Bayesian Optimisation: Mastering Sequential Decision Making

PhD. Juan Ungredda, ESTECO SpA

Content

- Recap
- Motivation & applications of Bayesian Optimisation.
- Surrogate model.
- Acquisition functions.
- Advanced topics.

Gaussian Process Regression: Recap

$$f(x) \sim GP(m(x), k(x, x'))$$

- Model is fully determined by $m(x)$ and $k(x, x')$
- Posterior can be computed in closed form.
- Provides mean predictions and uncertainty calibration.

Motivation of Bayesian Optimization.

Consider a function $f : \mathcal{X} \rightarrow \mathbb{R}$ in a bounded domain $\mathcal{X} \subseteq \mathbb{R}^D$. We aim to,

$$x^* = \mathrm{argmax}_{x \in \mathcal{X}} f(x)$$

- **Black-box Function**: Lacks a known analytical form. Only input-output pairs $(x, f(x))$ are observable.
- **Expensive**: There's a constraint on the number of function evaluations allowed.
- **Noisy or Uncertain Observations**: Function evaluations might have noise.

Applications

- **Model configuration in machine learning**: find optimal hyper-parameter values, learning rates, number of layers, etc.

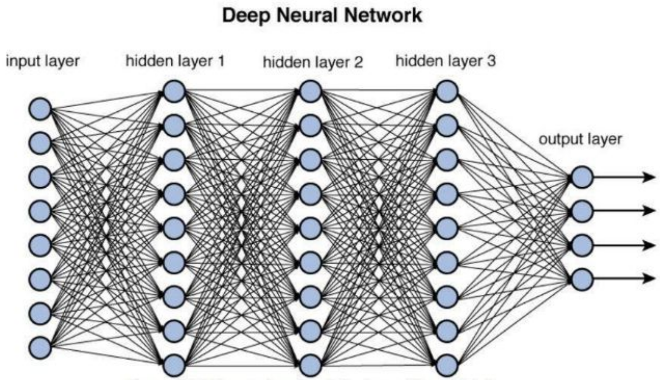
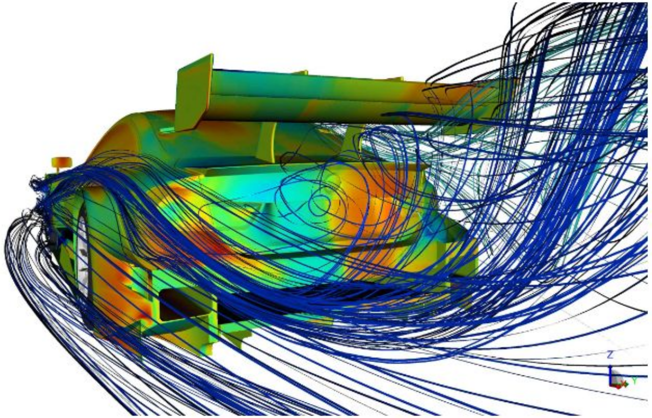


Figure 12.2 Deep network architecture with multiple layers.

Applications

- **Adaptive experimentation**: Optimize a function embodied in a physical process.



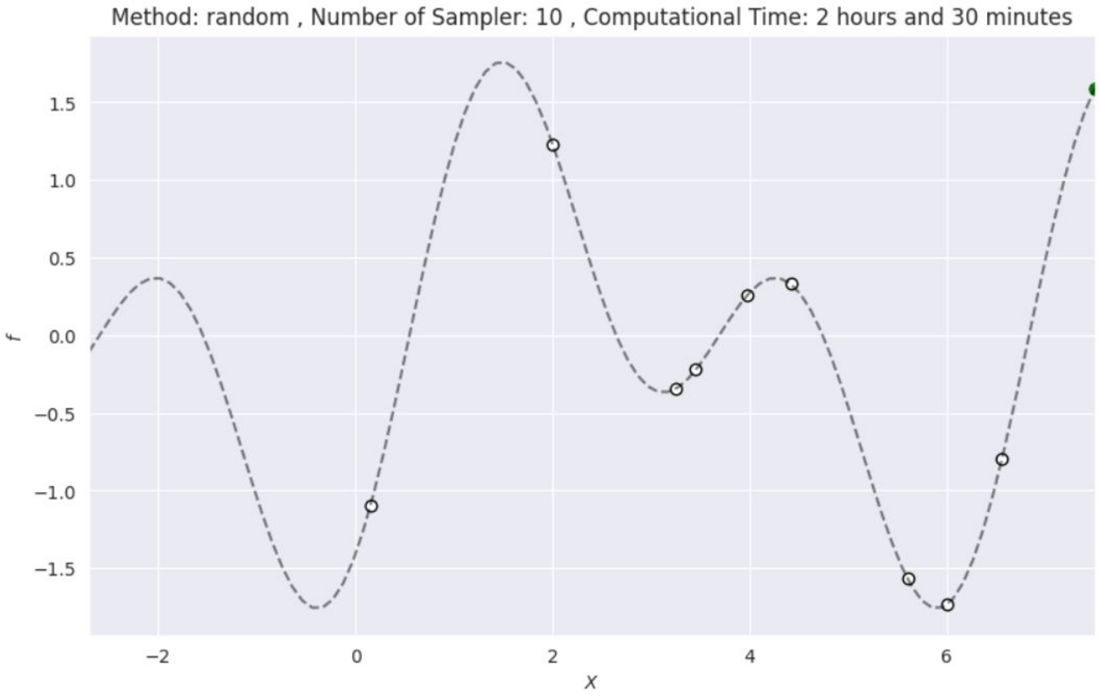
Applications

Many other problems:

- Robotics.
- Control, reinforcement learning.
- A/B testing.
- Scheduling, planning.
- Industrial design.
- Simulation-optimization.

optimiser random

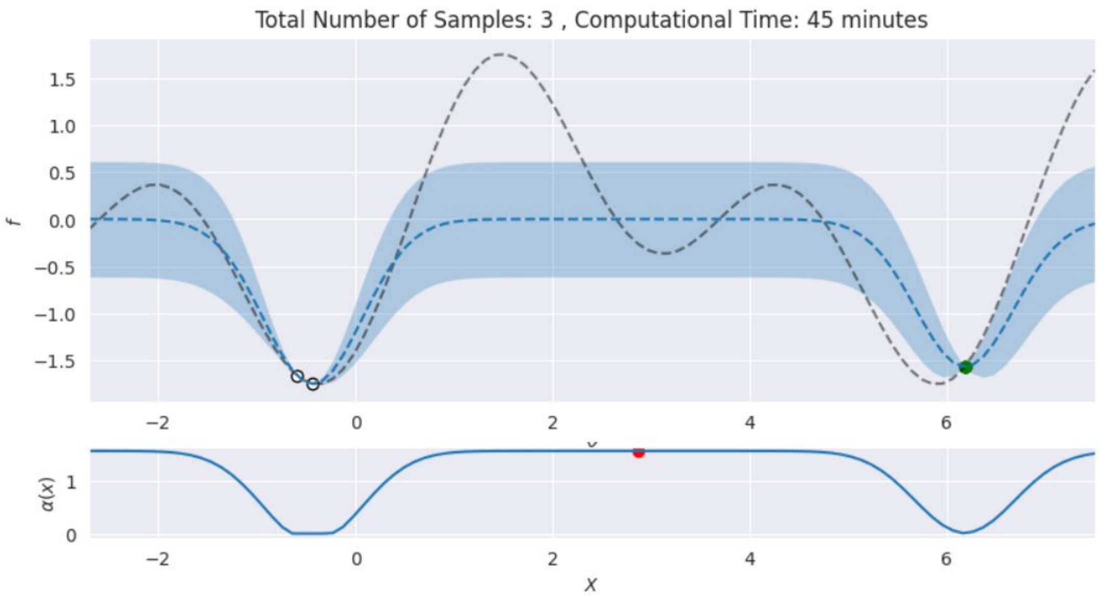
n: 7



Out[5]: Toggle show/hide

Bayesian Optimization Preview

BO samples: 0



Out[7]: Toggle show/hide

Ingredients of Bayesian Optimization

- **Surrogate Model**: Calibrates the prediction and uncertainty over the data.
- **Acquisition function**: Transform the surrogate model and decision maker's utility into a sampling decision.

Surrogate Model

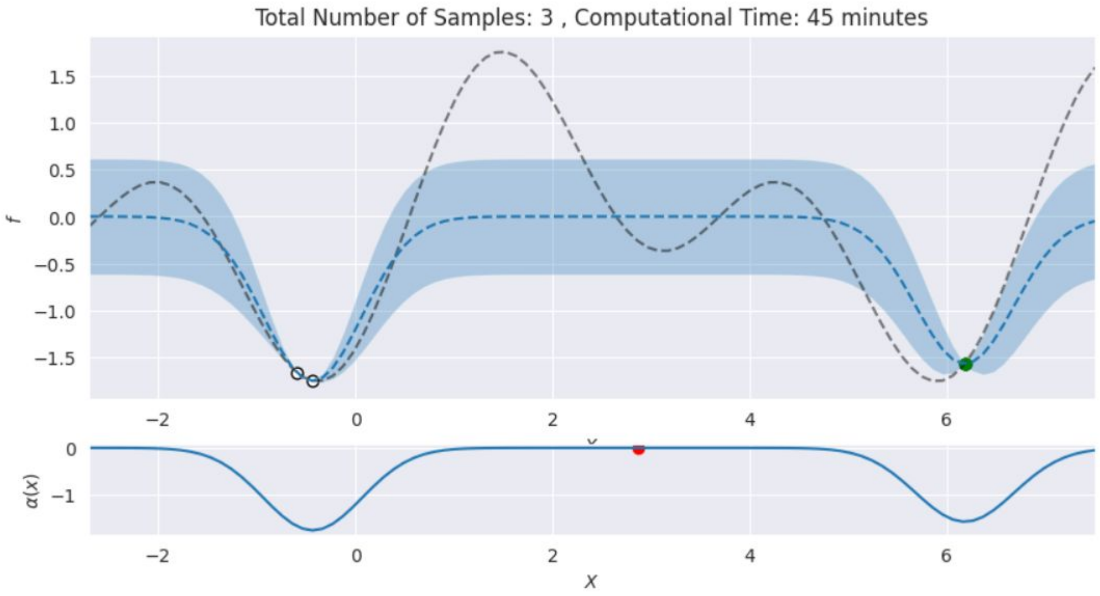
We typically use a Gaussian process (GP) but other models may be considered,

- T-Student processes.
- Random Forests.
- Bayesian neural networks.
- Trees of Parzen estimators.
- etc.

Any model able to calibrate uncertainty (needed for exploration) can be used in Bayesian optimization.

Exploration vs Exploitation

Strategy exploit BO samples: 0



Out[8]: Toggle show/hide

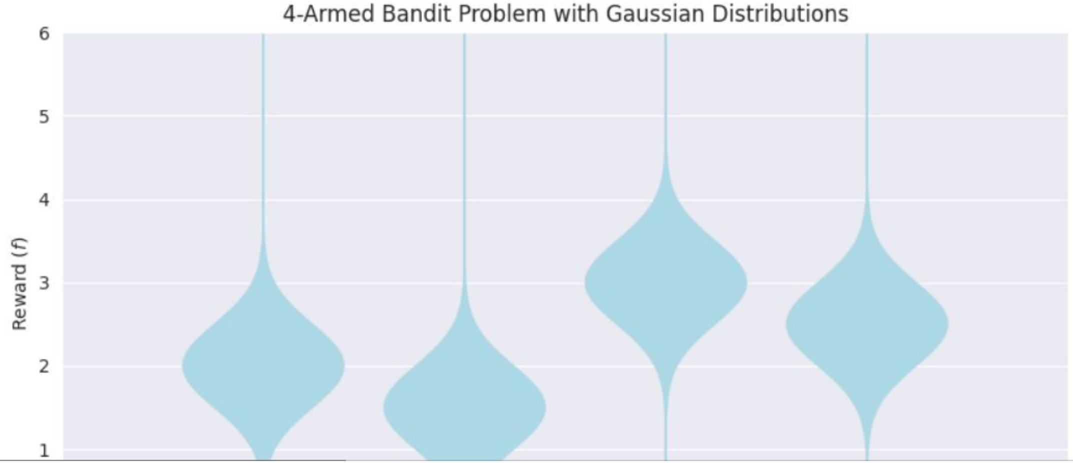
Bandit Problem

reset | Bandit 0 | bandit 1 | Bandit 2 | Bandit 3

Value for actions:

Ranking and Selection: $\mathbb{E}^{\pi}[f(x^B)] \approx 0.0$

Multi-Armed Bandits: $\mathbb{E}^{\pi}[\sum_{n=0}^{N} \lambda^n f(x^n)] \approx 0.0$



Acquisition Functions

$$\alpha(x) : \mathscr{X} \rightarrow \mathbb{R}$$

Aim:

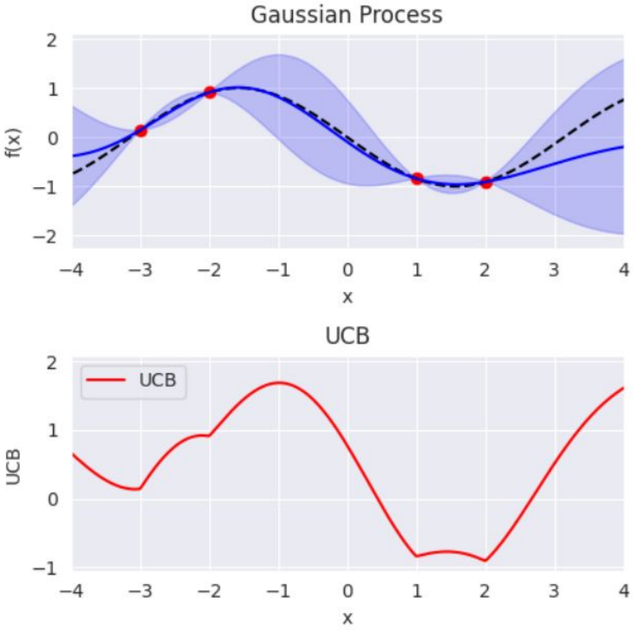
Maps an arbitrary query point x to a measure of quality of the experiment.

Important considerations:

1. **Computationally Efficiency**
2. **Consistency**

GP Upper (lower) Confidence Band

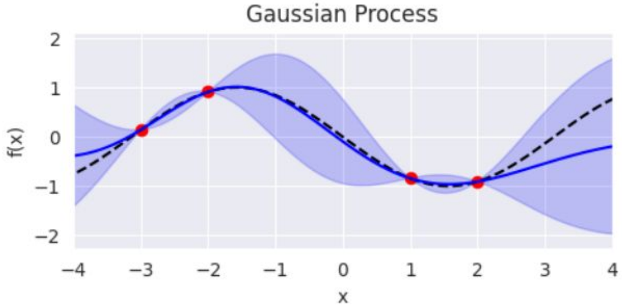
$$\alpha_{UCB}(x) = \mu(\mathbf{x}) + +\beta_n \sigma(\mathbf{x})$$



Expected Improvement

$$\alpha_{EI}(x) = \mathbb{E}_y[max(0, y - y_{best})]$$

$$= \alpha_{EI}(x) = (\mu(x) - f_{\text{best}})\Phi\left(\frac{\mu(x) - f_{\text{best}}}{\sigma(x)}\right) + \sigma(x)\phi\left(\frac{\mu(x) - f_{\text{best}}}{\sigma(x)}\right)$$



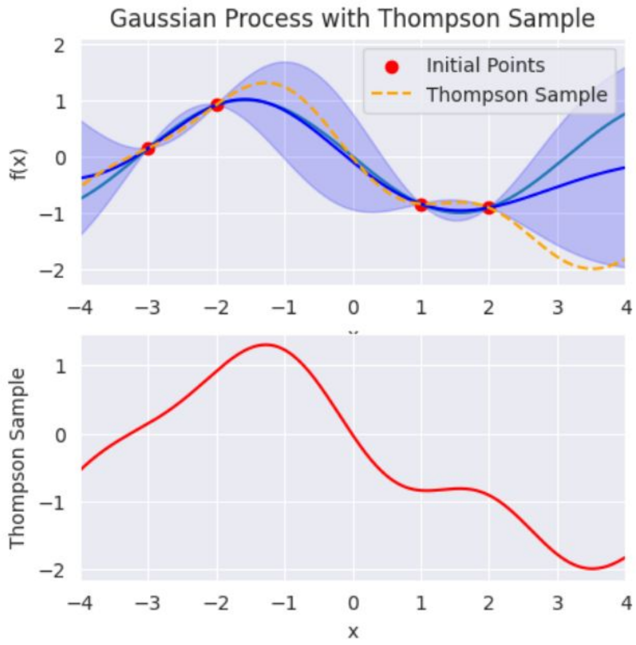
Entropy search and Predictive Entropy search

$$\alpha_{ES} = H[p(x_{max}|\mathscr{D})] - \mathbb{E}_{p(y|\mathscr{D},x)}[H[p(x_{max}|\mathscr{D} \cup \{x, y\})]]$$

- Information theoretic approaches: reduce the entropy of p(xmin).
- Same acquisition, two different approximations (ES, PES).
- Approximating p(x_{min}) is not trivial.

Thompson sampling

$$\alpha_{THOMP}(x) = g(x), \text{ where g(x) is sampled from } GP(\mu(x), k(x, x'))$$



Other acquisition functions

Each acquisition balances exploration-exploitation in a different way. No universal best method.

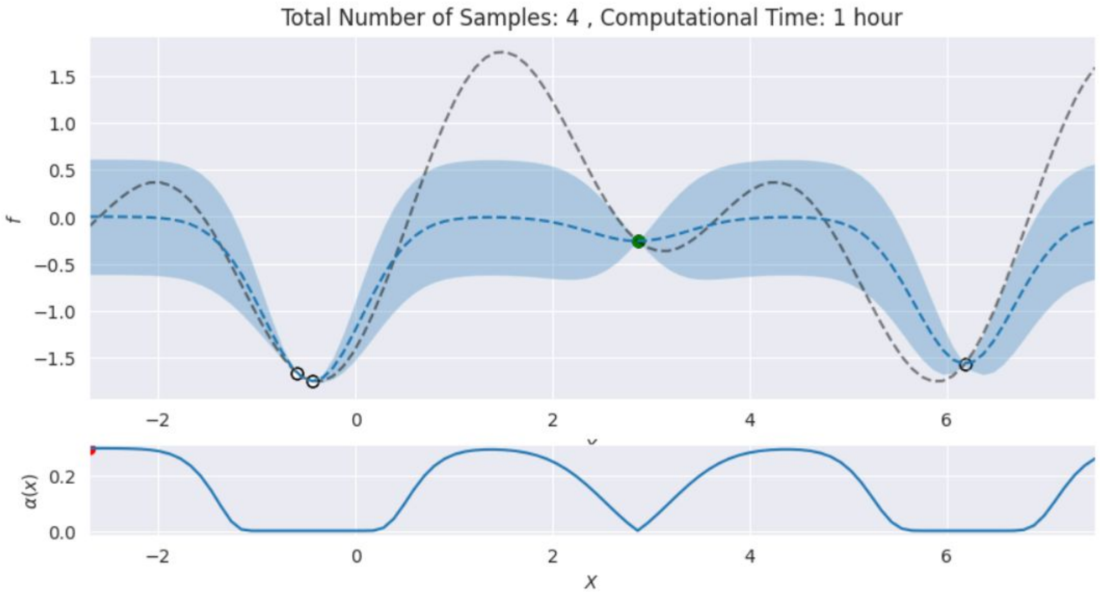
Others:

- Probability of improvement.
- Knowledge gradient.
- etc.

Bayesian Optimization Algorithm

0. Collect initial data and fit a Gaussian process.
1. While the budget is not over:
   - Compute $x_{new} = \text{argmax}_{x \in \mathcal{X}} \alpha(x)$
   - Update daset, $\mathcal{D}^{new} = \mathcal{D}^{old} \cup \{(x, y)_{new}\}$
   - Update Gaussian process to $\mathcal{D}^{new}$.
   - Update budget consumed
2. Recommend best found or estimated solution.

BO samples: 1



Out[10]: Toggle show/hide

Summary of Standard Bayesian Optimization

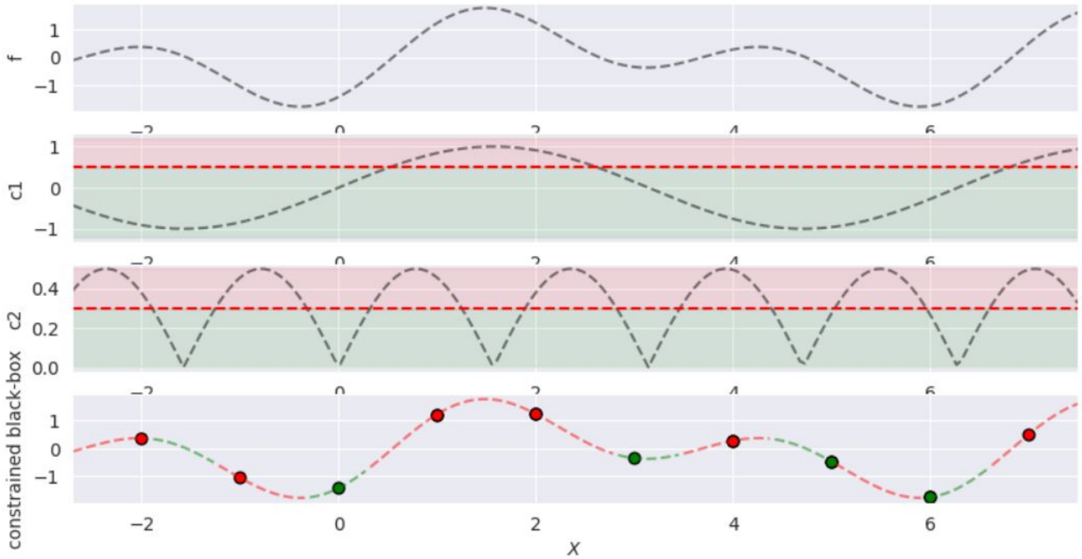
- Simple algorithm, multiple applications.
- Performs global optimization
- Decent calibration of exploration-exploitation

Bayesian Optimization with Constraints

Consider a function $f : \mathcal{X} \rightarrow \mathbb{R}$ in a bounded domain $\mathcal{X} \subseteq \mathbb{R}^D$. We aim to,

$$x^* = \text{argmax}_{x \in \mathcal{X}} f(x)$$
$$s.t. \; c_k(x) \leq v_k \text{ for all } k \in \{1, \ldots, K\}$$

Constraints are **black-box Function**, **expensive**, and potentially **noisy** x: -3

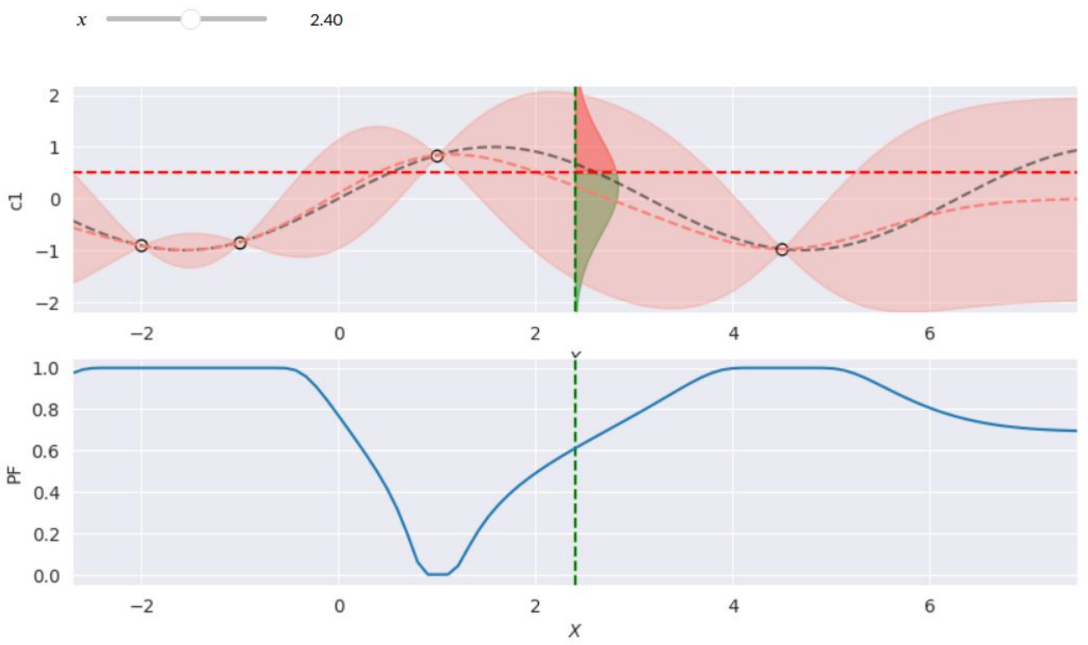


Out[12]: Toggle show/hide

Bayesian Optimization with Constraints

$$\mathrm{PF}(x) = \mathbb{P}(c_i(x) \leq v_i) = \Phi\left(\frac{v_i - \mu_k^n(x)}{\sqrt{k_k^n(x, x)}}\right)$$

- PF represents a score between (0, 1) that measures the feasibility of a point location x.
- More feasible designs would tend to 1.
- More infeasible designs would tend to 0.



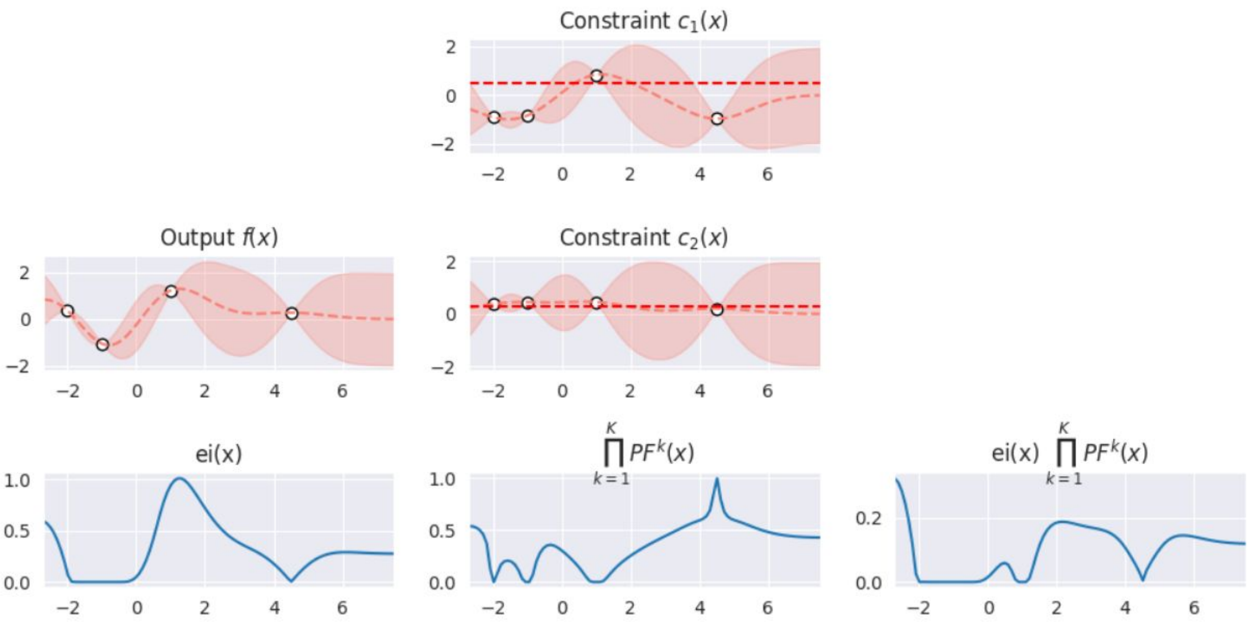
Out[26]: Toggle show/hide

Bayesian Optimization with Constraints

We may adapt any acquisition function to constrained problems by,

$$\alpha_{cons}(x) = \alpha_{uncons}(x) \prod_{k=1}^{K} \mathrm{PF}^k(x)$$

- Likely unfeasible regions are discouraged.
- Gives some importance to sampling unfeasible locations.
- Hard to sample on the feasibility boundary.



Out[15]: Toggle show/hide

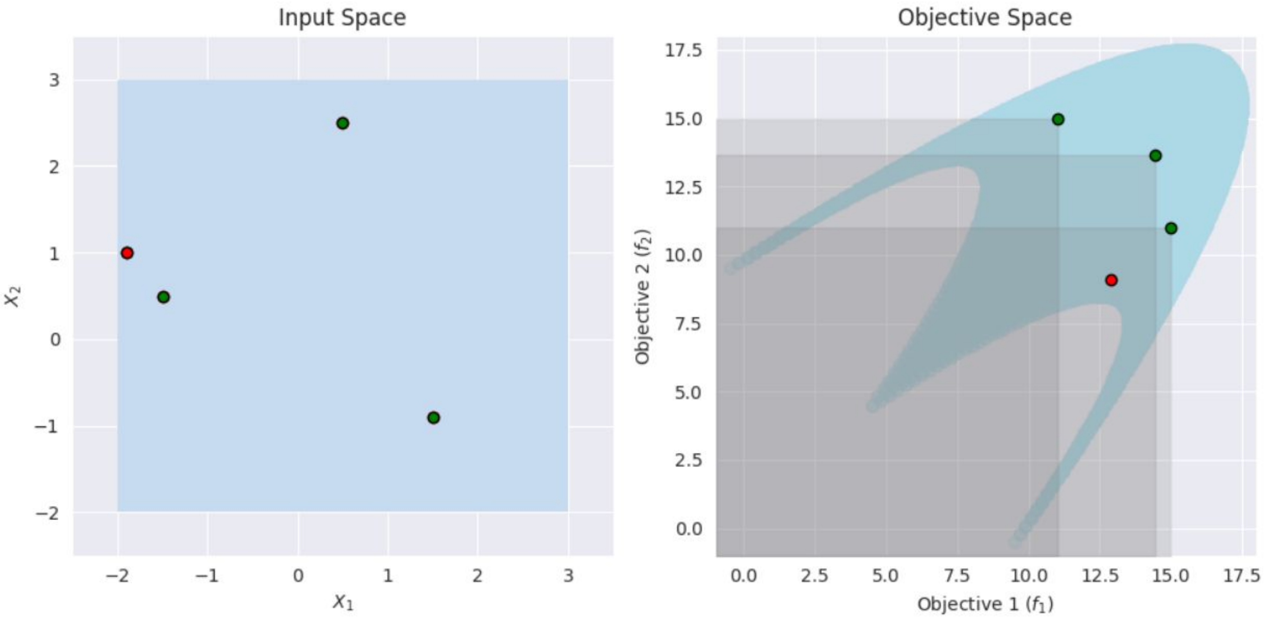
Bayesian Optimization with multiple objectives

$$\max_{x \in \mathcal{X}} f_1(x), \ldots, f_M(x)$$

- Trade-off between the different objectives.
- Each function may be an expensive-to-optimize function.

x_1 −1.90

x_2 1.00



Out[32]: Toggle show/hide

Bayesian Optimization with multiple objectives:

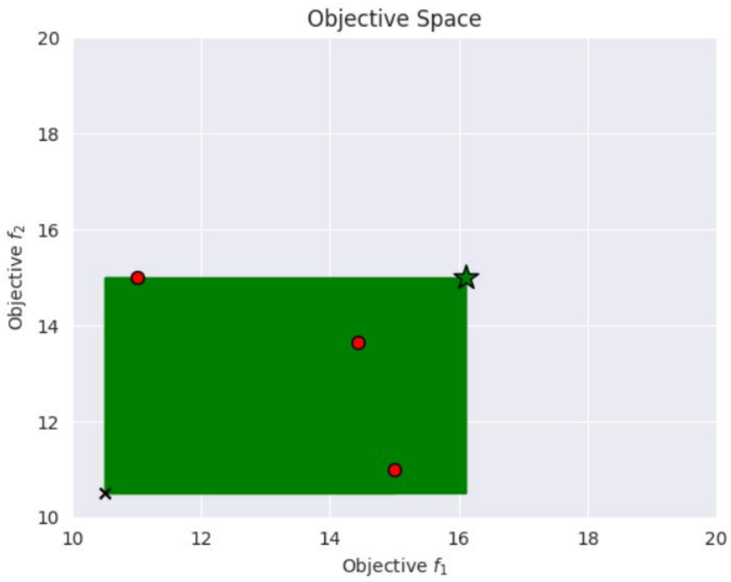
- **Hypervolume based**:
  - Computes the volume of the area enclosed by the Pareto front approximation and a reference point.
- **Scalarization based**:
  - Objectives may be aggreaged by an scalarization function, e.g.,

$$U(x) = \sum_{j=1,\ldots,M} \theta_j f_j(x) \quad , s.t. \sum_{j=1,\ldots,M} \theta_j = 1$$

metric hypervolume

f_1 16.10

f_2 15.00



Out[28]: Toggle show/hide

Bayesian Optimization with multiple objectives

- Hypervolume Based BO algorithms
  - SExI-EGO [Emmerich et al. (2011)].
  - EMO [Couckuyt et al. (2014)].
  - BMOO [Feliot et al. (2017)].
- Scalarization Based BO algorithms
  - ParEGO [Knowles (2006)].
  - EI-UU [Astudillo et al. (2020)].
  - MOEA/D-EGO [Zhang et al. (2010)].

see Rojas-Gonzalez, et al "A survey on kriging-based infill algorithms for multiobjective simulation optimization"

Other exotic settings.

- Optimize problems with a high number of input/output dimensions.
- Optimizing over non-euclidian spaces.
- Including user preferences for multi-objective problems.
- Acquisition functions with multiple steps.
- Bayesian optimization with heteroskedastic noise.

Benefits

- Global Optimization of black-box and (potentially) functions
- Sample Efficient

Difficulties

- Limited to Smooth Functions
- Struggles to scale to high number of dimensions or observations.