Optimization in Machine Learning

Bayesian Optimization Posterior Uncertainty and Acquisition Functions II





Learning goals

- Probability of improvement
- Expected improvement

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

$$a_{\mathsf{Pl}}(\mathbf{x}) = \mathbb{P}(Y(\mathbf{x}) < f_{\min}) = \Phi\left(rac{f_{\min} - \hat{f}(\mathbf{x})}{\hat{s}(\mathbf{x})}
ight)$$

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



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

 $\times \times$

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

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

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

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

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

Therefore:

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

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 $\arg \max_{\mathbf{x} \in S} a_{\mathsf{PI}}(\mathbf{x})$

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





(grey point = prev point from last iter)



× 0 0 × 0 × ×

...

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







× 0 0 × 0 × ×

...





× 0 0 × × ×





× 0 0 × × ×

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



× 0 0 × 0 × ×





× 0 0 × × ×

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





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

Х

×х

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



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



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

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

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



× × ×

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

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





(grey point = prev point from last iter)





× 0 0 × 0 × ×

...









× 0 0 × 0 × ×

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



× × 0 × × ×

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

• ...

× 0 0 × 0 × ×