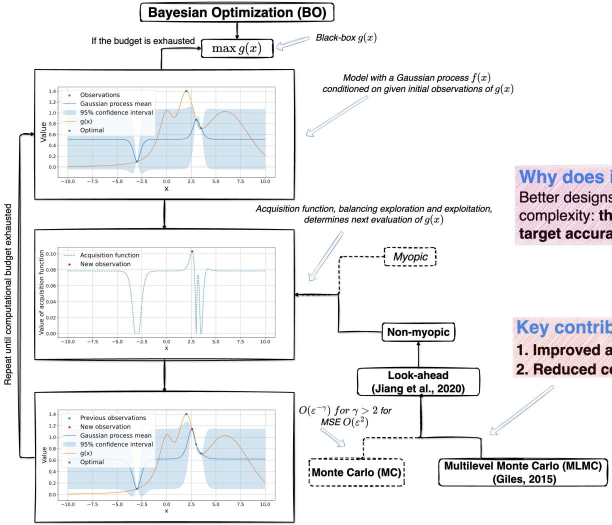
Accelerating Look-ahead in Bayesian Optimization: Multilevel Monte Carlo is All You Need

Shangda Yang[1], Vitaly Zankin[1], Maximilian Balandat[2], Stefan Scherer[2], Kevin Carlberg[2], Neil Walton[3,1], Kody J. H. Law[2,1]

1. University of Manchester 2. Meta 3. University of Durham

July 19, 2024

Overview



Why does it matter?

Better designs for a given budget, in terms of order of complexity: **the bigger the budget or higher the target accuracy, the more gain there is to be had.**

Key contributions:

1. **Improved asymptotic runtime $O(\varepsilon^{-2})$ for MSE $O(\varepsilon^2)$**
2. **Reduced cost of the whole BO**

Look-ahead acquisition function

Look-ahead construction:

$$\alpha_0(x;\mathcal{D}) := \mathbb{E}_{f(\cdot;\mathcal{D})}[r(f,x)] \approx \frac{1}{N}\sum_{i=1}^{N} r(f^i(x;\mathcal{D}))$$

$$\alpha_1(x;\mathcal{D}) := \mathbb{E}_{f(\cdot;\mathcal{D})}\left[r(f,x) + \max_{x_1} \mathbb{E}_{f(\cdot;\mathcal{D}_1(x))}\left[r(f,x_1)\right]\right]$$

$$\approx \frac{1}{N}\sum_{i=1}^{N}\left[r(f^i(x;\mathcal{D})) + \left(\max_{x_1^i} \frac{1}{M}\sum_{j=1}^{M} r(f^{ij}(x_1^i;\mathcal{D}_1^i(x)))\right)\right]$$

$$\vdots$$

where

- f is a Gaussian process given the current observation data $\mathcal{D}$
- $r(f,x)$ is a stage-wise reward characterizing the acquisition function
- $\mathcal{D}_1(x) = \mathcal{D} \cup \{(x, f(x;\mathcal{D}))\}$
- $\mathbb{E}_{f(\cdot;\mathcal{D})}$ denotes the expectations over the Gaussian process f given data $\mathcal{D}$.

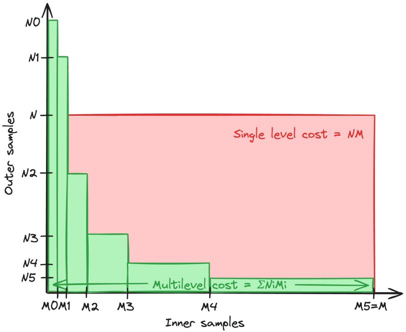
Complexity $O(\varepsilon^{-4}) \rightarrow O(\varepsilon^{-2})$ for MSE $O(\varepsilon^2)$ with MLMC

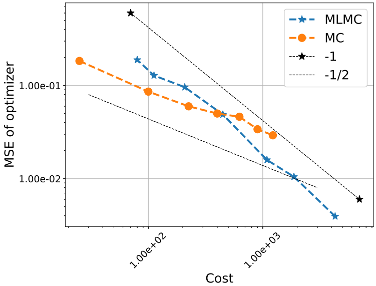
Multilevel Monte Carlo

The MLMC (Giles (2015)) approximation to $\mathbb{E}[\varphi]$ is

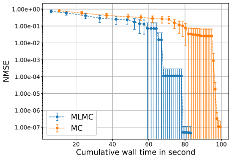
$$\mathbb{E}[\varphi] \approx \mathbb{E}[\varphi_L] = \sum_{l=0}^{L} \mathbb{E}[\varphi_l - \varphi_{l-1}] \approx \sum_{l=0}^{L} \frac{1}{N_l} \sum_{i=1}^{N_l} [\varphi_l^{(i)} - \varphi_{l-1}^{(i)}],$$

where $\varphi_{-1} = 0$ and $\varphi_0, \varphi_1, \varphi_2, ..., \varphi_L$ denotes the sequence of approximations with increasing accuracy and cost over levels l.

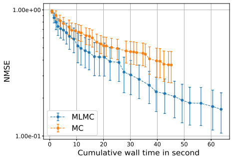




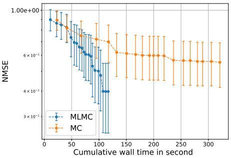
BO Results, MLMC & MC: NMSE vs time (↓)



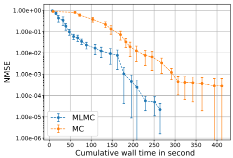
(a) 1D Toy Example (d=1)



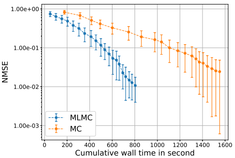
(b) Ackley (d=2)



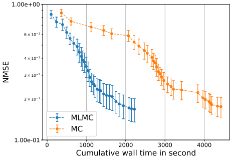
(c) DropWave (d=2)



(d) Branin (d=2)



(e) Hartmann6 (d=6)



(f) Cosine8 (d=8)

Figure: Convergence of the BO algorithm with respect to the cumulative wall time in seconds, with error bars (computed with 20 realizations). The Matérn kernel is applied. The initial BO run starts with $2 \times d$ observations.

Thank you for listening!

Scan me for the **Paper**

Scan me for the **Codes**