MALIBO: Meta-learning for Likelihood-free Bayesian Optimization

Jiarong Pan[1 2] Stefan Falkner[1] Felix Berkenkamp[1] Joaquin Vanschoren[2]

[1]Bosch Center for Artificial Intelligence [2]Eindhoven University of Technology

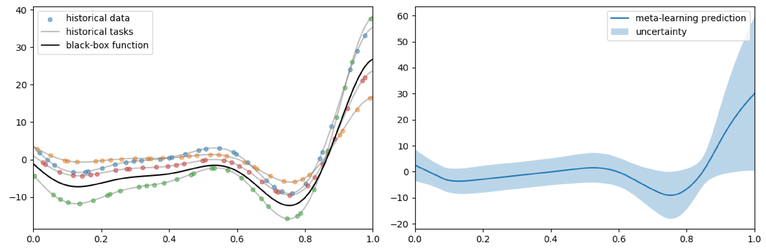
ICML International Conference On Machine Learning

Motivation and Setup

Bayesian optimization (BO) aims to optimize an expensive black-box function:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}\in\mathcal{X}} f(\mathbf{x}) \tag{1}$$

Meta-learning BO leverages information from past optimization experiences to accelerate the current optimization process.

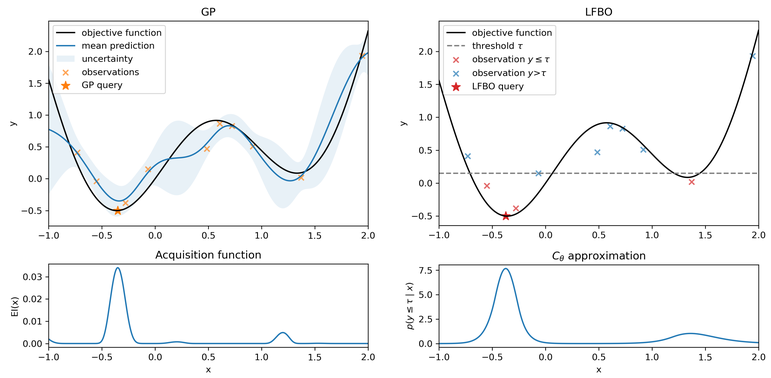


Contribution

- **Directly meta-learn the acquisition function** (optimization strategy) with robustness to heterogeneous scales and noises across tasks.
- Use a **probabilistic task adaptation** to account for task uncertainty.
- Propose a novel adaptation procedure with **residual prediction to ensure robust adaptation**.

Background: Likelihood-free Bayesian optimization (LFBO)

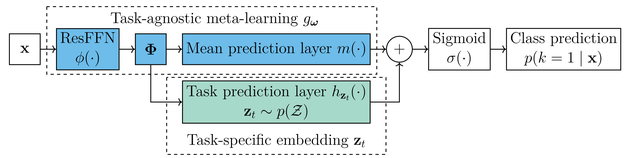
LFBO directly approximate the acquisition function by casting the approximation to a classification problem. The trained classifier $C_{\boldsymbol{\theta}}$ directly output the approximation.



$$\mathcal{L}^{\text{LFBO}}(\boldsymbol{\theta}; \mathcal{D}_N, \tau) = -\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}_N}\big[\max(\tau - y, 0)\ln C_{\boldsymbol{\theta}}(\mathbf{x}) + \ln(1 - C_{\boldsymbol{\theta}}(\mathbf{x}))\big]. \tag{2}$$

MALIBO

Meta-learning acquisition function



The meta-learning optimize the LFBO loss across tasks:

$$\mathcal{L}^{\text{meta}}(\boldsymbol{\omega}, \{\mathbf{z}_t\}_{t=1}^T) = \frac{1}{T}\sum_{t=1}^{T} \mathcal{L}^{\text{LFBO}}(\boldsymbol{\omega}, \mathbf{z}_t; \mathcal{D}^t, \tau) + \lambda\mathcal{R}(\{\mathbf{z}_t\}_{t=1}^T; p(\mathcal{Z})) \tag{3}$$

Directly optimizing $\mathcal{L}^{\text{LFBO}}$ leads to unreliable task adaptation. We regularize $\mathcal{Z}$ to follow a Gaussian prior with $\mathcal{R}$ for Bayesian task adaptation.

Probabilistic task adaptation

We need the posterior of task embedding $p(\mathbf{z} \mid \mathcal{D}_N)$ for prediction:

$$C(\mathbf{x}; \boldsymbol{\omega}, \mathcal{D}_N) = \int p(k = 1 \mid \boldsymbol{\omega}, \mathbf{z}) p(\mathbf{z} \mid \mathcal{D}_N)\,\mathrm{d}\mathbf{z} \tag{4}$$

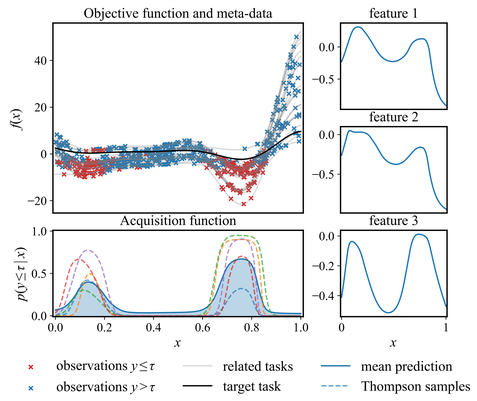
Obtaining the posterior is intractable due to $\sigma(\cdot)$ and the marginal. We use Laplace approximation to obtain $p(\mathbf{z} \mid \mathcal{D}_N) \simeq q(\mathbf{z}) = \mathcal{N}(\mathbf{z} \mid \mathbf{z}_{\text{MAP}}, \boldsymbol{\Sigma}_N)$:

$$\mathbf{z}_{\text{MAP}} = \arg\min_{\mathbf{z}} \frac{1}{2}\mathbf{z}^\mathsf{T}\mathbf{z} - \sum_{n=1}^{N}\Big(k_n(\tau - y)\ln\hat{k}_n + \ln(1 - \hat{k}_n)\Big) \tag{5}$$

$$\boldsymbol{\Sigma}_N^{-1} = \boldsymbol{\Sigma}_0^{-1} + \sum_{n=1}^{N}(k_n(\tau - y) + 1)\hat{k}_n(1 - \hat{k}_n)\boldsymbol{\Phi}_n\boldsymbol{\Phi}_n^\mathsf{T}, \tag{6}$$

The output is a sample of the approximated posterior.

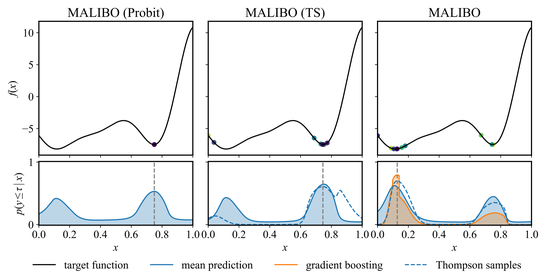
$$\hat{C}(\mathbf{x}) = \sigma\left(m(\phi(\mathbf{x})) + h_{\hat{\mathbf{z}}}(\phi(\mathbf{x}))\right), \quad \hat{\mathbf{z}} \sim q(\mathbf{z}) \tag{7}$$



Robust adaptation via residual prediction

To overcome the distribution shift between meta-data and the non-i.i.d. optimization data, we use a Gradient Boosting tree (GB) to model the residual between the model prediction and the black-box function.

$$C_{\text{GB}}(\mathbf{x}) = \sigma\left(m(\phi(\mathbf{x})) + h_{\hat{\mathbf{z}}}(\phi(\mathbf{x})) + \sum_{i=1}^{M} r_i(\mathbf{x})\right) \tag{8}$$



Experiments