Bayesian Optimization for Hyperparameter Optimization

Introduction to Bayesian Optimization

Bernd Bischl Frank Hutter Lars Kotthoff
Marius Lindauer Joaquin Vanschoren

Blackbox Optimization for Hyperparameter Optimization

- Consider the global optimization problem of finding:

$$\boldsymbol{\lambda}^* \in \underset{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}}{\arg\min}\, f(\boldsymbol{\lambda})$$

- In the most general form, function f is a blackbox function:

$$\boldsymbol{\lambda} \longrightarrow \blacksquare \longrightarrow f(\boldsymbol{\lambda})$$

  - Only mode of interaction with f: querying f's value at a given $\boldsymbol{\lambda}$
  - Function f may not be available in closed form, not differentiable, noisy, etc.

Blackbox Optimization for Hyperparameter Optimization

- Consider the global optimization problem of finding:

$$\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} f(\boldsymbol{\lambda})$$

- In the most general form, function f is a blackbox function:

$$\boldsymbol{\lambda} \longrightarrow \blacksquare \longrightarrow f(\boldsymbol{\lambda})$$

  - Only mode of interaction with f: querying f's value at a given $\boldsymbol{\lambda}$
  - Function f may not be available in closed form, not differentiable, noisy, etc.

- Today, we'll discuss a Bayesian approach for solving such blackbox optimization problems

Blackbox Optimization for Hyperparameter Optimization

- Consider the global optimization problem of finding:

$$\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} f(\boldsymbol{\lambda})$$

- In the most general form, function f is a blackbox function:

$$\boldsymbol{\lambda} \longrightarrow \blacksquare \longrightarrow f(\boldsymbol{\lambda})$$

  - Only mode of interaction with f: querying f's value at a given $\boldsymbol{\lambda}$
  - Function f may not be available in closed form, not differentiable, noisy, etc.

- Today, we'll discuss a Bayesian approach for solving such blackbox optimization problems

- Blackbox optimization can be used for hyperparameter optimization (HPO)
  - Define $f(\boldsymbol{\lambda}) := \mathcal{L}(\mathcal{A}_{\boldsymbol{\lambda}}, \mathcal{D}_{train}, \mathcal{D}_{valid})$

Blackbox Optimization for Hyperparameter Optimization

- Consider the global optimization problem of finding:

$$\boldsymbol{\lambda}^* \in \arg\min_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} f(\boldsymbol{\lambda})$$

- In the most general form, function f is a blackbox function:

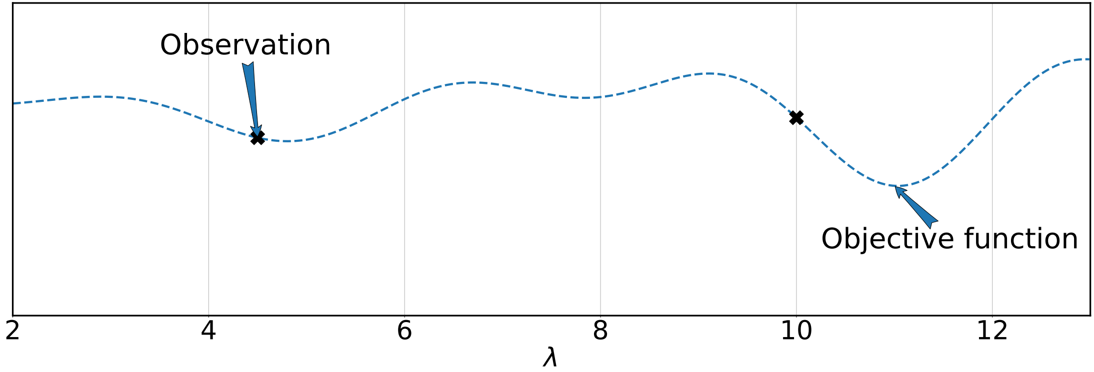
$$\boldsymbol{\lambda} \longrightarrow \blacksquare \longrightarrow f(\boldsymbol{\lambda})$$

  - Only mode of interaction with f: querying f's value at a given $\boldsymbol{\lambda}$
  - Function f may not be available in closed form, not differentiable, noisy, etc.

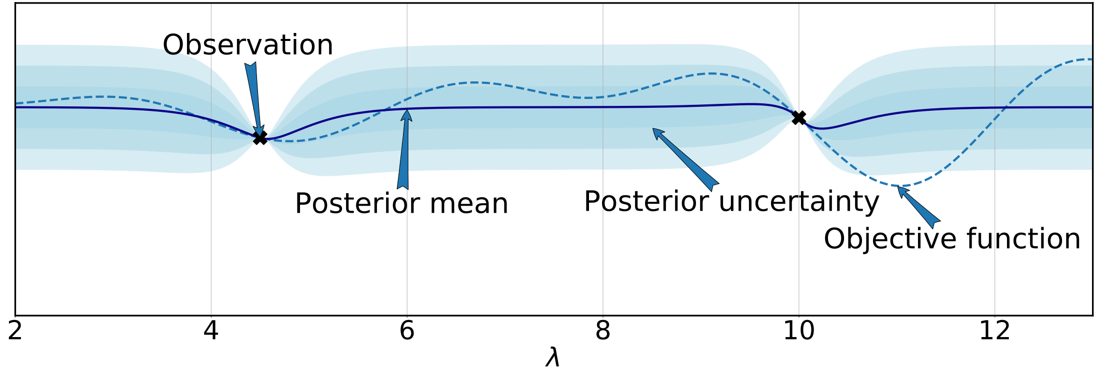
- Today, we'll discuss a Bayesian approach for solving such blackbox optimization problems

- Blackbox optimization can be used for hyperparameter optimization (HPO)
  - Define $f(\boldsymbol{\lambda}) := \mathcal{L}(\mathcal{A}_{\boldsymbol{\lambda}}, \mathcal{D}_{train}, \mathcal{D}_{valid})$
  - Note: for formulations of HPO that go beyond blackbox optimization, see next lecture

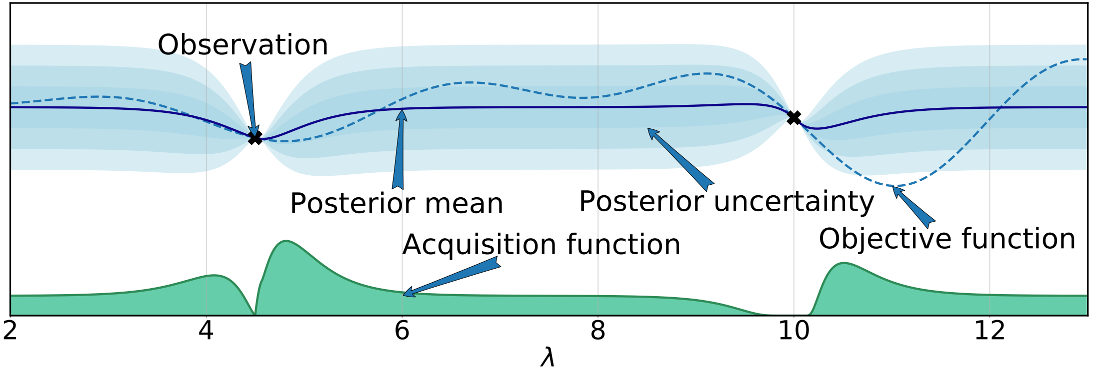
Bayesian Optimization of a blackbox function in a nutshell



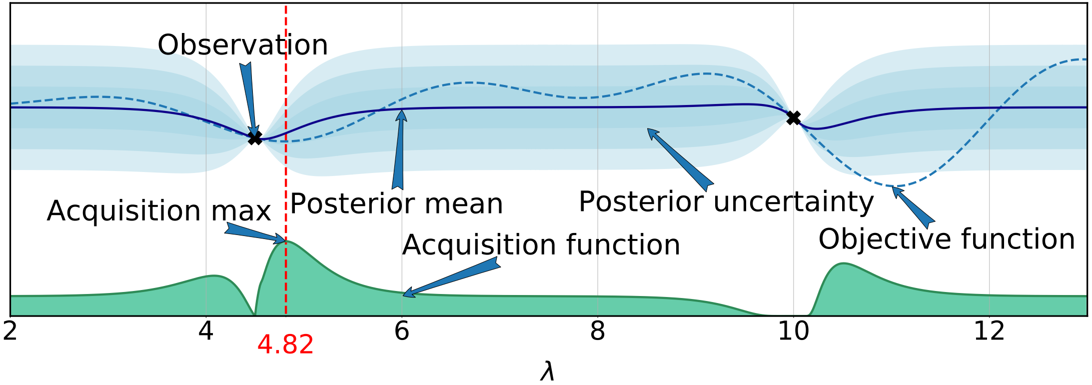
Bayesian Optimization of a blackbox function in a nutshell



Bayesian Optimization of a blackbox function in a nutshell



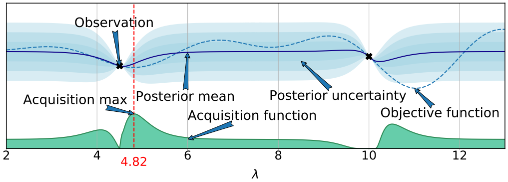
Bayesian Optimization of a blackbox function in a nutshell



Bayesian Optimization of a blackbox function in a nutshell

General approach

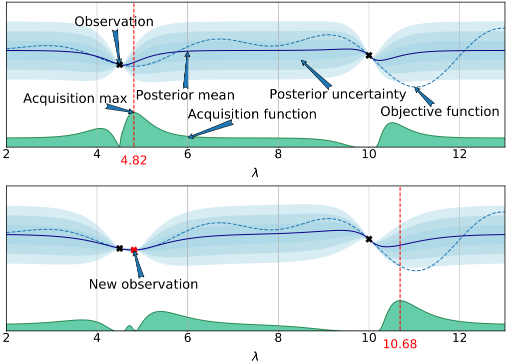
- Fit a probabilistic model to the collected function samples $\langle \boldsymbol{\lambda}, c(\boldsymbol{\lambda}) \rangle$
- Use the model to guide optimization, trading off exploration *vs* exploitation



Bayesian Optimization of a blackbox function in a nutshell

General approach

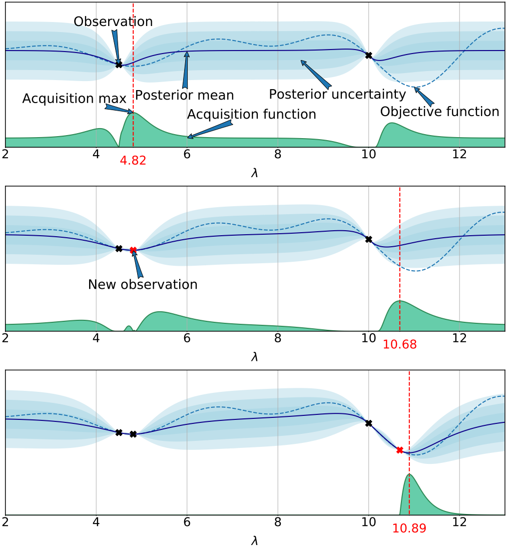
- Fit a probabilistic model to the collected function samples $\langle \boldsymbol{\lambda}, c(\boldsymbol{\lambda}) \rangle$
- Use the model to guide optimization, trading off exploration *vs* exploitation



Bayesian Optimization of a blackbox function in a nutshell

General approach

- Fit a probabilistic model to the collected function samples $\langle \boldsymbol{\lambda}, c(\boldsymbol{\lambda}) \rangle$
- Use the model to guide optimization, trading off exploration *vs* exploitation



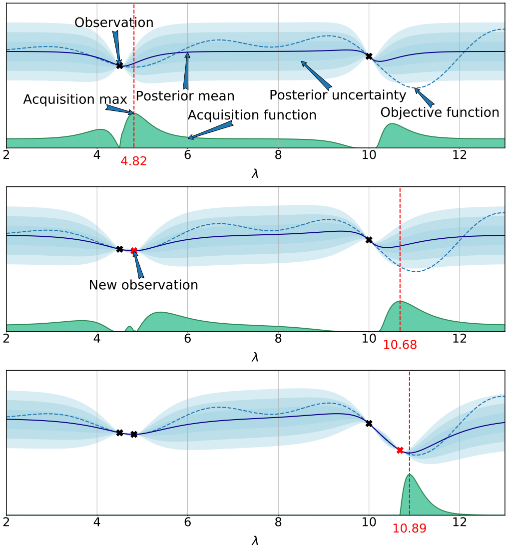
Bayesian Optimization of a blackbox function in a nutshell

General approach

- Fit a probabilistic model to the collected function samples $\langle \boldsymbol{\lambda}, c(\boldsymbol{\lambda}) \rangle$
- Use the model to guide optimization, trading off exploration *vs* exploitation

Popular approach in the statistics literature since Mockus et al. [1978]

- Efficient in #function evaluations
- Works when objective is nonconvex, noisy, has unknown derivatives, etc.
- Recent convergence results [Srinivas et al. 2009; Bull et al. 2011; de Freitas et al. 2012; Kawaguchi et al. 2015]



Bayesian Optimization: Pseudocode

BO loop

Require: Search space $\boldsymbol{\Lambda}$, cost function c, acquisition function u, predictive model $\hat{c}$, maximal number of function evaluations T

Result : Best configuration $\hat{\boldsymbol{\lambda}}$ (according to $\mathcal{D}$ or $\hat{c}$)

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

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

  Fit predictive model $\hat{c}^{(t)}$ on $\mathcal{D}^{(t-1)}$

  Select next query point: $\boldsymbol{\lambda}^{(t)} \in \arg\max_{\boldsymbol{\lambda} \in \boldsymbol{\Lambda}} u(\boldsymbol{\lambda}; \mathcal{D}^{(t-1)}, \hat{c}^{(t)})$

  Query $c(\boldsymbol{\lambda}^{(t)})$

  Update data: $\mathcal{D}^{(t)} \leftarrow \mathcal{D}^{(t-1)} \cup \{\langle \boldsymbol{\lambda}^{(t)}, c(\boldsymbol{\lambda}^{(t)}) \rangle\}$

Bayesian Optimization: Origin of the Name

- Bayesian optimization uses Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \propto P(B|A) \times P(A)$$

- Bayesian optimization uses this to compute a posterior over functions:

$$P(f|\mathcal{D}_{1:t}) \propto P(\mathcal{D}_{1:t}|f) \times P(f), \qquad \text{where } \mathcal{D}_{1:t} = \{\boldsymbol{\lambda}_{1:t}, c(\boldsymbol{\lambda}_{1:t})\}$$

Bayesian Optimization: Origin of the Name

- Bayesian optimization uses Bayes' theorem:

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)} \propto P(B|A) \times P(A)$$

- Bayesian optimization uses this to compute a posterior over functions:

$$P(f|\mathcal{D}_{1:t}) \propto P(\mathcal{D}_{1:t}|f) \times P(f), \qquad \text{where } \mathcal{D}_{1:t} = \{\boldsymbol{\lambda}_{1:t}, c(\boldsymbol{\lambda}_{1:t})\}$$

- Meaning of the individual terms:
  - $P(f)$ is the prior over functions, which represents our belief about the space of possible objective functions before we see any data
  - $\mathcal{D}_{1:t}$ is the data (or observations, evidence)
  - $P(\mathcal{D}_{1:t}|f)$ is the likelihood of the data given a function
  - $P(f|\mathcal{D}_{1:t})$ is the posterior probability over functions given the data

Bayesian Optimization: Advantages and Disadvantages

Advantages

- Sample efficient
- Can handle noise
- Native incorporation of priors
- Does not require gradients
- Theoretical guarantees

Bayesian Optimization: Advantages and Disadvantages

Advantages

- Sample efficient
- Can handle noise
- Native incorporation of priors
- Does not require gradients
- Theoretical guarantees

Disadvantages

- Overhead because of model training in each iteration
- Crucially relies on robust surrogate model
- Inherently sequential (in its basic form)

Learning Goals of this Lecture

After this lecture, students can ...

- Explain the basics of Bayesian optimization
- Derive simple acquisition functions
- Describe advanced acquisition functions
- Describe possible surrogate models and their pros and cons
- Discuss the limits of Bayesian optimization and extensions to tackle these
- Describe the alternative Bayesian optimization approach of TPE
- Discuss success stories of Bayesian optimization