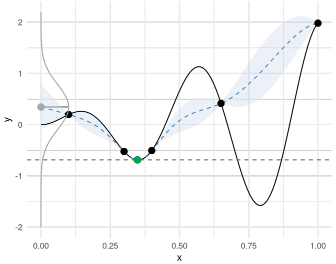


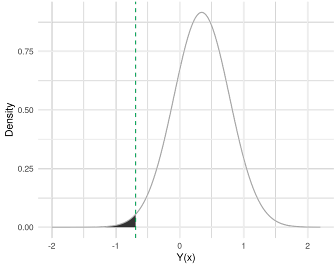
EXPECTED IMPROVEMENT



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

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



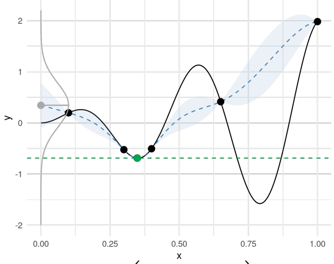


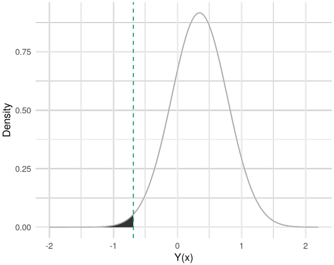
- We now take the expectation in the tail, instead of the prob as in PI.
- Improvement is always assumed ≥ 0.

EXPECTED IMPROVEMENT

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

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





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

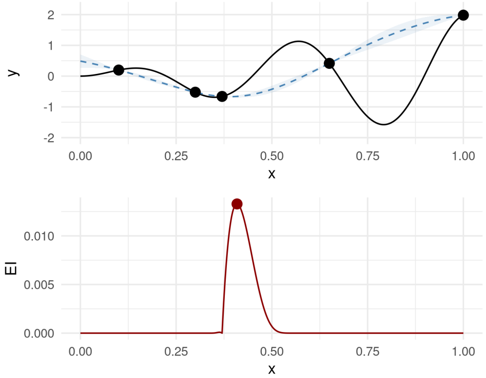
$$a_{\text{EI}}(\mathbf{x}) = (f_{\min} - \hat{f}(\mathbf{x}))\Phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right) + \hat{s}(\mathbf{x})\phi\left(\frac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}\right),$$

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

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

EXPECTED IMPROVEMENT

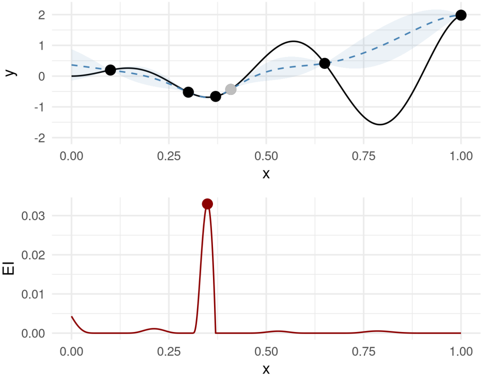
We use the EI (red line) to propose the next point ...



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

EXPECTED IMPROVEMENT / 2

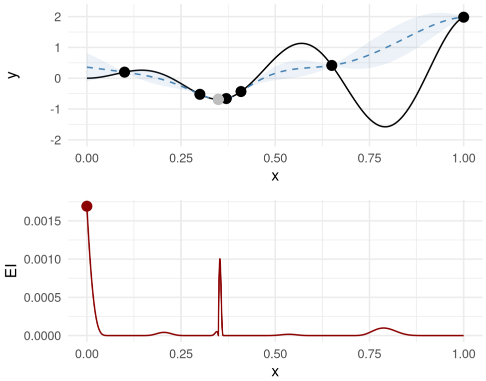
... evaluate that point, refit the SM and propose the next point



(grey point = prev point from last iter)

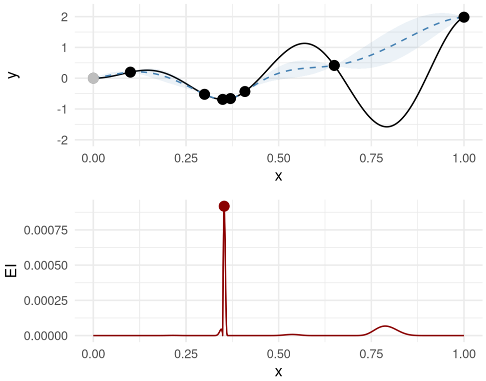
EXPECTED IMPROVEMENT

...



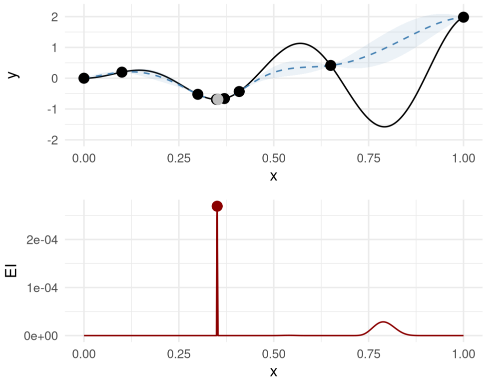
EXPECTED IMPROVEMENT

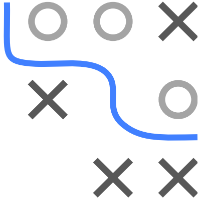
...



EXPECTED IMPROVEMENT

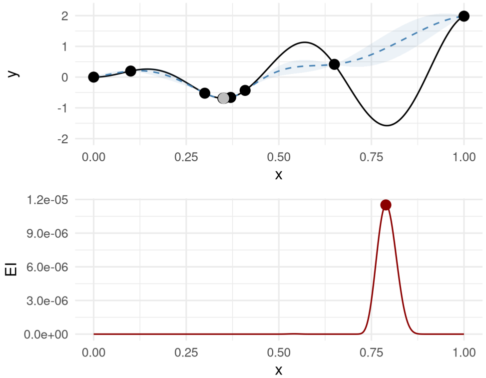
...





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
- Cannot be proven for the PI or the LCB
- In theory, this suggests choosing the EI as ACQF
- In practice, LCB works quite well, and EI generates a very multi-modal landscape

Other ACQFs:

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