Bayesian Optimization for Hyperparameter Optimization

Computationally Cheap Acquisition Functions

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

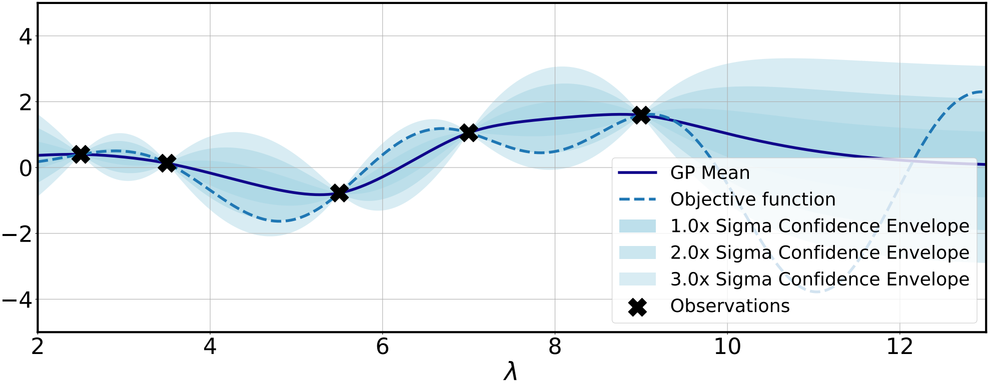
Acquisition Functions: the Basics

- Given the surrogate model $\hat{c}^{(t)}$ at the t-th iteration of BO, the acquisition function $u(\cdot)$ judges the utility (or usefulness) of evaluating f at $\boldsymbol{\lambda}^{(t)} \in \boldsymbol{\Lambda}$ next

Acquisition Functions: the Basics

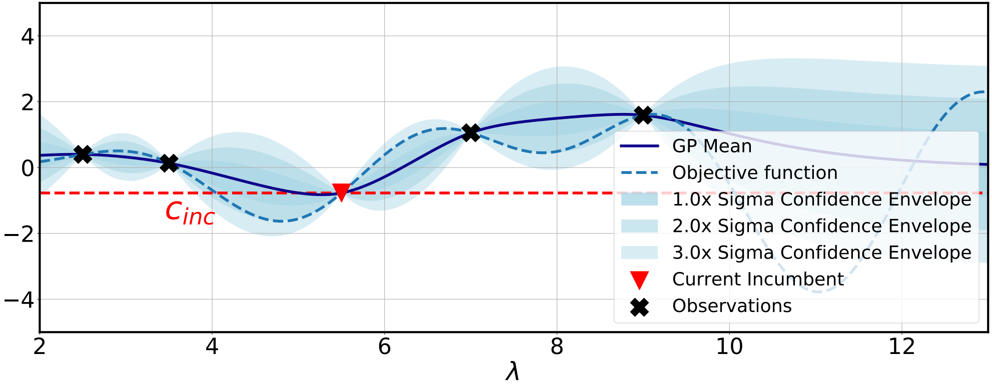
- Given the surrogate model $\hat{c}^{(t)}$ at the t-th iteration of BO, the acquisition function $u(\cdot)$ judges the utility (or usefulness) of evaluating f at $\boldsymbol{\lambda}^{(t)} \in \boldsymbol{\Lambda}$ next
- The acquisition function needs to trade off exploration and exploitation
  - E.g., just picking the $\boldsymbol{\lambda}$ with lowest predicted mean would be too greedy
  - We also need to take into account the uncertainty of the surrogate model $\hat{c}^{(t)}$ to explore

Probability of Improvement (PI): Concept



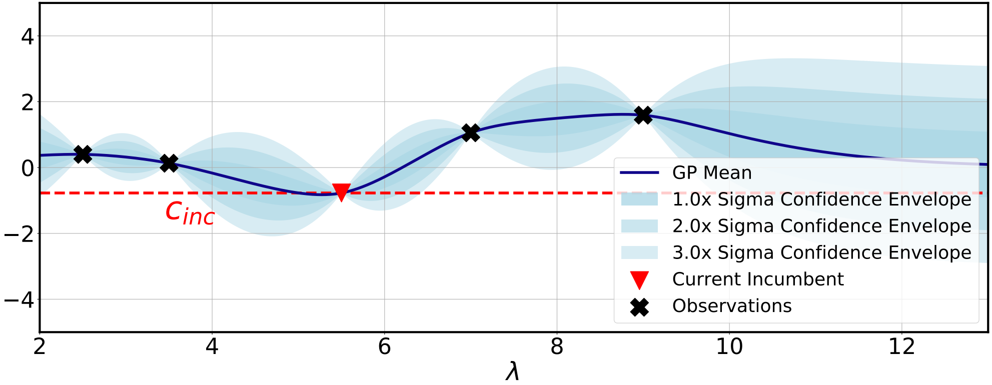
Given the surrogate fit at iteration t

Probability of Improvement (PI): Concept



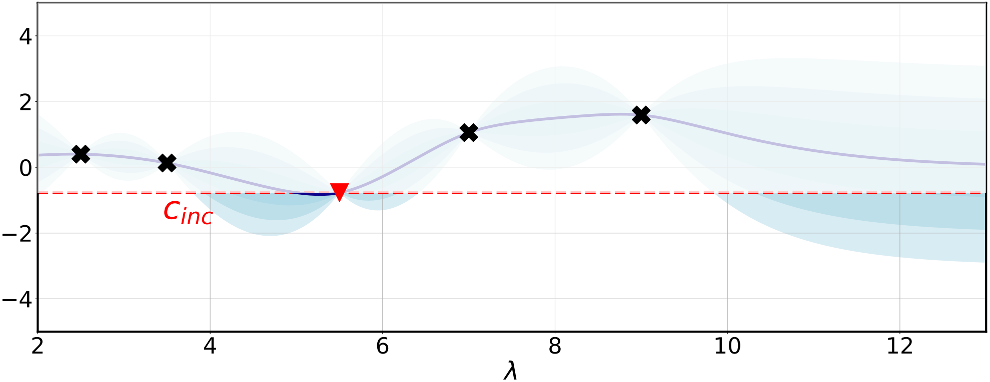
Current incumbent $\hat{\lambda}$ and its observed cost c_{inc}

Probability of Improvement (PI): Concept



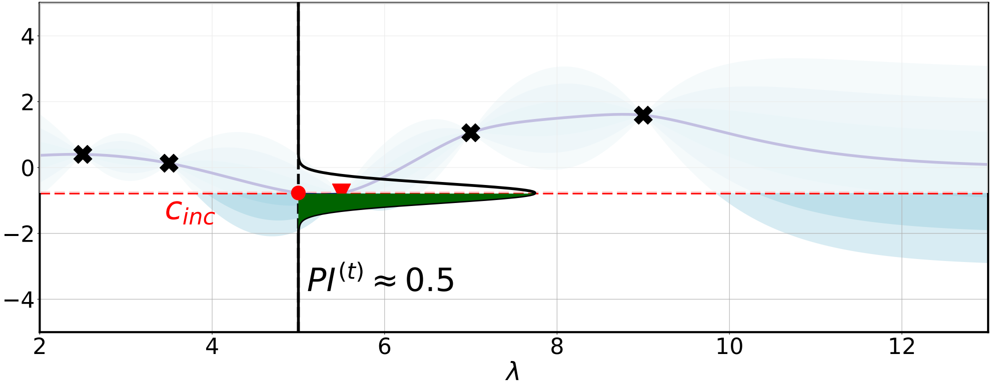
Now let's drop the objective function - it's unknown after all!

Probability of Improvement (PI): Concept



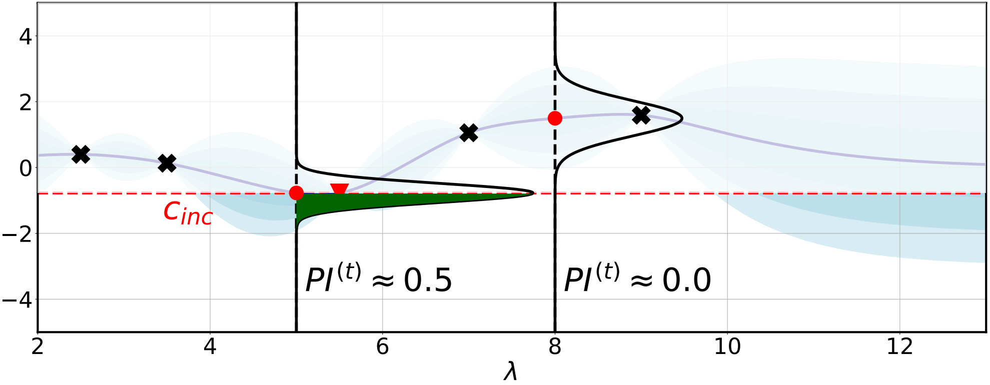
Intuitively, we care about the probability of improving over the current incumbent

Probability of Improvement (PI): Concept



PDF of a good candidate configuration. Only the green area is an improvement.

Probability of Improvement (PI): Concept



PDF of a bad candidate configuration

Probability of Improvement (PI): Formal Definition

- We define the current incumbent at time step t as: $\hat{\boldsymbol{\lambda}}^{(t-1)} \in \arg\min_{\boldsymbol{\lambda}' \in \mathcal{D}^{(t-1)}} c(\boldsymbol{\lambda}')$
- We write c_{inc} shorthand for the cost of the current incumbent: $c_{inc} = c(\hat{\boldsymbol{\lambda}}^{(t-1)})$
- The probability of improvement $u_{PI}(\boldsymbol{\lambda})$ at a configuration $\boldsymbol{\lambda}$ is then defined as:

$$u^{(t)}_{PI}(\boldsymbol{\lambda}) = P(c(\boldsymbol{\lambda}) \leq c_{inc}).$$

Probability of Improvement (PI): Formal Definition

- We define the current incumbent at time step t as: $\hat{\boldsymbol{\lambda}}^{(t-1)} \in \arg\min_{\boldsymbol{\lambda}' \in \mathcal{D}^{(t-1)}} c(\boldsymbol{\lambda}')$
- We write c_{inc} shorthand for the cost of the current incumbent: $c_{inc} = c(\hat{\boldsymbol{\lambda}}^{(t-1)})$
- The probability of improvement $u_{PI}(\boldsymbol{\lambda})$ at a configuration $\boldsymbol{\lambda}$ is then defined as:

$$u_{PI}^{(t)}(\boldsymbol{\lambda}) = P(c(\boldsymbol{\lambda}) \leq c_{inc}).$$

- Since the predictive distribution for $c(\boldsymbol{\lambda})$ is a Gaussian $\mathcal{N}(\mu^{(t-1)}(\boldsymbol{\lambda}), {\sigma^2}^{(t-1)}(\boldsymbol{\lambda}))$, this can be written as:

$$u_{PI}^{(t)}(\boldsymbol{\lambda}) = \Phi[Z], \quad \text{with } Z = \frac{c_{inc} - \mu^{(t-1)}(\boldsymbol{\lambda}) - \xi}{\sigma^{(t-1)}(\boldsymbol{\lambda})},$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and ξ is an optional exploration parameter

Probability of Improvement (PI): Formal Definition

- We define the current incumbent at time step t as: $\hat{\boldsymbol{\lambda}}^{(t-1)} \in \arg\min_{\boldsymbol{\lambda}' \in \mathcal{D}^{(t-1)}} c(\boldsymbol{\lambda}')$
- We write c_{inc} shorthand for the cost of the current incumbent: $c_{inc} = c(\hat{\boldsymbol{\lambda}}^{(t-1)})$
- The probability of improvement $u_{PI}(\boldsymbol{\lambda})$ at a configuration $\boldsymbol{\lambda}$ is then defined as:

$$u_{PI}^{(t)}(\boldsymbol{\lambda}) = P(c(\boldsymbol{\lambda}) \leq c_{inc}).$$

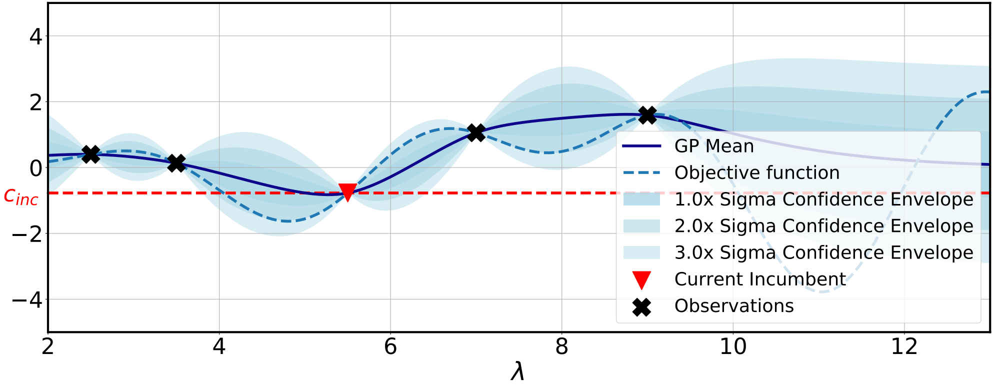
- Since the predictive distribution for $c(\boldsymbol{\lambda})$ is a Gaussian $\mathcal{N}(\mu^{(t-1)}(\boldsymbol{\lambda}), {\sigma^2}^{(t-1)}(\boldsymbol{\lambda}))$, this can be written as:

$$u_{PI}^{(t)}(\boldsymbol{\lambda}) = \Phi[Z], \quad \text{with } Z = \frac{c_{inc} - \mu^{(t-1)}(\boldsymbol{\lambda}) - \xi}{\sigma^{(t-1)}(\boldsymbol{\lambda})},$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and ξ is an optional exploration parameter

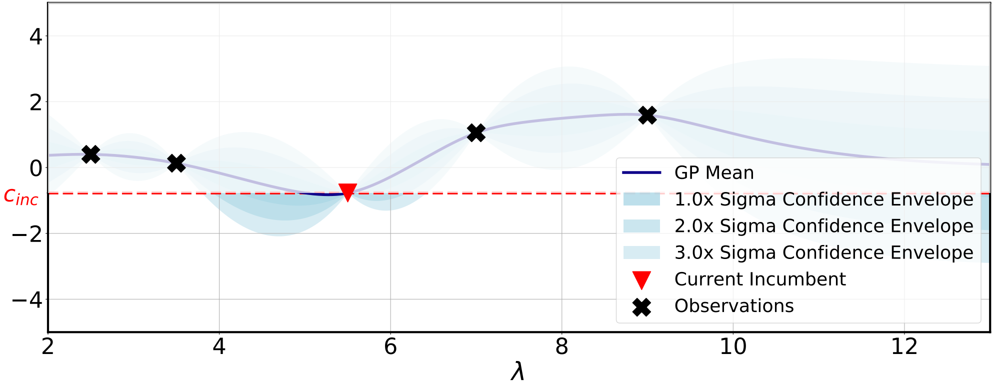
$$\boxed{\text{Choose } \boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}}(u_{PI}^{(t)}(\boldsymbol{\lambda}))}$$

Expected Improvement (EI): Concept



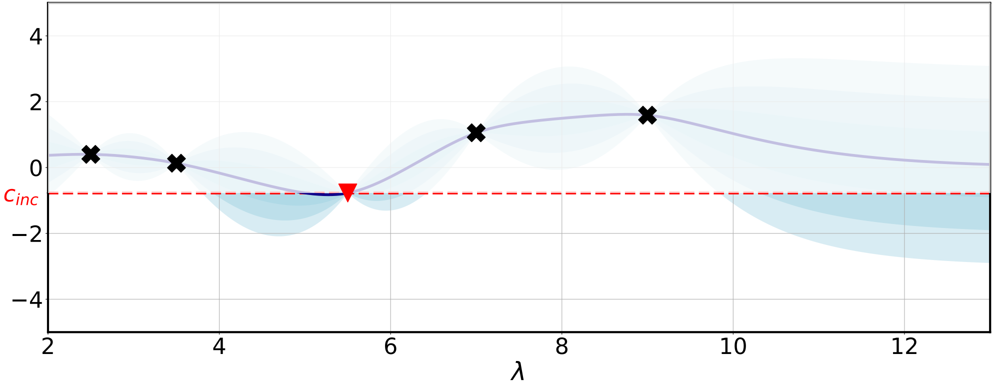
Given the surrogate fit at iteration t

Expected Improvement (EI): Concept



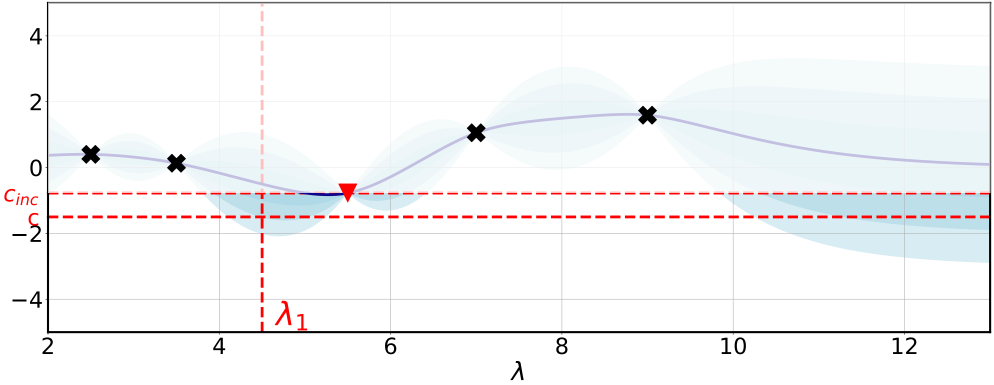
Region of probable improvement – but how large is the improvement?

Expected Improvement (EI): Concept



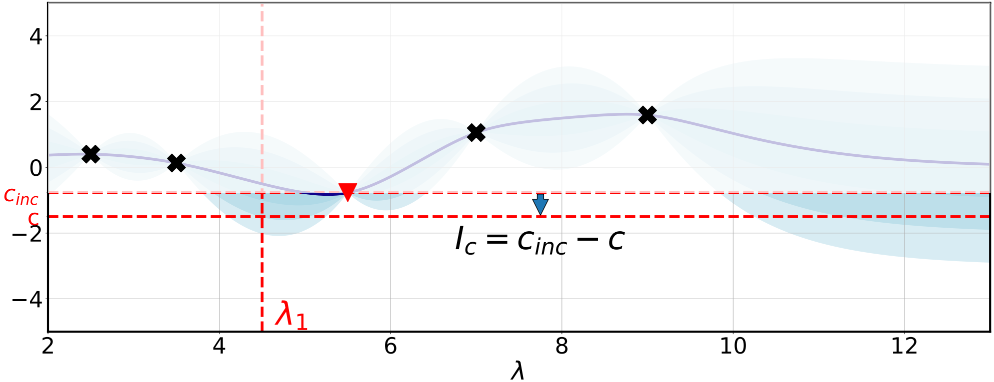
Region of probable improvement – but how large is the improvement?

Expected Improvement (EI): Concept



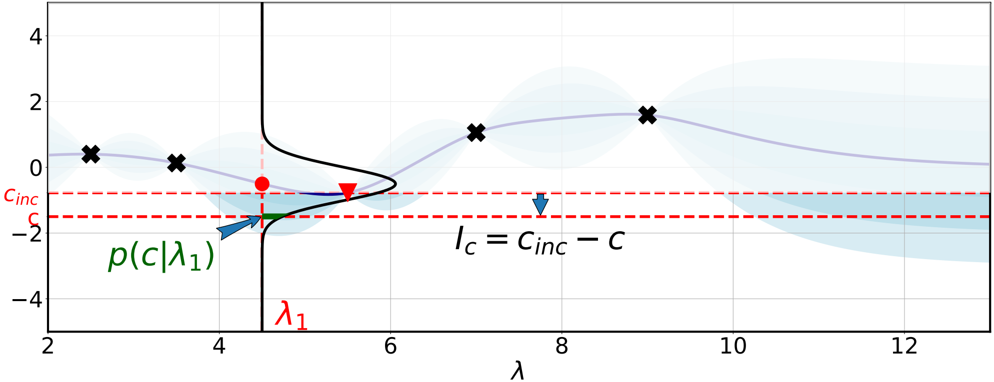
Hypothetical *real* cost c at a given $\boldsymbol{\lambda}$ - unknown in practice without evaluating

Expected Improvement (EI): Concept



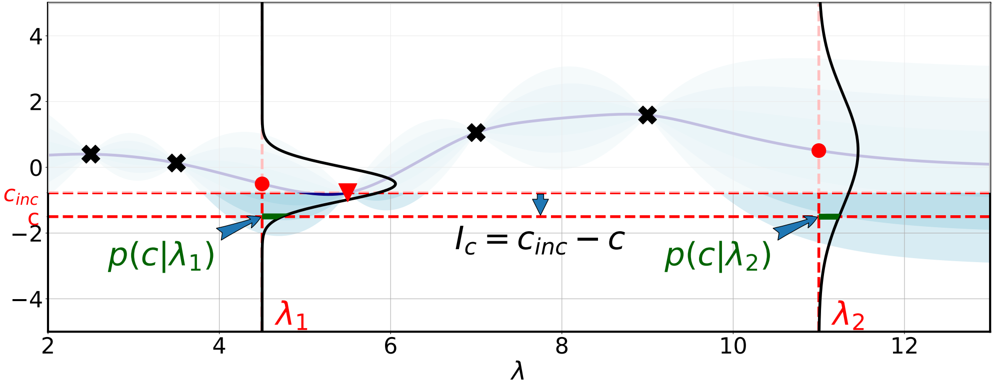
Given a hypothetical c, we can improve the improvement $I_c(\boldsymbol{\lambda})$

Expected Improvement (EI): Concept



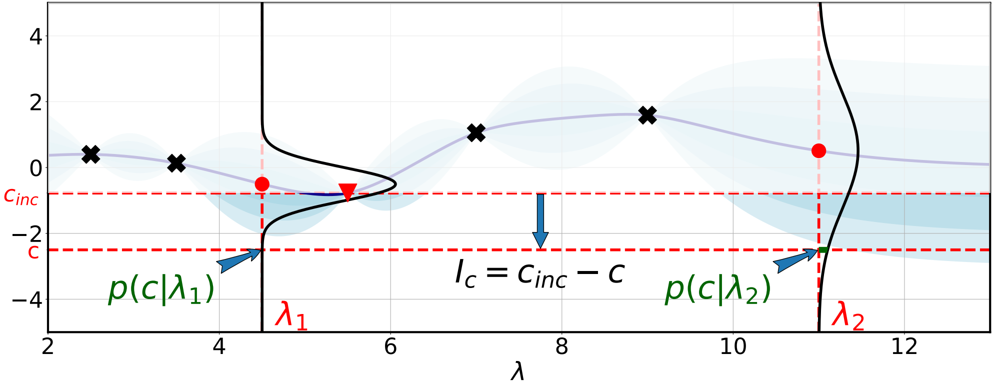
Given $\hat{c}(\boldsymbol{\lambda}) = \mathcal{N}(\mu(\boldsymbol{\lambda}), \sigma^2(\boldsymbol{\lambda}))$, we can also compute $p(c|\boldsymbol{\lambda})$.

Expected Improvement (EI): Concept



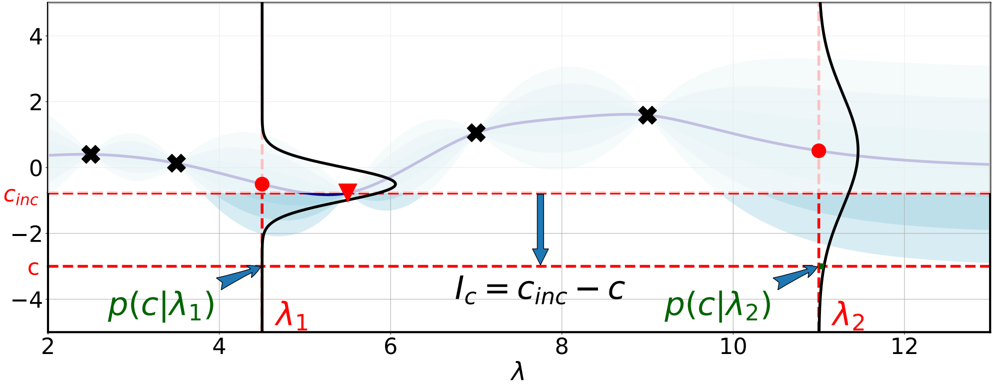
Compare the likelihood of a given improvement for two different configurations $\boldsymbol{\lambda}_1$ and $\boldsymbol{\lambda}_2$

Expected Improvement (EI): Concept



Now consider the likelihood of a larger improvement.

Expected Improvement (EI): Concept



Larger improvements are more likely in areas of high uncertainty.

To compute $\mathbb{E}[I(\boldsymbol{\lambda})]$, intuitively, we sum $p(c \mid \boldsymbol{\lambda}) \times I_c$ over all possibles values of c.

Expected Improvement (EI): Formal Definition

- We define the one-step positive improvement over the current incumbent as
$$I^{(t)}(\boldsymbol{\lambda}) = \max(0, c_{inc} - c(\boldsymbol{\lambda}))$$

- Expected Improvement is then defined as
$$u_{EI}^{(t)}(\boldsymbol{\lambda}) = \mathbb{E}[I^{(t)}(\boldsymbol{\lambda})] = \int_{-\infty}^{\infty} p^{(t)}(c \mid \boldsymbol{\lambda}) \times I^{(t)}(\boldsymbol{\lambda}) \; dc.$$

Expected Improvement (EI): Formal Definition

- We define the one-step positive improvement over the current incumbent as
$$I^{(t)}(\boldsymbol{\lambda}) = \max(0, c_{inc} - c(\boldsymbol{\lambda}))$$

- Expected Improvement is then defined as
$$u_{EI}^{(t)}(\boldsymbol{\lambda}) = \mathbb{E}[I^{(t)}(\boldsymbol{\lambda})] = \int_{-\infty}^{\infty} p^{(t)}(c \mid \boldsymbol{\lambda}) \times I^{(t)}(\boldsymbol{\lambda}) \; dc.$$

- Since the posterior distribution of $\hat{c}(\boldsymbol{\lambda})$ is a Gaussian, EI can be computed in closed form (see exercise):

$$u_{EI}^{(t)}(\boldsymbol{\lambda}) = \begin{cases} \sigma^{(t)}(\boldsymbol{\lambda})[Z\Phi(Z) + \phi(Z)], & \text{if } \sigma^{(t)}(\boldsymbol{\lambda}) > 0 \\ 0 & \text{if } \sigma^{(t)}(\boldsymbol{\lambda}) = 0, \end{cases}$$

where $Z = \dfrac{c_{inc} - \mu^{(t)}(\boldsymbol{\lambda}) - \xi}{\sigma^{(t)}(\boldsymbol{\lambda})}$ and ξ is an optional exploration parameter.

Expected Improvement (EI): Formal Definition

- We define the one-step positive improvement over the current incumbent as
$$I^{(t)}(\boldsymbol{\lambda}) = \max(0, c_{inc} - c(\boldsymbol{\lambda}))$$

- Expected Improvement is then defined as
$$u_{EI}^{(t)}(\boldsymbol{\lambda}) = \mathbb{E}[I^{(t)}(\boldsymbol{\lambda})] = \int_{-\infty}^{\infty} p^{(t)}(c \mid \boldsymbol{\lambda}) \times I^{(t)}(\boldsymbol{\lambda}) \; dc.$$

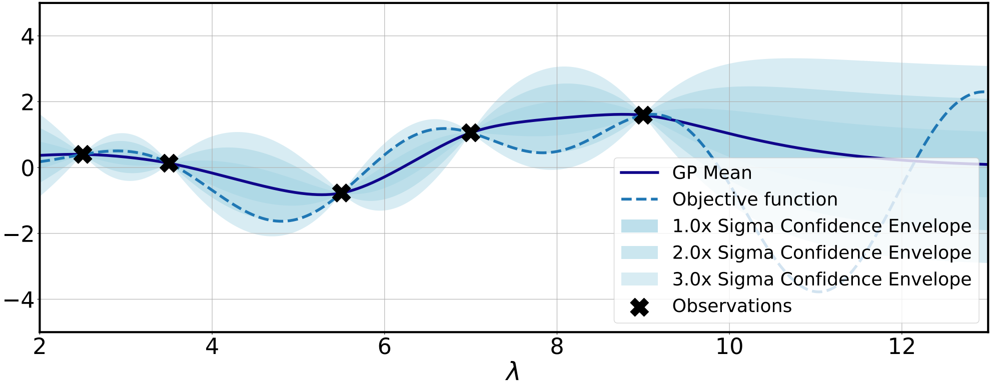
- Since the posterior distribution of $\hat{c}(\boldsymbol{\lambda})$ is a Gaussian, EI can be computed in closed form (see exercise):

$$u_{EI}^{(t)}(\boldsymbol{\lambda}) = \begin{cases} \sigma^{(t)}(\boldsymbol{\lambda})[Z\Phi(Z) + \phi(Z)], & \text{if } \sigma^{(t)}(\boldsymbol{\lambda}) > 0 \\ 0 & \text{if } \sigma^{(t)}(\boldsymbol{\lambda}) = 0, \end{cases}$$

where $Z = \dfrac{c_{inc} - \mu^{(t)}(\boldsymbol{\lambda}) - \xi}{\sigma^{(t)}(\boldsymbol{\lambda})}$ and ξ is an optional exploration parameter.

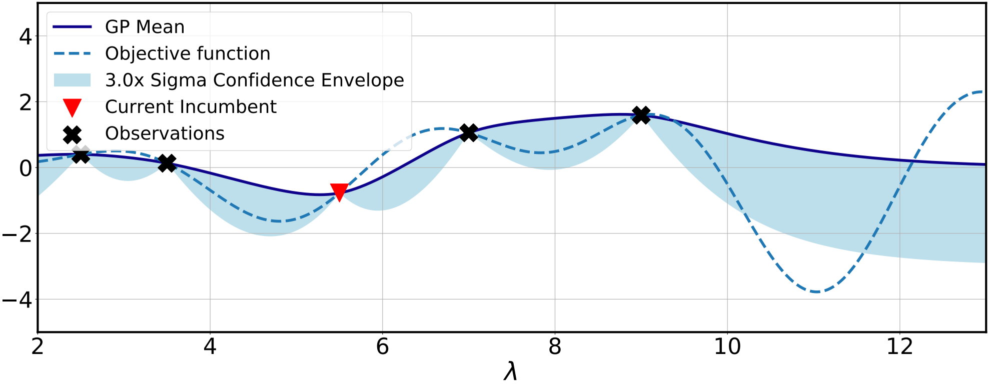
$$\boxed{\text{Choose } \; \boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} (u_{EI}^{(t)}(\boldsymbol{\lambda}))}$$

Lower/Upper Confidence Bounds (LCB/UCB): Concept



Given the surrogate fit at iteration t

Lower/Upper Confidence Bounds (LCB/UCB): Concept



Lower Confidence Bound, $\mu(\boldsymbol{\lambda}) - \alpha\sigma(\boldsymbol{\lambda})$ (here, for $\alpha = 3$)

Lower/Upper Confidence Bounds (LCB/UCB): Formal Definition

- We define the Lower Confidence Bound as

$$u_{LCB}^{(t)}(\boldsymbol{\lambda}) = \mu^{(t)}(\boldsymbol{\lambda}) - \alpha\sigma^{(t)}(\boldsymbol{\lambda}), \quad \alpha \geq 0$$

- One can schedule α (e.g., increase it over time [Srinivas et al. 2009])

$$\boxed{\text{Choose } \boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} \left(-u_{LCB}^{(t)}(\boldsymbol{\lambda})\right)}$$

Lower/Upper Confidence Bounds (LCB/UCB): Formal Definition

- We define the Lower Confidence Bound as

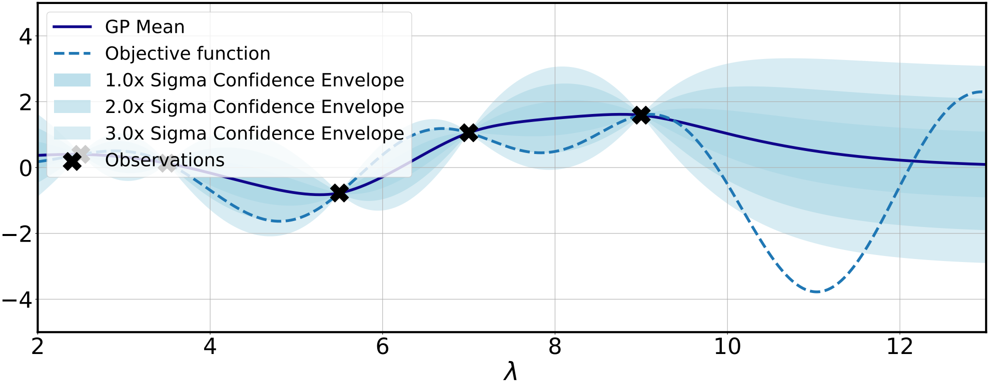
$$u^{(t)}_{LCB}(\boldsymbol{\lambda}) = \mu^{(t)}(\boldsymbol{\lambda}) - \alpha\sigma^{(t)}(\boldsymbol{\lambda}), \quad \alpha \geq 0$$

- One can schedule α (e.g., increase it over time [Srinivas et al. 2009])

$$\boxed{\text{Choose } \boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}} \left(-u^{(t)}_{LCB}(\boldsymbol{\lambda})\right)}$$

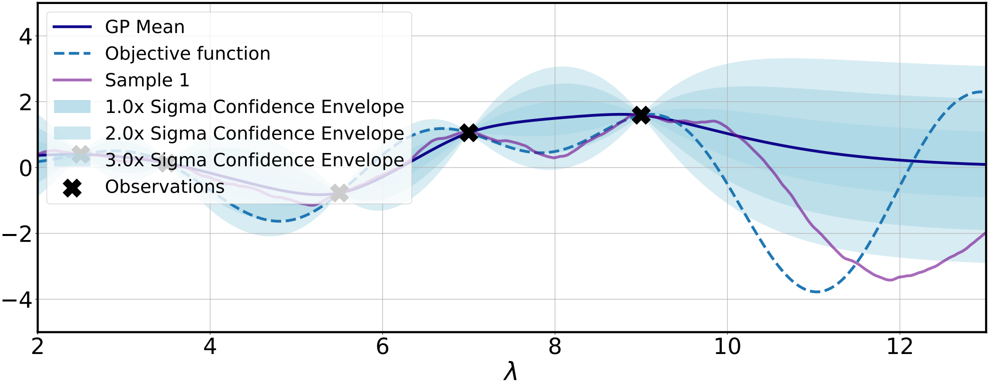
- Note: when one aims to maximize the objective function, one would use UCB instead
  - $u^{(t)}_{UCB}(\boldsymbol{\lambda})) = \mu^{(t)}(\boldsymbol{\lambda}) + \alpha\sigma^{(t)}(\boldsymbol{\lambda})$
  - For UCB, one would choose $\boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda}\in\boldsymbol{\Lambda}}(u^{(t)}_{UCB}(\boldsymbol{\lambda}))$

Thompson Sampling (TS): Concept



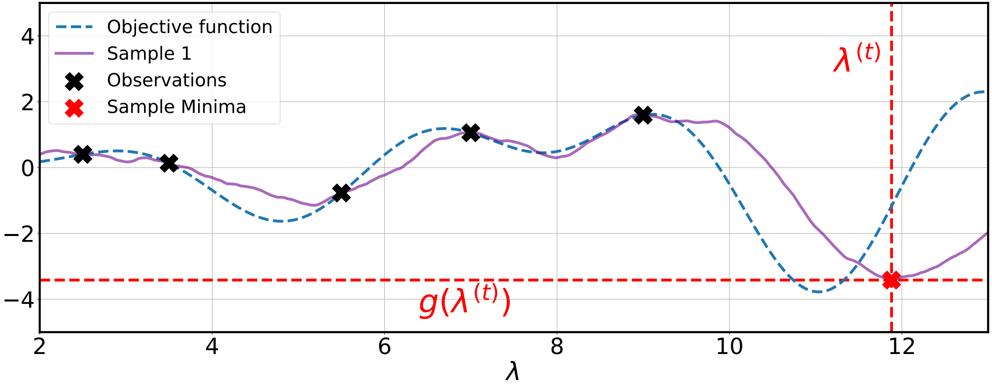
Given the surrogate at iteration t fit on dataset $\mathcal{D}^{(t-1)}$

Thompson Sampling (TS): Concept



Draw a sample g from the predictive surrogate model

Thompson Sampling (TS): Concept



Then choose the minimum of this sample to evaluate at next

Thompson Sampling (TS): Pseudocode

Bayesian Optimization using Thompson Sampling

Require: Search space $\mathbf{\Lambda}$, cost function c, surrogate model $\hat{c}$, maximal number of function evaluations T

Result : Best observed configuration $\hat{\boldsymbol{\lambda}}$ according to $\mathcal{D}^{(T)}$ or $\mathcal{G}$

```
Initialize data D^(0) with initial observations
for t = 1 to T do
    Fit predictive model ĉ^(t) on D^(t−1)
    Sample a function from the surrogate: g ~ ĉ^(t)
    Select next query point: λ^(t) ∈ arg min_{λ∈Λ} g(λ)
    Query c(λ^(t))
    Update data: D^(t) ← D^(t−1) ∪ {⟨λ^(t), c(λ^(t))⟩}
```

Initialize data $\mathcal{D}^{(0)}$ with initial observations

for $t = 1$ **to** T **do**

- Fit predictive model $\hat{c}^{(t)}$ on $\mathcal{D}^{(t-1)}$
- Sample a function from the surrogate: $g \sim \hat{c}^{(t)}$
- Select next query point: $\boldsymbol{\lambda}^{(t)} \in \arg\min_{\boldsymbol{\lambda} \in \mathbf{\Lambda}} g(\boldsymbol{\lambda})$
- Query $c(\boldsymbol{\lambda}^{(t)})$
- Update data: $\mathcal{D}^{(t)} \leftarrow \mathcal{D}^{(t-1)} \cup \{\langle \boldsymbol{\lambda}^{(t)}, c(\boldsymbol{\lambda}^{(t)}) \rangle\}$

Questions to Answer for Yourself / Discuss with Friends

- Discussion. How would you set the exploration parameter ξ for PI if you want to avoid too incremental improvements?
- Derivation. Derive the closed form solution of expected improvement.
- Discussion. In which situations would EI perform substantially differently than PI?