Bayesian Optimization of Expensive Cost Functions

Archit Dabral

22124011
Department of Mathematical Sciences
Indian Institute of Technology (BHU)
Varanasi-221 005
(Under Dr. Amit Kumar)

April 21, 2025

Definition 1.1

An enormous body of scientific literature has been devoted to the problem of optimizing a nonlinear function $f(x)$ over a *compact set* A. In realm of optimization, this problem is formulated concisely as follows: That is,

$$\max_{x \in A \subseteq \mathbb{R}^d} f(x)$$

Now consider the following assumptions with out objective function f(x) :-

(1) Evaluating Objective function is **expensive**

(2) The derivatives and convexity properties are **unknown**

$f(x)$ is a **black-box function**

Bayes' Theorem

Bayes' theorem provides a principled way to update our belief about a hypothesis based on new evidence.

$$P(\theta \mid D) = \frac{P(D \mid \theta) \cdot P(\theta)}{P(D)}$$

- $P(\theta \mid D)$: Posterior – updated belief after seeing data
- $P(D \mid \theta)$: Likelihood – probability of data given hypothesis
- $P(\theta)$: Prior – initial belief before seeing data
- $P(D)$: Evidence – overall probability of the data

Bayes' Theorem is the foundation of **Bayesian Optimization**, where we treat the objective function as a **probabilistic model** and update our beliefs as we collect more data.

$$P(\theta \mid D) \propto P(D \mid \theta) \cdot P(\theta)$$

Notation

- Let x_i denote the i-th input sample.
- Let $f(x_i)$ denote the observed value of the objective function at x_i.
- Let $\mathcal{D}_{1:t} = \{x_{1:t}, f(x_{1:t})\}$ represent the dataset of observations collected up to time step t.

Bayesian Inference

- We assume a **prior distribution** over functions: $P(f)$.
- We compute a **likelihood** of observing data $\mathcal{D}_{1:t}$ under function f: $P(\mathcal{D}_{1:t} \mid f)$.
- Bayes' Theorem gives us the **posterior distribution**:

$$P(f \mid \mathcal{D}_{1:t}) \propto P(\mathcal{D}_{1:t} \mid f) \cdot P(f)$$

Interpretation

If our prior belief is that f is smooth and low-noise, then a dataset with large variance or sudden changes is considered less likely. The posterior balances prior assumptions with observed data.

Posterior and Surrogate Modeling

- The **posterior distribution** reflects our updated belief about the unknown objective function f after observing data.
- This posterior allows us to construct a **surrogate function** (also called a **response surface**).
- In Gaussian Process (GP) modeling, the surrogate is typically the **posterior mean** of the GP.

Acquisition Function and Sampling

- An **acquisition function** is used to decide the next sampling point $x_{t+1} \in A$.
- It encodes a trade-off between:
  - **Exploration**: sampling in regions with high uncertainty.
  - **Exploitation**: sampling where the function is expected to be high.
- The acquisition function uses both the mean and variance of the GP to model the utility of sampling.

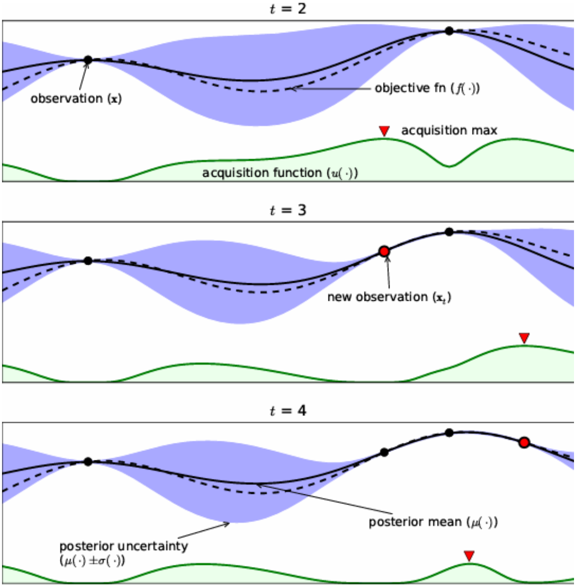
Iterative Optimization Process

1. Start with an initial set of observations (typically two points).
2. At each iteration:
    - Maximize the acquisition function to choose x_{t+1}.
    - Evaluate the true objective function at x_{t+1}.
    - Update the Gaussian Process with the new data.
3. Repeat until the evaluation budget is exhausted or convergence is reached.

Bayesian Optimization is particularly powerful when function evaluations are expensive and the function is non-convex or multimodal.

Introduction and Motivation

Bayesian Optimization Overview



Bayesian Optimization

Assumptions

- Optimization is a broad and foundational field in mathematics.
- In this context, we narrow it to:
  - **Maximization problems** instead of the more common minimization.
  - **Real-valued objective function:** $x = \arg\max_x f(x)$
  - Minimization of $-f(x)$ is equivalent to maximization of $f(x)$.
- The objective function $f(x)$ is assumed to be **Lipschitz-continuous**.
- There exists a constant C such that for all $x_1, x_2 \in A$:

$$|f(x_1) - f(x_2)| \leq C \|x_1 - x_2\|$$

- C may be unknown in practice.
- We are interested in **global** optimization, not local.
- A local maximum x^* satisfies:

$$f(x^*) \geq f(x) \quad \text{for all } x \text{ near } x^*$$

- If $f(x)$ is convex, local maxima are also global.
- However, we do **not** assume convexity of $f(x)$.

- Many global optimization methods exist:
  - **Deterministic:** Interval optimization, branch-and-bound.
  - **Stochastic:** Stochastic approximation (e.g., RL, deep learning).
- **Problem:** Most methods require **many expensive evaluations**, making them unsuitable for domains like active user-modelling.
- Even in a noise-free domain, evaluating an objective function with Lipschitz continuity C on a d-dimensional unit hypercube, guaranteeing the best observation

  $$f(x^+) \geq f(x^*) - \epsilon$$

  requires at least $\left(\frac{C}{2\epsilon}\right)^d$ samples[1] — exponential in d!
- **Bayesian Optimization:**
  - Relaxes worst-case guarantees.
  - Uses prior belief and evidence to update a probabilistic model.
  - Selects the next evaluation by maximizing the **expected gain**.
  - Focuses on **average-case performance** — often faster in practice.
- Ideal when:
  - Evaluations are expensive.
  - Objective is a black-box.
  - Sample efficiency is critical.

[1] [Betro, 1991] B. Betro. Bayesian methods in global optimization. *J. Global Optimization*, 1:114, 1991.

- **Original problem:**

$$\max_{x \in \mathcal{X}} f(x)$$

  - Black-box function $f(x)$: expensive, unknown, noisy.

- **Bayesian optimization reframes this as:**

$$x_t = \arg\max_{x \in \mathcal{X}} u(x \mid D_{1:t-1})$$

  - Where $u(x)$ is the **acquisition function** — the utility of sampling x.
  - Acquisition is computed using a **posterior distribution** over f.

- **Principle of Maximum Expected Utility:**
  - Choose x_t to maximize the **expected utility** under the posterior.

$$\mathbb{E}_{f \sim P(f \mid D_{1:t-1})}[u(x)]$$

  - Equivalent to minimizing expected risk.
  - Utility encodes goals like "likely to improve" or "reduce uncertainty".

- **Two-level optimization:**
  1. Learn posterior $P(f \mid D_{1:t})$ using prior $P(f)$ and likelihood.
  2. Optimize $u(x)$ (acquisition) to choose next query point x_t.

- **Surrogate function:** The posterior defines a cheap-to-evaluate *surrogate* for $f(x)$ (e.g., Gaussian Process).

- **Noisy setting:** Observations are $y_i = f(x_i) + \epsilon_i$, with $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

- Any Bayesian method depends on a prior distribution. A Bayesian optimization method will converge to the optimum if:
  1. The acquisition function is continuous and approximately minimizes the risk (expected deviation from the global minimum at a fixed point x).
  2. The conditional variance converges to zero (or a positive minimum under noise) **iff** the distance to the nearest observation is zero [Mokus1982, Mokus19194].
  3. The objective function is continuous.
  4. The prior is homogeneous (stationary covariance structure).
  5. Optimization is independent of the mth differences (e.g., linear trends do not affect inference).
- GP priors for Bayesian optimization date back to the late 1970s [ohagan1978,zilinskas1980].
- A Gaussian Process (GP) is a distribution over functions:

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

- A GP returns a mean and variance at any x, modeling uncertainty over function values.

Bayesian Optimization

Prior over functions

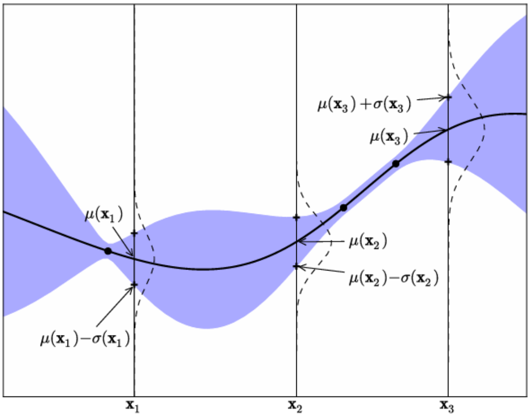


Figure: Simple 1D Gaussian process with three observations. The solid black line is the GP surrogate mean prediction of the objective function given the data, and the shaded area shows the mean plus and minus the variance. The superimposed Gaussians correspond to the GP mean and standard deviation of prediction at the points, x 1:3

Bayesian Optimization

Posterior and Predictive Distribution

- We use a zero-mean function $m(x) = 0$ for simplicity.[1]
- The covariance function (kernel) defines similarity between points. A popular choice is the **squared exponential kernel**:

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}\|x_i - x_j\|^2\right)$$

- This kernel leads to a smooth function prior and satisfies the convergence requirements in [4].
- The prior over function values at points $\mathbf{x}_{1:t}$ is:

$$\mathbf{f}_{1:t} \sim \mathcal{N}(0, K)$$

where K is the kernel matrix.

- Given past observations $\mathcal{D}_{1:t} = \{\mathbf{x}_{1:t}, \mathbf{f}_{1:t}\}$, we predict the function value at a new point x_{t+1} using the **posterior distribution**:

$$f_{t+1} \mid \mathcal{D}_{1:t}, x_{t+1} \sim \mathcal{N}\left(\mu_t(x_{t+1}), \sigma_t^2(x_{t+1})\right)$$

- Where:

$$\mu_t(x_{t+1}) = \mathbf{k}^\top K^{-1}\mathbf{f}_{1:t}, \quad \sigma_t^2(x_{t+1}) = k(x_{t+1}, x_{t+1}) - \mathbf{k}^\top K^{-1}\mathbf{k}$$

- $\mathbf{k}$ is the vector of kernel evaluations between x_{t+1} and $\mathbf{x}_{1:t}$.

[1] Alternative priors for the mean can be found in [3, 1].

Why are Acquisition Functions Important?

In Bayesian Optimization, the acquisition function determines where to sample next. It balances **exploration** (sampling where the model is uncertain) with **exploitation** (sampling where the model predicts high values). Rather than optimizing the expensive black-box function directly, we optimize a cheap-to-evaluate acquisition function:

$$x_{t+1} = \arg\max_{x} u(x \mid \mathcal{D})$$

- The acquisition function $u(x)$ reflects the potential utility of evaluating at point x.
- It is high where:
    - The predicted mean is high (exploitation),
    - The uncertainty is large (exploration),
    - Or both.
- It converts the GP posterior into a criterion for selecting the next sample point.

We will study three popular acquisition functions:

1. **Probability of Improvement (PI)**
2. **Expected Improvement (EI)**
3. **Upper Confidence Bound (UCB)**

Acquisition Functions in Bayesian Optimization

Probability of Improvement

Definition

The Probability of Improvement (PI) selects the next point to maximize the probability of improving over the current best value $f(x^+)$, where:

$$x^+ = \arg \max_{x_i \in \{x_1, \ldots, x_t\}} f(x_i)$$

$$PI(x) = P(f(x) \geq f(x^+)) = \Phi\left(\frac{\mu(x) - f(x^+)}{\sigma(x)}\right)$$

where $\mu(x)$ and $\sigma(x)$ are the GP predictive mean and standard deviation, and $\Phi(\cdot)$ is the standard normal CDF.

Introducing the Exploration Term ε

PI tends to be overly exploitative. To encourage exploration, a small threshold $\varepsilon > 0$ is introduced:

$$PI_\varepsilon(x) = \Phi\left(\frac{\mu(x) - f(x^+) - \varepsilon}{\sigma(x)}\right)$$

- A larger ε: more exploration.
- A smaller ε: more exploitation.

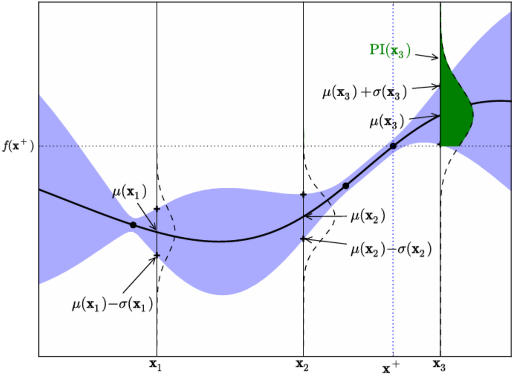


Figure: Gaussian process from the above Figure, additionally showing the region of probable improvement. The maximum observation is at x^+. The darkly-shaded area in the superimposed Gaussian above the dashed line represents the improvement region $I(x)$. The model predicts almost no possibility of improvement at x_1 or x_2, while sampling at x_3 is more likely to improve on $f(x^+)$.

Acquisition Functions in Bayesian Optimization

Expected Improvement

Definition

Expected Improvement (EI) selects the next query point by not only considering the probability of improvement but also the **magnitude** of the improvement over the current best observation $f(x^+)$.
The improvement function is defined as:

$$I(x) = \max(0, f(x) - f(x^+))$$

Then the expected improvement is:

$$EI(x) = \mathbb{E}[I(x)] = (\mu(x) - f(x^+))\Phi(Z) + \sigma(x)\phi(Z)$$

where

$$Z = \frac{\mu(x) - f(x^+)}{\sigma(x)}$$

and $\mu(x)$, $\sigma(x)$ are the predictive mean and standard deviation of the Gaussian Process, Φ is the CDF and ϕ is the PDF of the standard normal distribution.

Interpretation and Properties

- Balances **exploration** and **exploitation** in a natural, principled way.
- Encourages sampling points that are likely to yield large improvements.

Acquisition Functions in Bayesian Optimization

Upper Confidence Bound (UCB)

Definition

Upper Confidence Bound (UCB) selects points for evaluation based on the upper confidence bound of the prediction site. It was introduced by Cox and John (1992, 1997) in their "Sequential Design for Optimization" (SDO) algorithm.
For maximization problems, the UCB acquisition function is defined as:

$$\mathrm{UCB}(x) = \mu(x) + \kappa\sigma(x)$$

where:

- $\mu(x)$ is the predictive mean of the Gaussian Process
- $\sigma(x)$ is the predictive standard deviation
- $\kappa \geq 0$ is the parameter that controls the exploration-exploitation trade-off

Variants

The GP-UCB variant proposed by Srinivas et al. (2010) is defined as:

$$\text{GP-UCB}(x) = \mu(x) + \sqrt{\nu\tau}\sigma(x)$$

where $\nu > 0$ is a hyperparameter and $\tau = \sqrt{\nu\tau}$ is typically set to control the confidence level.

Modeling Observation Noise

In real-world settings, observations are noisy:

$$y_i = f(x_i) + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2_{\text{noise}})$$

This modifies the GP's predictive variance and mean:

$$\mu_t(x) = k^T[K + \sigma^2_{\text{noise}} I]^{-1} y \qquad \sigma_t^2(x) = k(x,x) - k^T[K + \sigma^2_{\text{noise}} I]^{-1} k$$

Applications of Bayesian Optimization

Key Application Areas

- **Hyperparameter Tuning in Machine Learning**
  - Deep neural networks (e.g., learning rate, dropout)
  - SVMs, XGBoost, etc.
- **Automated Machine Learning (AutoML)**
  - Model selection, pipeline optimization
  - Tools: Auto-WEKA, SMAC, Hyperopt, BOHB
- **Experimental Design**
  - Physics, chemistry, biology lab experiments
- **Robotics**
  - Policy optimization, trajectory tuning
- **Reinforcement Learning**
  - Policy parameter optimization under limited trials
- **Engineering Optimization**
  - Aerospace design, CFD simulations

Takeaway

Bayesian Optimization excels where every query is expensive, noisy, or risky—bridging the gap between theory and practical deployment.

References (1/2)

Brochu, E., Cora, V. M., and de Freitas, N. (2010).
A Tutorial on Bayesian Optimization of Expensive Cost Functions, with Application to Active User Modeling and Hierarchical Reinforcement Learning.
arXiv preprint arXiv:1012.2599.

Betro, B. (1991).
Bayesian methods in global optimization.
Journal of Global Optimization, 1:1–14.

Mockus, J. (1982).
The Bayesian Approach to Global Optimization.
In *System Modeling and Optimization*.

Mockus, J. (1994).
Application of Bayesian Approach to Numerical Methods of Global and Stochastic Optimization.
Journal of Global Optimization.

References (2/2)

O'Hagan, A. (1978).
Curve fitting and optimal design for prediction.
Journal of the Royal Statistical Society: Series B (Methodological).

Zilinskas, A. (1980).
Optimization of computer programs using statistical models of functions.
In *System Identification and Optimization*.

Martinez-Cantin, R., de Freitas, N., Brochu, E., Castellanos, J., and Doucet, A. (2009).
A Bayesian exploration-exploitation approach for optimal online sensing and planning with a visually guided mobile robot.
Autonomous Robots, 27(2):93–103.

Thank You!

Questions or Comments?

Archit Dabral `22124011`