Enhancing Gaussian Process Surrogates for Optimization via Random Exploration

Hwanwoo Kim

with Daniel Sanz-Alonso (U.Chicago)

Department of Statistical Science, Duke University

May 7, 2025

Gaussian Process

- GP(0,k) is a prior distribution over the space of continuous functions, which reflects our initial knowledge on f
- Conditioning on $\mathcal{D}_t = \{(\theta_1, f(\theta_1)), \cdots, (\theta_t, f(\theta_t))\}$, the posterior mean and variance of the Gaussian process posterior are given by

$$\mu_t(\theta) = k_t(\theta)^\top K_{tt}^{-1} F_t,$$
$$\sigma_t^2(\theta) = k(\theta, \theta) - k_t(\theta)^\top K_{tt}^{-1} k_t(\theta),$$

  where $k_t(\theta) = [k(\theta, \theta_1), \ldots, k(\theta, \theta_t)]^\top$, K_{tt} is a $t \times t$ matrix with entries $(K_{tt})_{i,j} = k(\theta_i, \theta_j)$ and $F_t = [f(\theta_1), \ldots, f(\theta_t)]^\top$
- The posterior mean serves as our approximation for the objective function f and the posterior variance quantifies how accurate the approximation is

Covariance Kernels

- *Matérn kernels* with smoothness parameter ν and length scale parameter ℓ, given by

$$k(\theta, \theta') = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\frac{\sqrt{2\nu}\|\theta - \theta'\|}{\ell} \right)^{\nu} B_\nu \left(\frac{\sqrt{2\nu}\|\theta - \theta'\|}{\ell} \right),$$

  where B_ν is a modified Bessel function of the second kind

Optimization Problem

Want to solve

$$\max_{\theta \in \Theta} f(\theta)$$

where $f : \Theta \subset \mathbb{R}^d \to \mathbb{R}$, where

- Θ is a compact search space
- Only source of information on f is through its evaluation, which is computationally expensive
- $\Theta_t = \{\theta_1, \ldots, \theta_t\}$: function evaluation locations
- $F_t = [f(\theta_1), \ldots, f(\theta_t)]^\top$: a vector of function evaluations on Θ_t
- Examples: hyper-parameter tuning for large-scale neural network models, computer model calibrations, MAP estimates under intractable likelihoods and etc

Gaussian Process Upper Confidence Bound (GP-UCB)

Input: Kernel k; Total number of evaluations T; Initial query locations Θ_0; Initial noise-free observations F_0
For $t = 1, \cdots, T$:

1. Compute posterior mean/variance using all query locations and function evaluations (Θ_{t-1}, F_{t-1})
2. Select a subsequent location to evaluate function f:

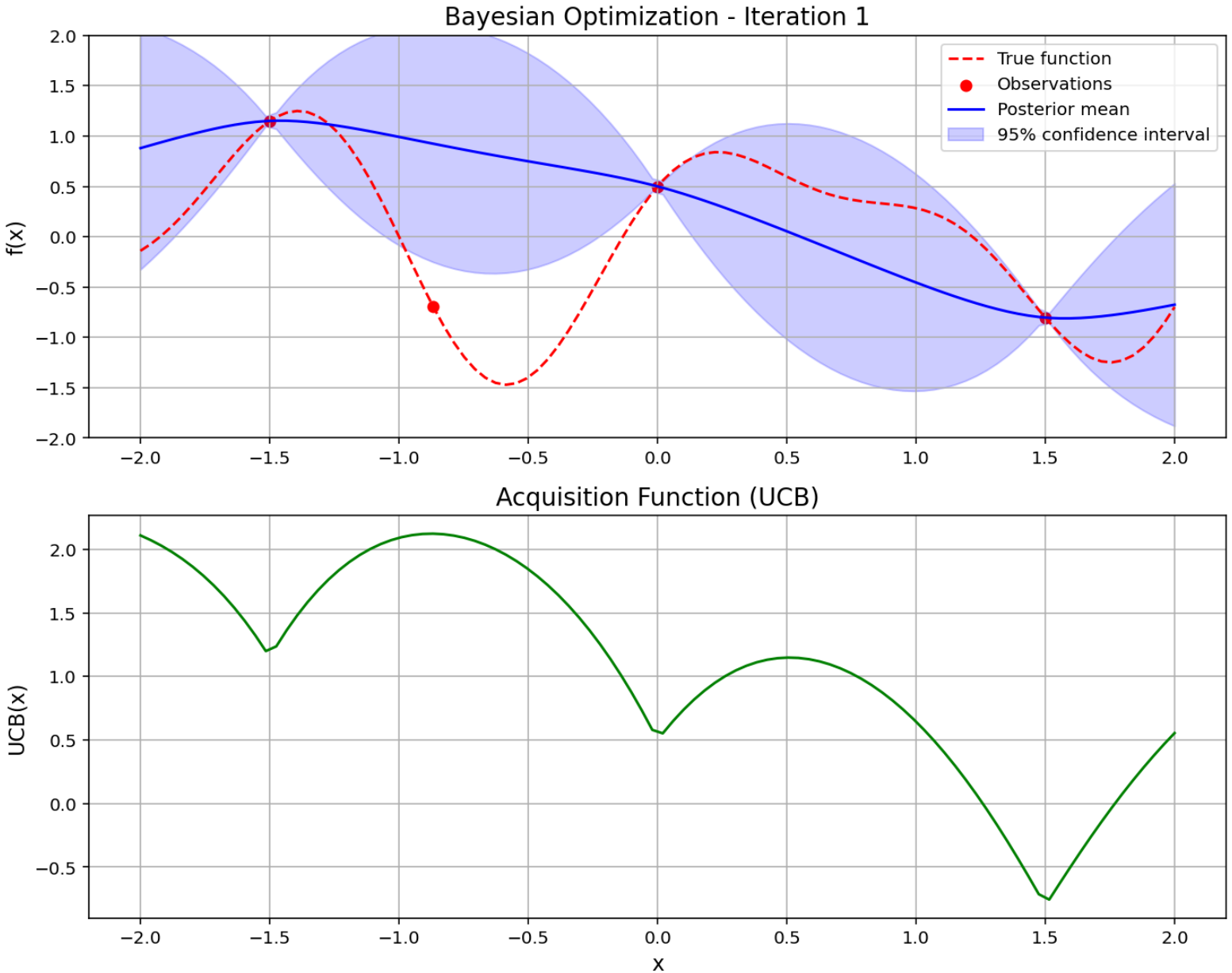
$$\theta_t = \arg\max_{\theta \in \Theta} \mu_{t-1}(\theta) + \beta_t \sigma_{t-1}(\theta),$$

   where $\beta_t \in \mathbb{R}_+$
3. Set $\Theta_t = \Theta_{t-1} \cup \{\theta_t\}$, $F_t = F_{t-1} \cup \{f(\theta_t)\}$

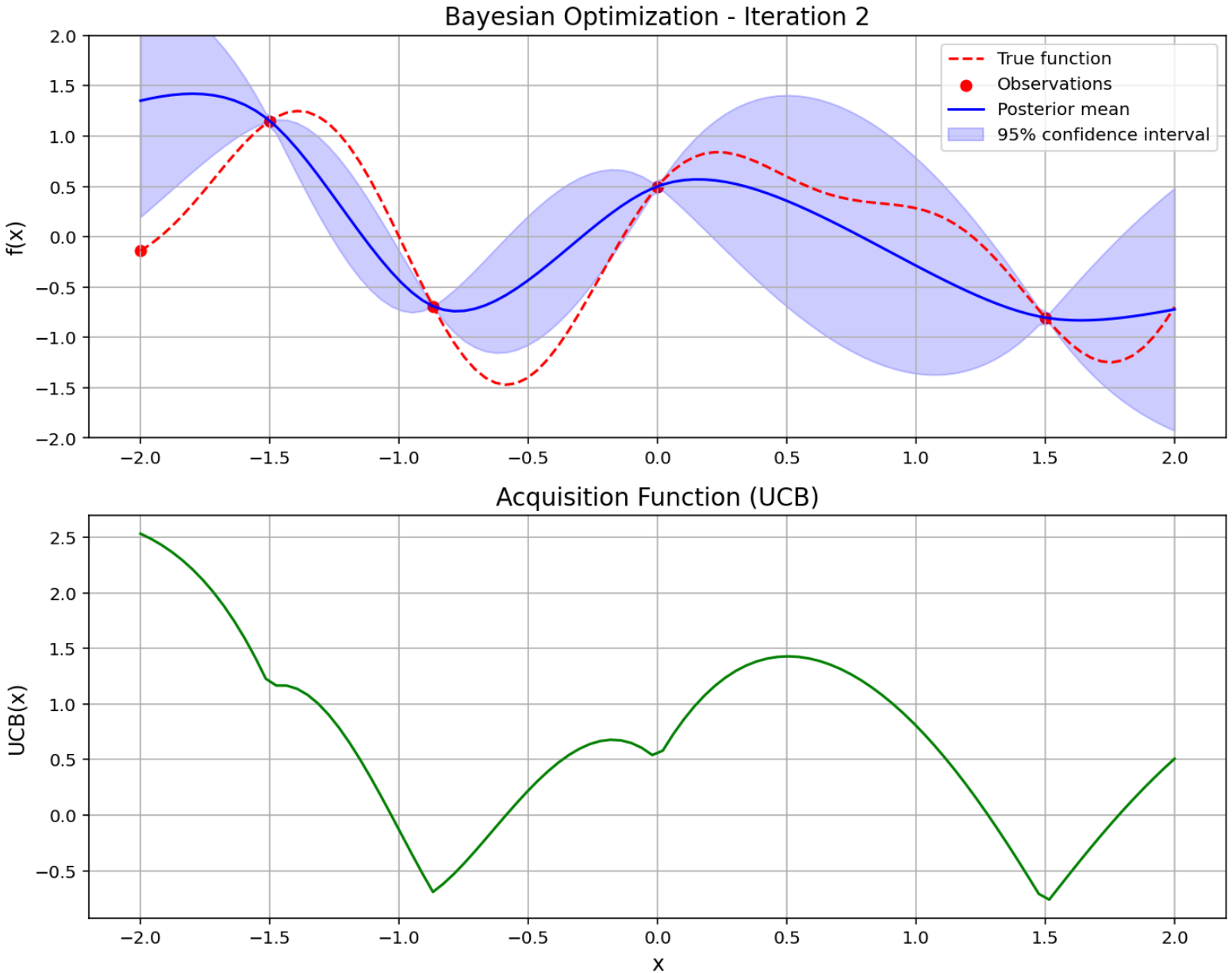
Output: $\arg\max_{\theta \in \Theta_T} f(\theta)$

More details in (Srinivas, Krause, Kakade, & Seeger, 2009)

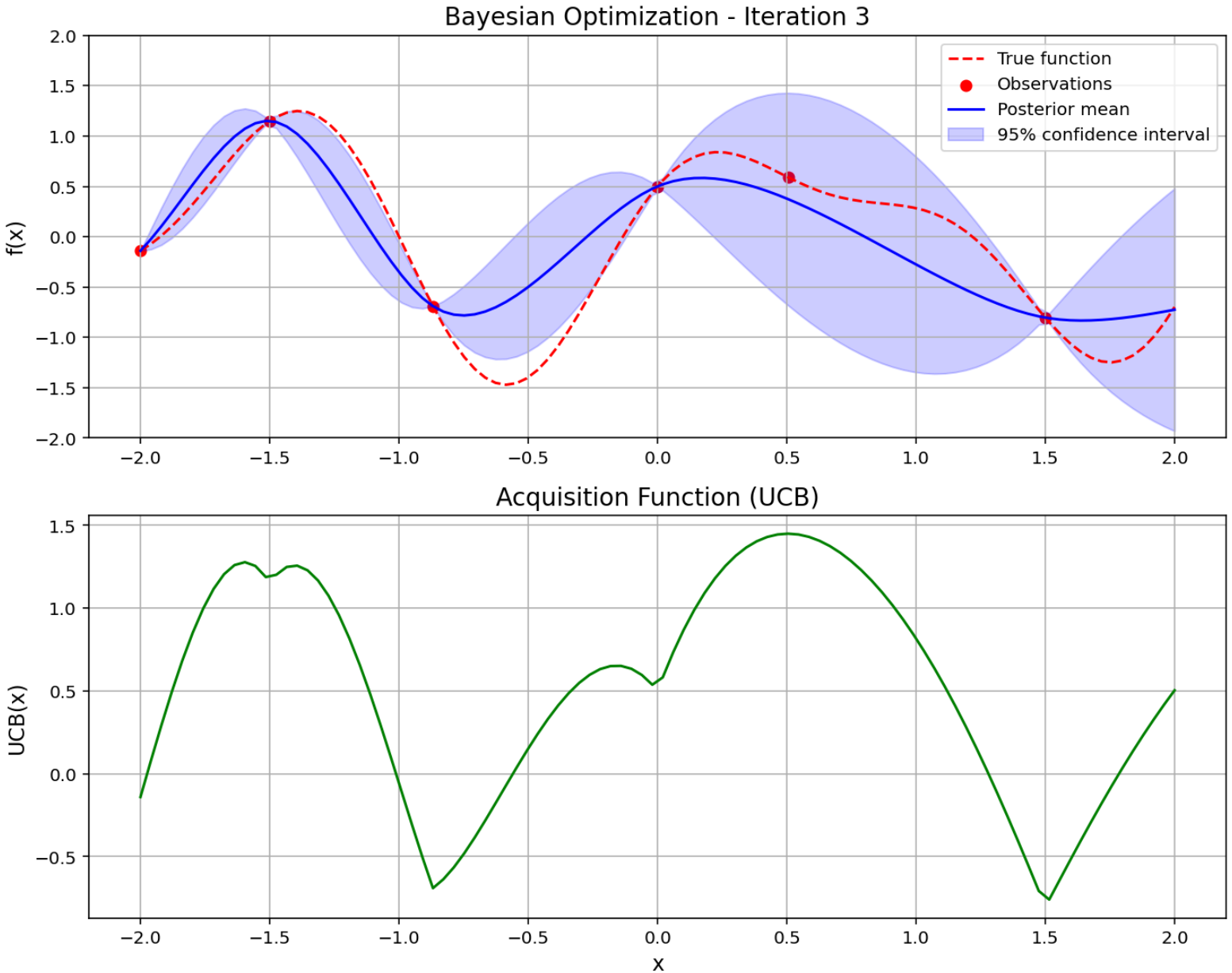
Bayesian Optimization: First Iteration



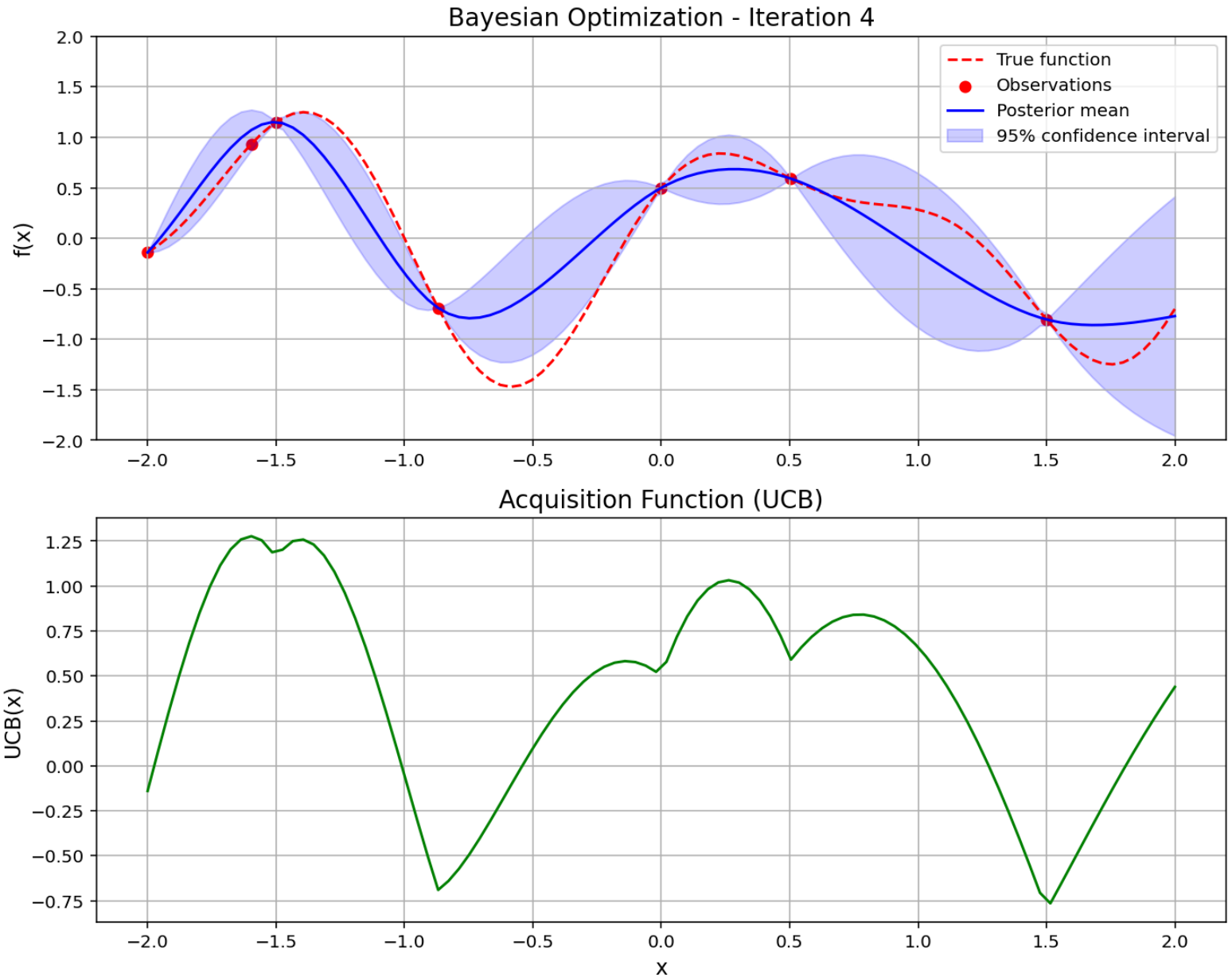
Bayesian Optimization: Second Iteration



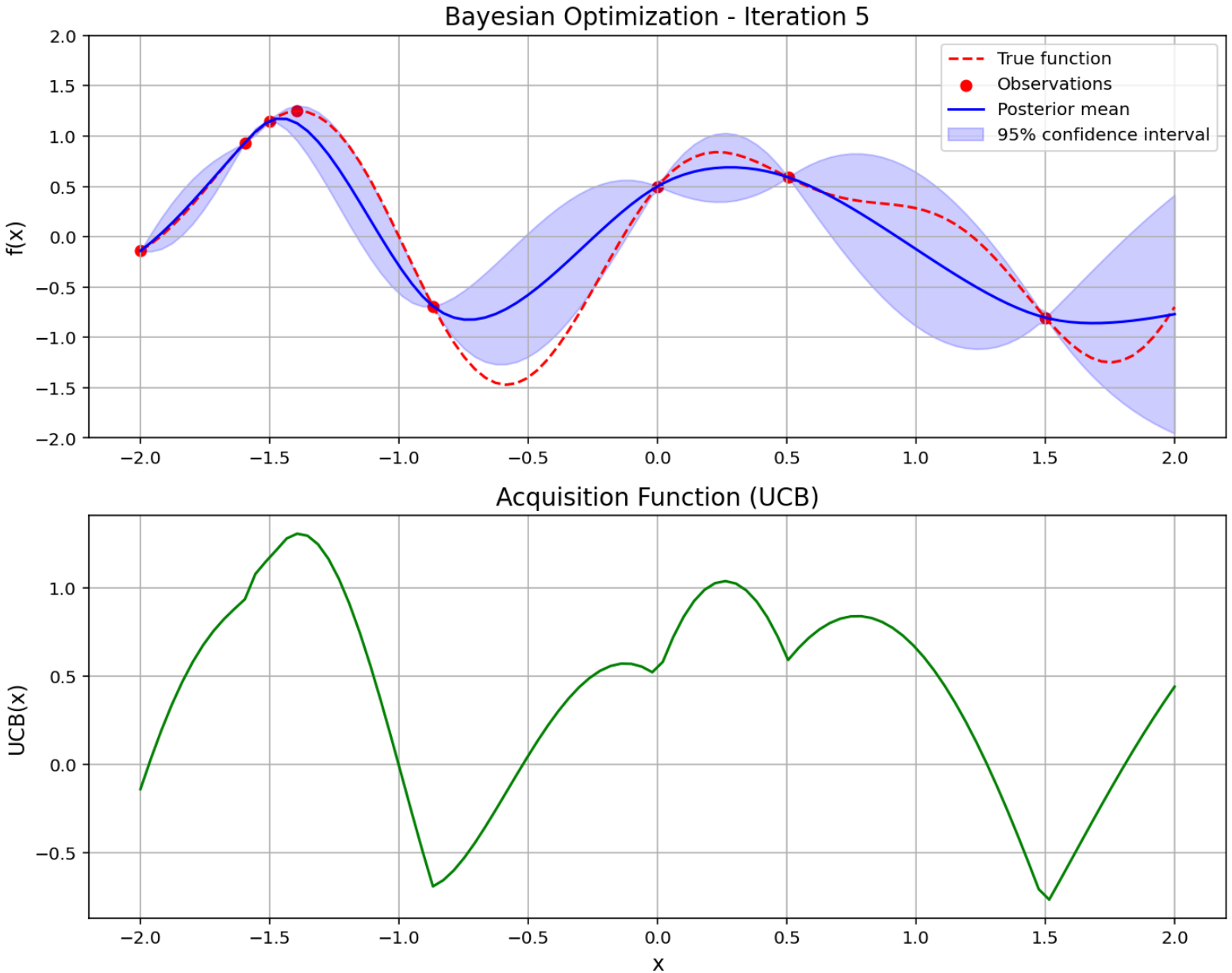
Bayesian Optimization: Third Iteration



Bayesian Optimization: Fourth Iteration



Bayesian Optimization: Fifth Iteration



Performance Metric

- Cumulative Regret: $R_T = \sum_{t=1}^{T} (\max_{\Theta} f(\theta) - f(\theta_t))$
- Simple Regret: $S_T = \max_{\Theta} f(\theta) - \max_{t=1,\cdots,T} f(\theta_t)$
- Simple Regret $(S_T) \leq$ Averaged Cumulative Regret (R_T/T)
- Sublinear growth of R_T leads to the convergence of S_T
- Convergence rate of S_T is bounded by that of R_T/T

Existing Regret Bounds

- (Lyu, Yuan, & Tsang, 2019) Under the reproducible kernel Hilbert space (RKHS) objective function assumption, GP-UCB with $\beta_t = \|f\|_{\mathcal{H}_k}$ yields

$$R_T = \tilde{\mathcal{O}}\left(T^{\frac{\nu+d}{2\nu+d}}\right) \text{ for Matérn kernel,}$$

  which implies,

$$S_T = \tilde{\mathcal{O}}\left(T^{\frac{-\nu}{2\nu+d}}\right)$$

- Optimal convergence rate (Bull, 2011): Under RKHS assumption, for Matérn kernels with smoothness parameter $\nu > 0$,

$$S_T = \Theta\left(T^{-\frac{\nu}{d}}\right),$$

  for the best strategy

GP-UCB+ & EXPLOIT+ : Algorithm

Input: Kernel k; Total number of evaluations T; Initial query locations Θ_0; Initial noise-free observations F_0; Probability distribution P on Θ.
For $t = 1, \cdots, T/2$:

1. Compute posterior mean/variance using all query locations and function evaluations (Θ_{t-1}, F_{t-1})
2. Obtain one location to evaluate a function f either through
   - GP-UCB+: $\theta_t = \arg\max \mu_{t-1}(\theta) + \beta_t \sigma_{t-1}(\theta)$
   - EXPLOIT+: $\theta_t = \arg\max \mu_{t-1}(\theta)$
3. Acquire another location to evaluate a function f via random sampling $\tilde{\theta}_t \sim P$
4. Set $\Theta_t = \Theta_{t-1} \cup \{\theta_t, \tilde{\theta}_t\}$, $F_t = F_{t-1} \cup \{f(\theta_t), f(\tilde{\theta}_t)\}$

Output: $\arg\max_{\theta \in \Theta_T} f(\theta)$

GP-UCB+ & EXPLOIT+ : Theory

Theorem (Kim and Sanz-Alonso, 2024)

Under RKHS assumption: With $\beta_t = \|f\|_{\mathcal{H}_k}$*, Matérn kernels with a smoothness parameter* $\nu > 0$*,*

$$\mathbb{E}_P[S_T] = \mathcal{O}\left(T^{-\frac{\nu}{d}+\varepsilon}\right)$$

where $\varepsilon > 0$ *can be arbitrarily small. For squared exponential kernels,*

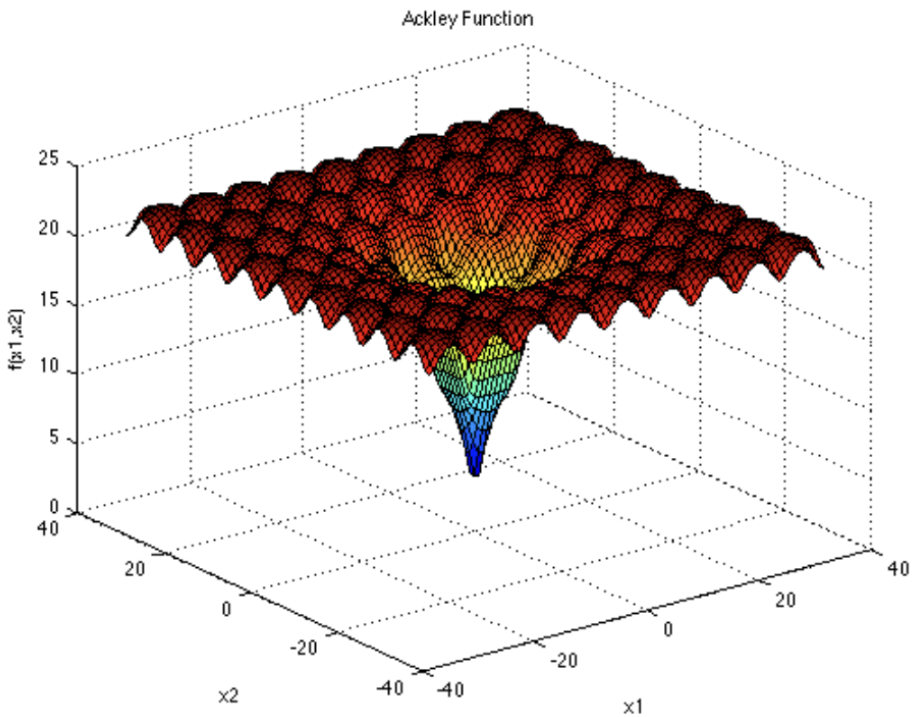
$$\mathbb{E}_P[S_T] = \mathcal{O}\left(\exp\left(-CT^{\frac{1}{d}-\varepsilon}\right)\right),$$

for some constant $C > 0$ *with an arbitrarily small* $\varepsilon > 0$

Benchmark Function

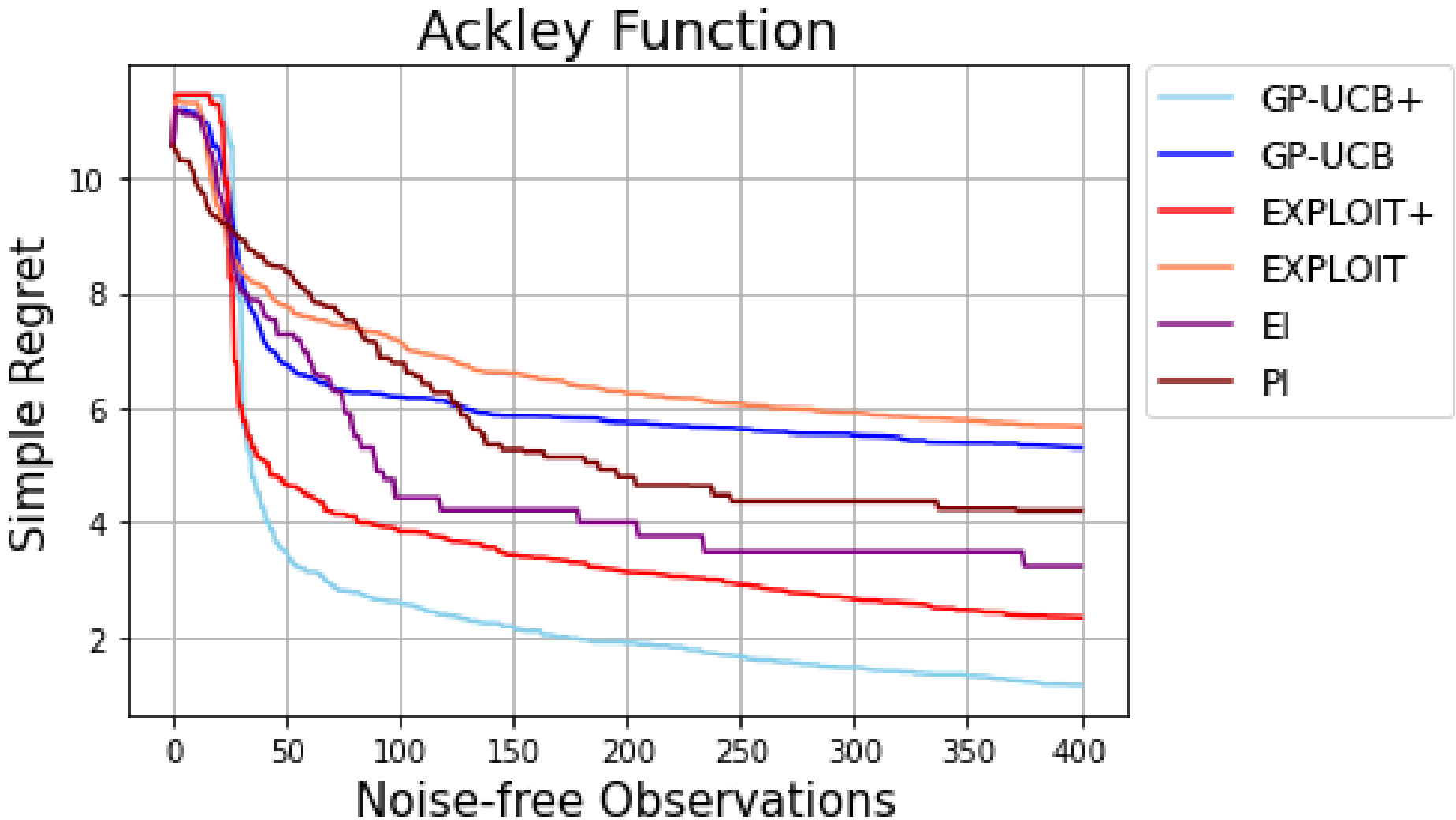
- Ackley function:

$$f(\theta) = 20 \exp\left(-\frac{1}{5}\sqrt{\frac{1}{d}\sum_{i=1}^{d}(\theta^i)^2}\right) + \exp\left(\frac{1}{d}\sum_{i=1}^{d}\cos(2\pi\theta^i)\right)$$



Numerical Results

Benchmark Ackley function (10-dim):

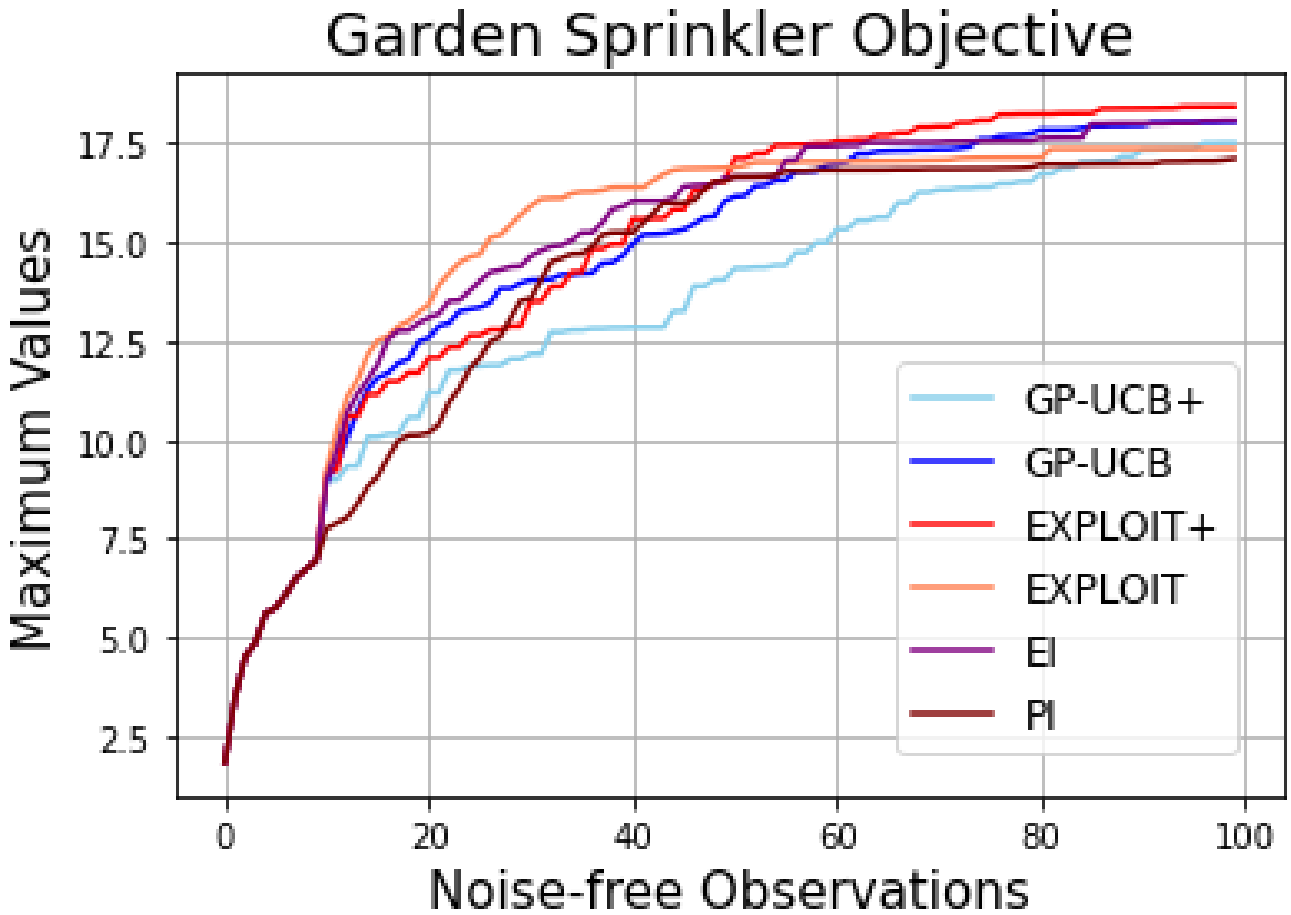


Engineering Design Problem

- Maximize the range of Garden sprinkler can water
- How to tune parameters of the Garden sprinkler design? e.g., Vertical nozzle angle, Tangential nozzle angle, Nozzle profile, Diameter of the sprinkler head, Dynamic friction moment, Static friction moment, Entrance pressure, Diameter flow line

Numerical Results: Engineering Design

Blackbox Garden springkler function (8-dim):



Optimization Algorithm as a Design Tool

- Conditioning on the set of T points acquired from the Bayesian optimization, approximate log-posterior through the posterior mean of GP surrogate
- What's good about such approximation?
  - The set of T points is more concentrated in the region of maxima due to the nature of the optimization algorithm (compared to pure random sampling)
  - The set of T points explores a wider region of search space due to a random sampling step (in comparison to standard Bayesian optimization algorithms)
  - Facilitates a cost-efficient Bayesian inference for parameters of differential equations

Example: Lorenz-63 Dynamics

Consider the Lorenz dynamics, given by

$$\begin{aligned}\frac{dx}{dt} &= \sigma(y - x),\\ \frac{dy}{dt} &= x(\rho - z) - y,\\ \frac{dz}{dt} &= xy - \beta z,\end{aligned}$$

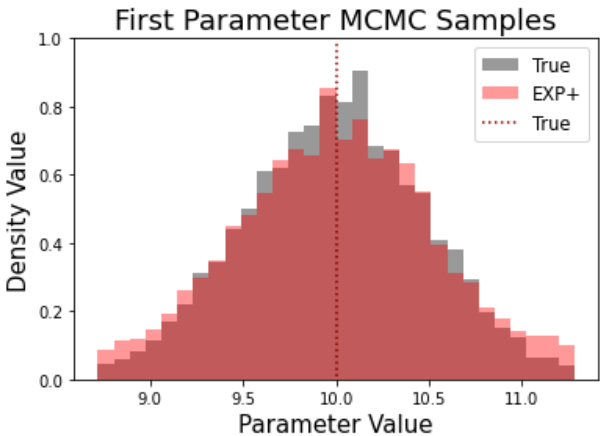
over time window [10, 200].

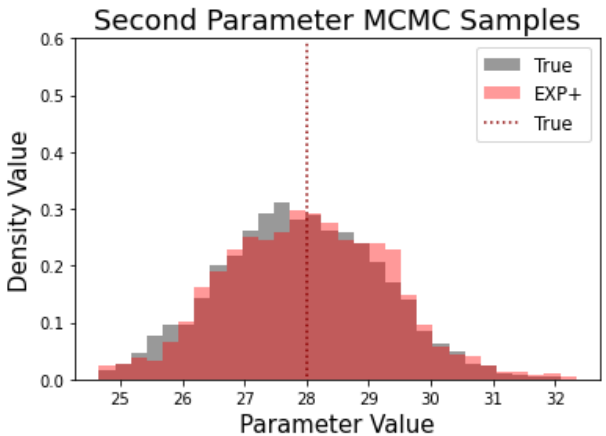
- $\theta = (\sigma, \rho, \beta)$ is an unknown parameter we wish to infer.
- Data given were generated with $\theta^* = (10, 28, 8/3)$.
- Imposed a Gaussian prior θ:

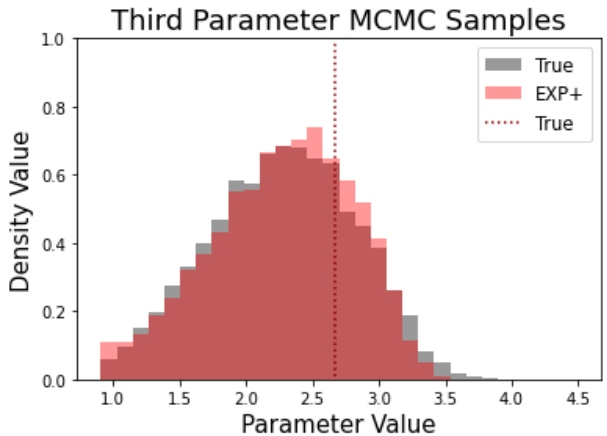
$$\theta \sim \mathcal{N}([10, 28.5, 2.7], \text{diag}([0.25, 2.25, 0.49])).$$

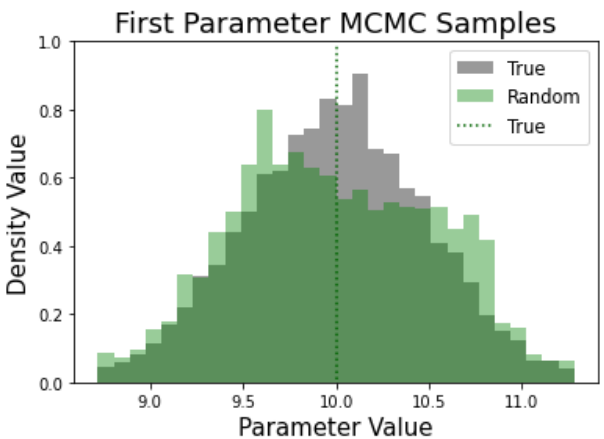
- Total 400 design points were used.

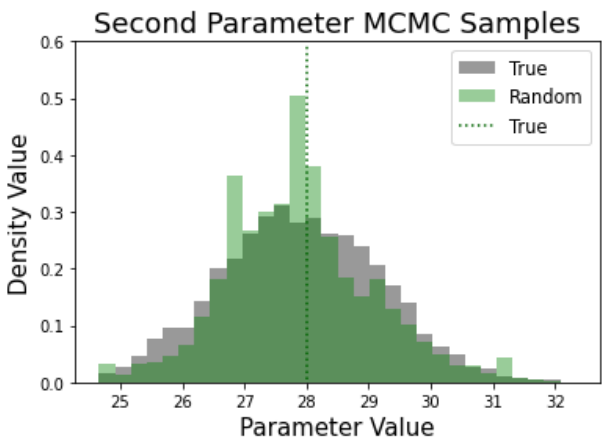
Lorenz-63 Dynamical System: Posterior MCMC Samples I

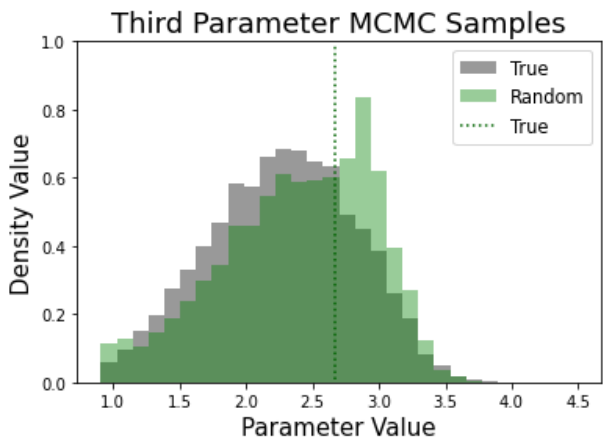




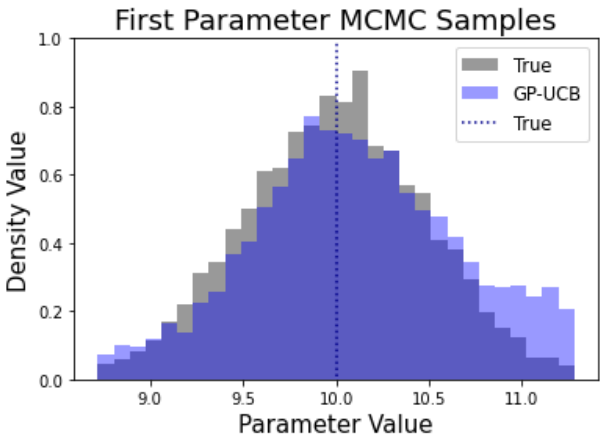


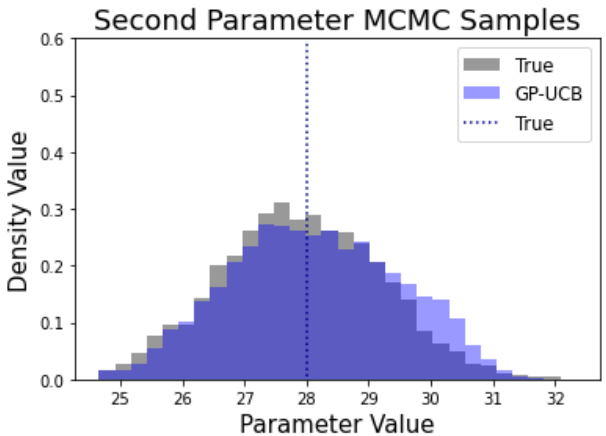


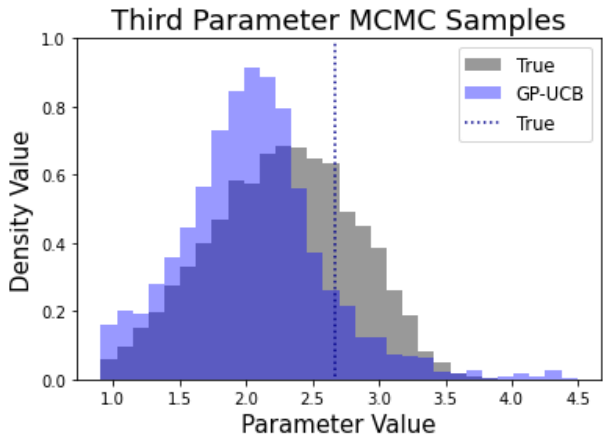


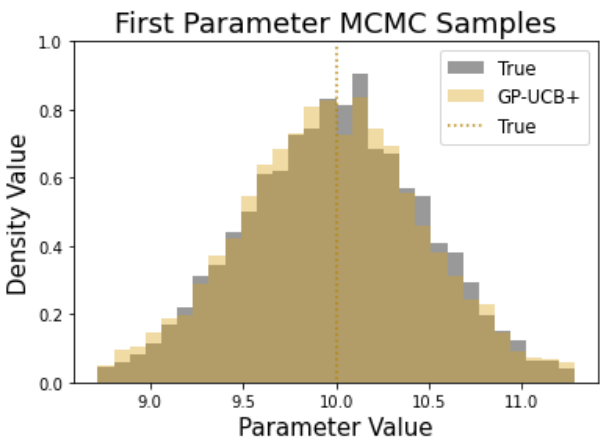


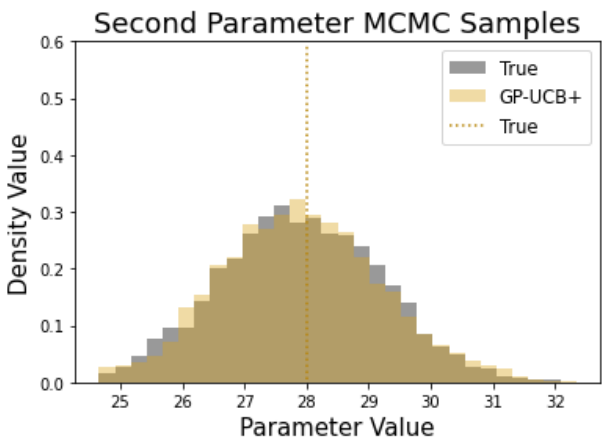
Lorenz-63 Dynamical System: Posterior MCMC Samples II

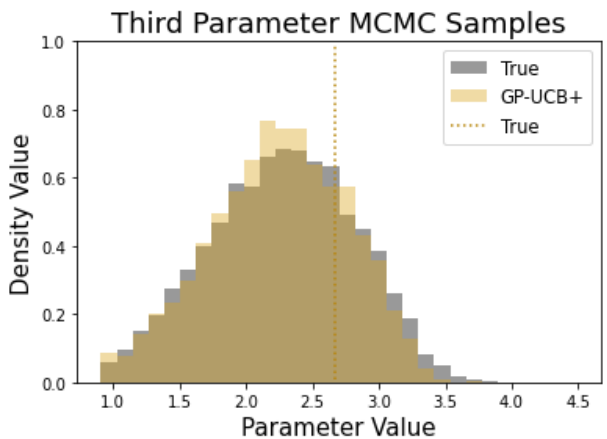












References I

Bull, A. D. (2011).
Convergence rates of efficient global optimization algorithms..
Journal of Machine Learning Research, *12*(10).

Kim, H., & Sanz-Alonso, D. (2024).
Enhancing gaussian process surrogates for optimization and posterior approximation via random exploration.
https://arxiv.org/abs/2401.17037.

Lyu, Y., Yuan, Y., & Tsang, I. W. (2019). Efficient batch black-box optimization with deterministic regret bounds.
arXiv preprint arXiv:1905.10041.

Srinivas, N., Krause, A., Kakade, S. M., & Seeger, M. (2009).
Gaussian process optimization in the bandit setting: No regret and experimental design. *arXiv preprint arXiv:0912.3995*.