Objective Bound Conditional Gaussian Process for Bayesian Optimization

Taewon Joeng, Heeyoung Kim

{papilion89, heeyoungkim} @ kaist.ac.kr

Industrial Statistics Lab, KAIST

ICML 2021

Bayesian Optimization

❖Bayesian optimization (BO)

BO is a widely used technique for black-box optimization when the objective function is expensive to evaluate.

- Strategy of BO:

Step 1: Construct a surrogate model of the black-box function. In general, a Gaussian process (GP) is used as the prior over the objective function, and the posterior GP is used as a surrogate model.

Step 2: Select the next query point based on the surrogate model using an acquisition function.

Step 3: Augment the data with the new point from Step2 and repeat Steps 1 and 2 for a sequential design process.

Algorithm 1 Basic pseudo-code for BO

Place GP prior on $f \sim GP(\mu, k)$

Observe f at n_0 points according to an initial space-filling experimental design. Set $n = n_0$ and $D_n = ((x_1, f(x_1)), ...(x_n, f(x_n)))$

while $n \leq N$ **do**

  Update the posterior probability distribution on f using D_n

  $x_{n+1} = arg\,\max_{x \in \Omega} Acq(x; D_n)$

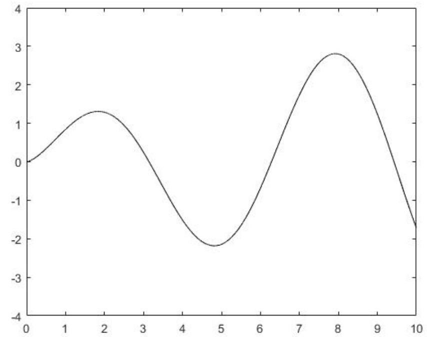
  observe $f(x_{n+1})$ and set $D_n \leftarrow D_n + ((x_{n+1}, f(x_{n+1})))$

  Increment n

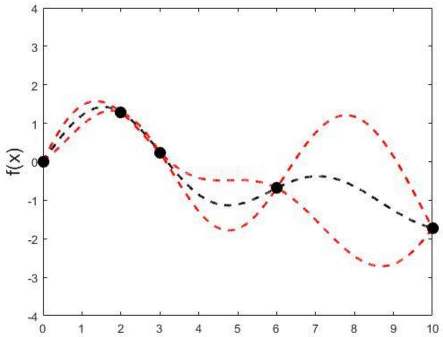
end while

Bayesian Optimization

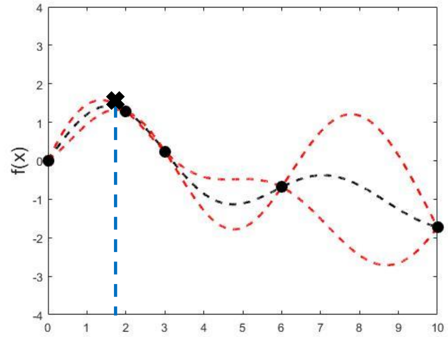
❖ Example



True objective function



Step1: Construct a surrogate model (posterior GP)



Step3: Evaluate the function value at the point from Step 2, augment the observation, and repeat steps 1 & 2.

❖Acquisition functions in BO

[1] Expected Improvement (EI)

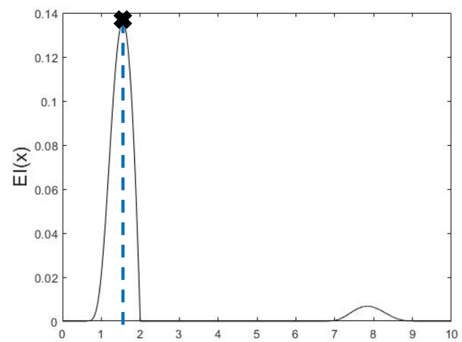
$$Acq(x; D_n) = E[(f(x) - f_{best})^+ | D_n]$$

[2] GP-UCB

$$Acq(x; D_n) = E[f(x)|D_n] + \beta\sqrt{Var[f(x)|D_n]}$$

[3] Predictive Entropy Search

$$Acq(x; D_n) = I(\{x, f(x)\}; x_{opt} | D_n)$$



Step2: Select the next query point by optimizing an acquisition function.

Motivation for New Surrogate Model

❖Bound on the optimal function value

Example 1

Past process:

$$f^{past}(\alpha_1, \alpha_2)$$

- A lot of data were observed.
- Near optimal value f_{best}^{past} was found.

Technology improvement

Current process:

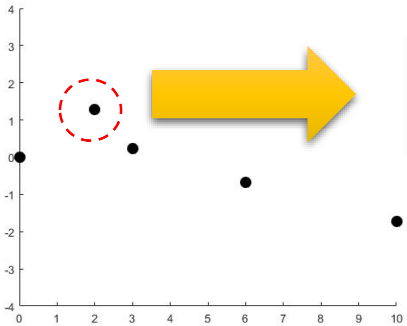
$$f^{curr}(\alpha_1, \alpha_2)$$

- Only small amount of data are observed.
- Expert knowledge: $\max f^{curr} \geq \max f^{past}$

Prior knowledge about the optimal function value:

$$\max f^{curr} \geq f_{best}^{past}$$

Example 2



Bound on the optimal function value:
$\max f \geq f_{best}$(maximum value among the observations D_n)

Motivation for new surrogate model:
We propose a novel surrogate model to incorporate the information of "existence of $x \in \Omega$, where $f(x) \geq l_p$ (or $l_p \leq f(x) \leq u_p$)"

Objective Bound Conditional GP (OBCGP)

❖Inducing parameter and inducing variable

We introduce an ***inducing parameter*** $\boldsymbol{x_M}$, which means "candidate optimal location," set its function value $\boldsymbol{f(x_M)}$ **to be an *inducing variable (latent variable)*,** and then construct the surrogate model for $f(x_M)$

[1] l_p, lower bound on the optimal function value, is given:

$$f(x_M) = l_p + Z_M, \quad Z_M \sim Exp(\lambda)$$ ➔ **support of $\boldsymbol{f(x_M)}$ becomes $[\boldsymbol{l_p}, \infty]$**

[2] not only l_p, but also u_p, the upper bound of optimal value, is given:

$$f(x_M) = l_p + (u_p - l_p)Z_M, \quad Z_M \sim Beta(1, \alpha)$$ ➔ **support of $\boldsymbol{f(x_M)}$ becomes $[\boldsymbol{l_p}, \boldsymbol{u_p}]$**

Knowledge of a bound on the optimal function value: "existence of $\boldsymbol{x_M} \in \boldsymbol{\Omega}$, where $\boldsymbol{f(x_M) \geq l_p}$ (or $\boldsymbol{l_p \leq f(x) \leq u_p}$)"

❖OBCGP as an alternative to GP to incorporate a bound on the optimal function value

Construct the conditional GP, where the formulae for mean and covariance are consistent with those for the posterior GP given $(x_M, f(x_M))$.

$$p(f_n | x_M, f(x_M)) \sim N(f_n | \mu_n, \Sigma_{n\times n})$$

$$\mu_n = f(x_M)\frac{k_M}{k(x_M, x_M)}, \qquad \Sigma_{n\times n} = K - \frac{k_M k_M^T}{k(x_M, x_M)}$$

$$f_n = \left(f(x_1), f(x_2), \dots f(x_n)\right)^T, \; k_M = \left(k(x_1, x_M), \dots k(x_n, x_M)\right)^T$$

Inference and Acquisition Functions for OBCGP

❖Parameter estimation via variational inference (VI)

$$\log p_\theta(\mathbf{f}_n; X_n) \geq L(\theta, \phi; \mathbf{f}_n, X_n)$$
$$= E_{q_\phi(Z_M)}[\log p_\theta(\mathbf{f}_n|Z_M)] - KL(q_\phi(Z_M)||p(Z_M))$$

❖Computing posterior moments for OBCGP

$$E[f(x^*)|D_n] = E[E[f(x^*)|D_n, Z_M]|D_n]$$
$$\approx \mathcal{A} + \tau E[Z_M|D_n] = \mathcal{A} + E_1^q$$

$$Var[f(x^*)|D_n] = E[f(x^*)^2|D_n] - (E[f(x^*)|D_n])^2$$
$$\approx \hat{\sigma}^2(x^*; D_n, Z_M) + \tau^2(E_2^q - \left(E_1^q\right)^2)$$

❖Acquisition functions for OBCGP

[1] Moment matching: Gaussian approximation for the OBCGP posterior based on moment matching

➔ Apply acquisition functions that are composed of only the mean and variance (e.g. EI, UCB)

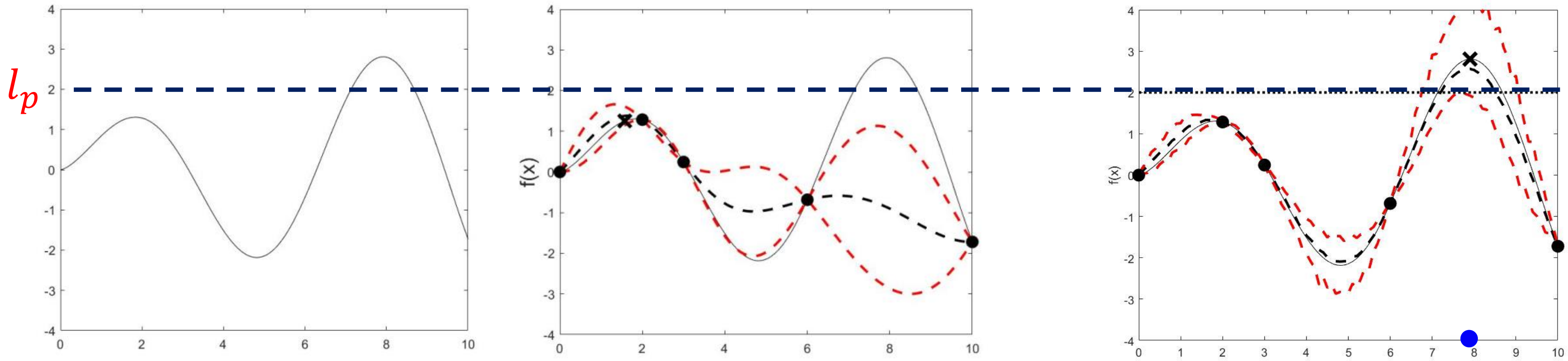
[2] Sampling methods: Sampling from the estimated variational distribution $q(f(x_M)|D_n)$

➔ Apply Monte Carlo sampling for computing acquisition functions.

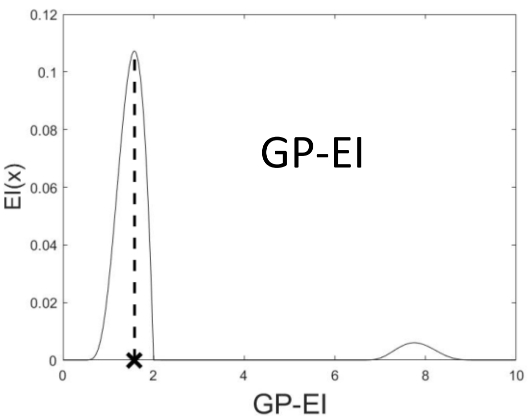
Comparison: GP vs OBCGP

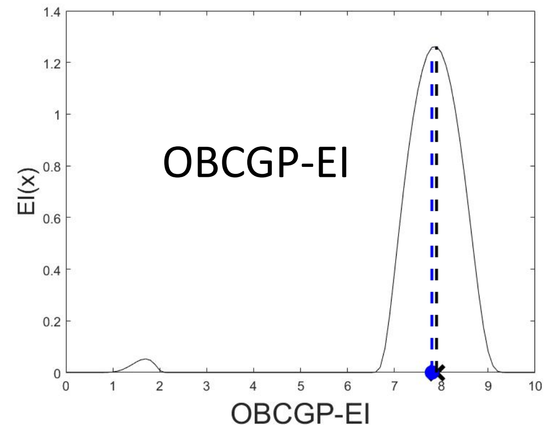
❖BO with GP vs OBCGP

Example: lower bound l_p is available



True function GP posterior OBCGP posterior





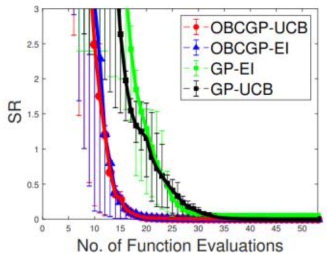
Numerical Study Results

Simple regret (SR): $|f_{opt} - f_{best}|$

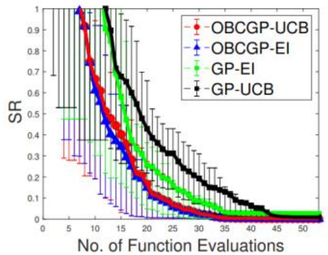
f_{opt}: True optimal function value

f_{best}: The best function value among the observations

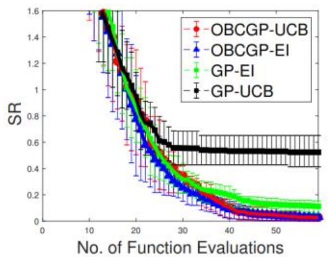
- ❖ GP-BO vs OBCGP-BO: SR comparison on test functions



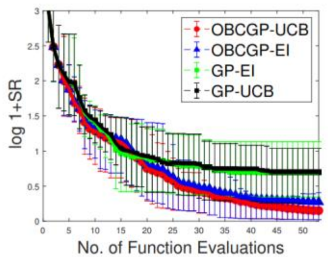
(a) Branin



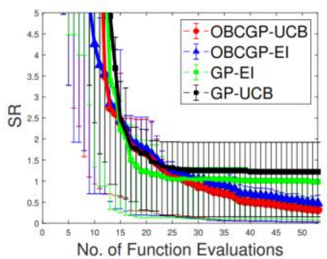
(b) Camel



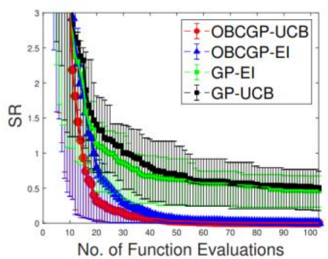
(c) Hartmann 6



(d) Goldstein

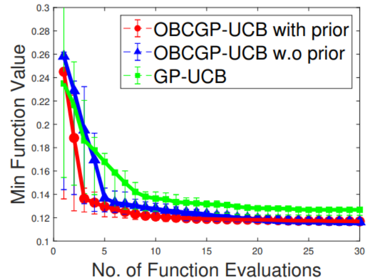


(e) Rosenbrock

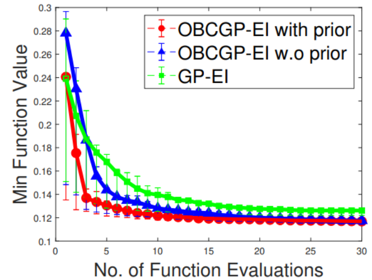


(f) Branin-20D

- ❖ GP-BO vs OBCGP-BO: SR comparison on hyper-parameter optimization for MLP



(a) MLP example: UCB



(b) MLP example: EI