Bayesian Optimisation with Surrogate model

Heteroscedastic Evolutionary Bayesian Optimisation

Nadhir Hassen
nadhir.hassen@mila.quebec

MILA Quebec
Polytechnique Montreal

Table of contents

Intro

Common approaches

- Bayesian optimization addresses problems where the aim is to find the parameters $\hat{\mathbf{x}}$ that maximize a function $\mathbf{f}(\hat{\mathbf{x}})$ over some domain $\mathcal{X}$ consisting of finite lower and upper bounds on every variable

$$\hat{\mathbf{x}} = \underset{\mathbf{x} \in \mathcal{X}}{\operatorname{argmax}} \left[\mathbf{f}(\mathbf{x}) \right].$$

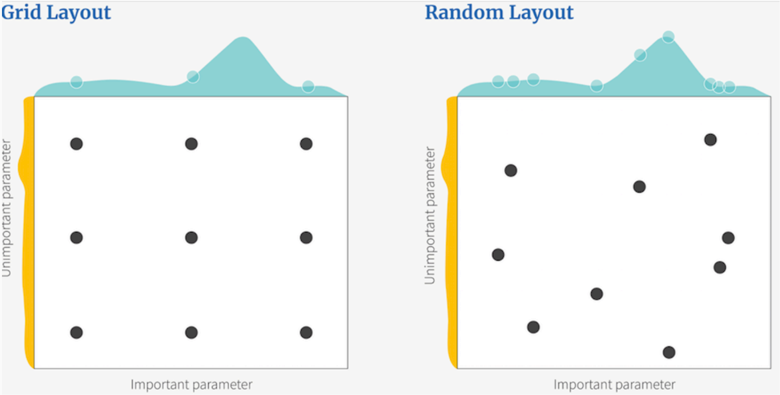
- The goal of Bayesian optimization is to find the maximum point on the function using the minimum number of function evaluations. More formally, we want to minimize the number of iterations t before we can guarantee that we find parameters $\hat{\mathbf{x}}$ such $\mathbf{f}(\hat{\mathbf{x}})$ is less than ϵ from the true maximum $\hat{\mathbf{f}}$

Strategies

Strategies

- Grid Search: Quantize each dimension of to form an input grid and then evaluate each point in the grid.
- Simple and easily parallelizable, but suffers from the curse of dimensionality
- The size of the grid grows exponentially in the number of dimensions.

Strategies


Grid Layout
Unimportant parameter
Important parameter
Random Layout
Unimportant parameter
Important parameter

More advanced Strategies

- Random Search: Specify probability distributions for each dimension of and then randomly sample from these distributions (Bergstra and Bengio, 2012).
- Sequential search strategies: Take into account previous measurements.
- Exploration: One idea is that we could explore areas where there are few samples so that we are less likely to miss the global maximum entirely.

Bayesian Optimisation: Surrogate Model

Bayesian Optimisation : BO deals with uncertainty

- BO: Sequential search framework
- Incorporates both exploration and exploitation and can be considerably more efficient than either grid search or random search.
- Goal: Build a probabilistic model of the underlying function that will know both $\mathbf{x}_1$ is a good place to sample and $\mathbf{x}_2$ has a high uncertainity.
- A Bayesian optimization algorithm has two main components:
- A probabilistic model of the function: Gaussian processes.
- Acquisition function: Posterior distribution over the function and is defined on the same domain.
- Determine how to favor exploration vs exploitation.

Acquisition Functions

Upper confidence bound : UCB

$$\text{UCB}[\mathbf{x}^*] = \mu(\mathbf{x}^*) + \beta^{1/2}\sigma(\mathbf{x}^*).$$

- Exploitation: This favors either regions where $\mu[\mathbf{x}^*]$ is large.
- Exploration : Regions where $\sigma[\mathbf{x}^*]$ is large.
- Positive parameter β trades off these two tendencies.

Probability of improvement: PI

$$\text{PI}[\mathbf{x}^*] = \int_{\mathsf{f}(\hat{\mathbf{x}})}^{\infty} \text{Norm}_{\mathsf{f}(\mathbf{x}^*)}(\mu(\mathbf{x}^*), \sigma(\mathbf{x}^*)) d\mathsf{f}(\mathbf{x}^*)$$

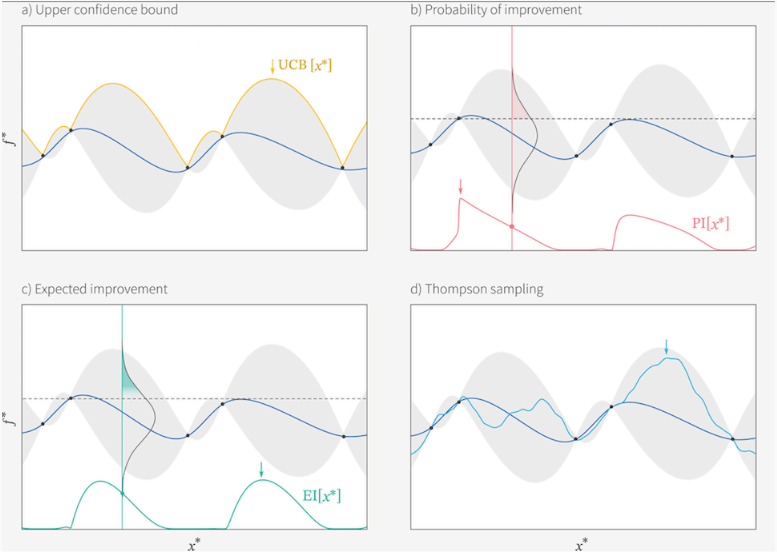
- Acquisition function retains the likelihood of the function at $\mathbf{x}^*$ higher than the current maximum $\mathbf{f}(\hat{\mathbf{x}})$.
- For each point $\mathbf{x}^*$, we integrate the part of the associated normal distribution that is above the current maximum

Expected improvement: EI

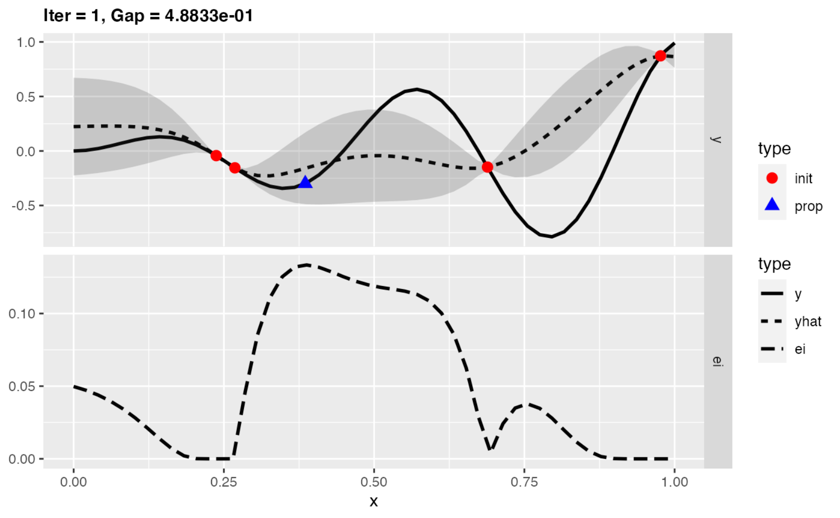
$$\mathrm{EI}[\mathbf{x}^*] = \int_{\mathrm{f}(\hat{\mathbf{x}})}^{\infty} (\mathrm{Relu}[f(\mathbf{x}^*)] - f(\hat{\mathbf{x}}))\mathrm{Norm}_{\mathrm{f}(\mathbf{x}^*))}(\mu(\mathbf{x}^*), \sigma(\mathbf{x}^*))d\mathrm{f}(\mathbf{x}^*).$$

- Disadvantage: does not take into account how much the improvement will be;
- We do not want to favor small improvements over larger ones.
- EI computes the expectation of the improvement of $\mathbf{f}(\mathbf{x}^\star) - f(\hat{\mathbf{x}})$

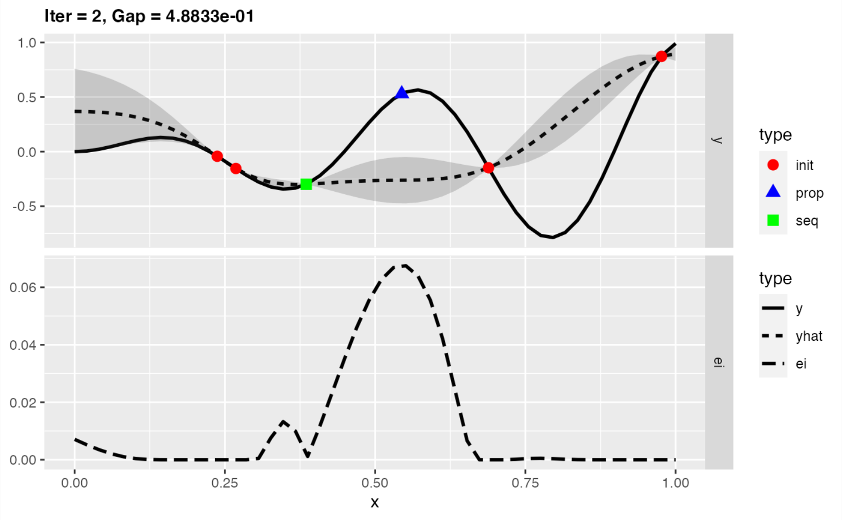
ALL Startegies



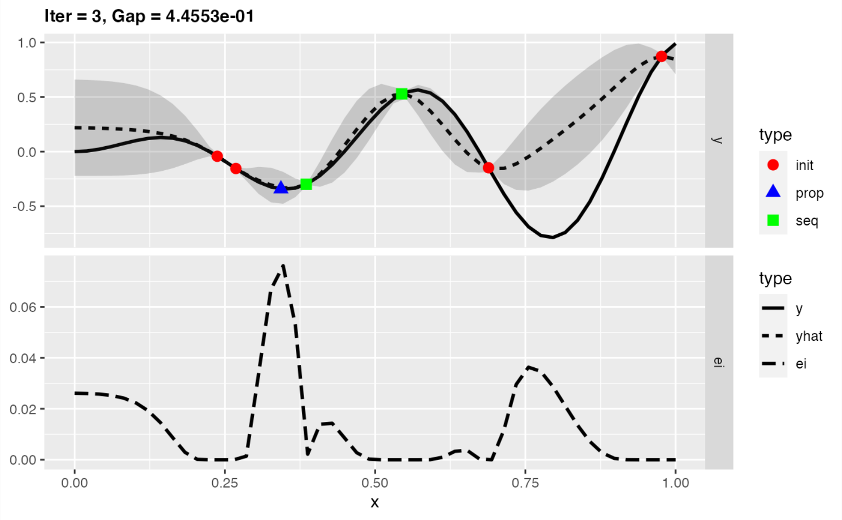
Expected Improvement in Action



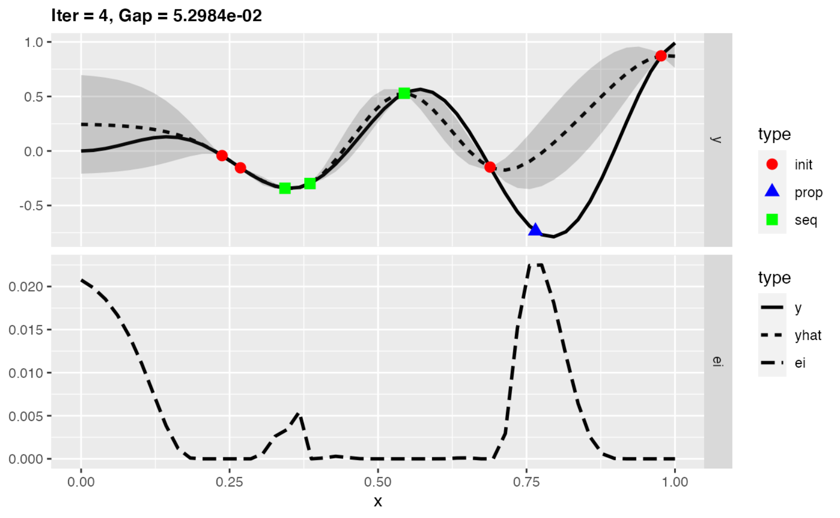
Expected Improvement in Action



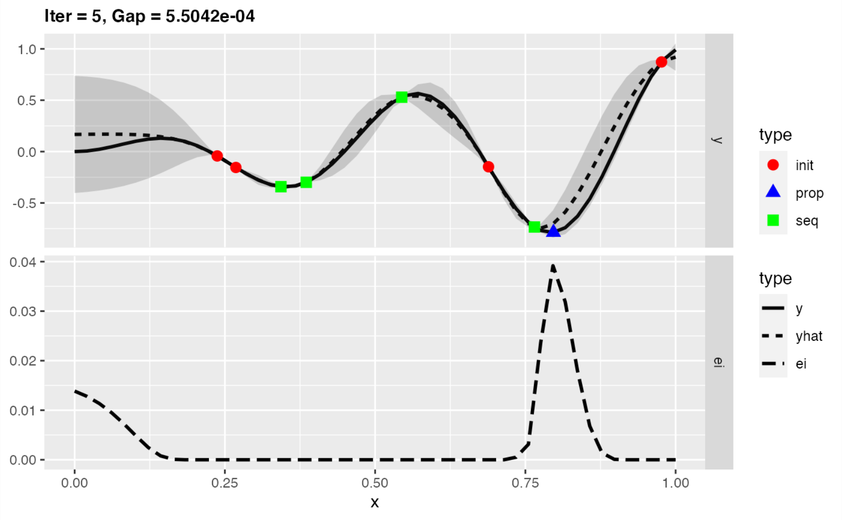
Expected Improvement in Action



Expected Improvement in Action



Expected Improvement in Action

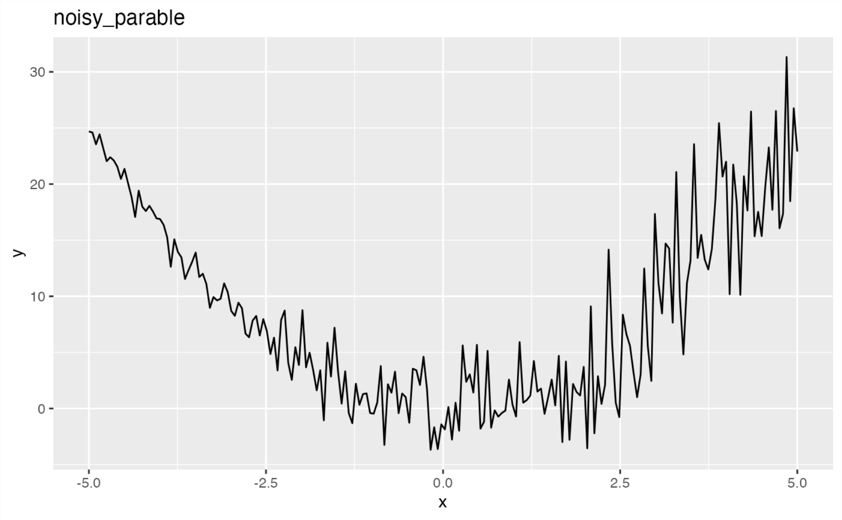


HEBO for Heteroscedasticity

HEBO: Heteroscedastic Evolutionary Bayesian Optimisation

- Step1: Fit a surgogate model [3]: Competition data can be noisy(non Gaussian) and potentially heteroscdeastic
- Step2: Maximize acquisition function:
- Step3: Evaluate the Black Box [2]

Noisy data



Step1 : Dealing with Heteroscdasticity

BBO challenge contributions

Key Idea: Power transformation $\Gamma(.)$ on the output and Warping $(\Psi_\theta(.))$ on the input, dependent on parameters θ.

- Wraped gaussian Processes [4]
- Power law transformation: Apply Box Cox [1] on the target label.
- For output objectives: Instead of the commonly used linear normalisation, Use power transformation to reduce heteroscedasticity.
- Evaluate the Black Box

Step2 : From single to multi-objective acquisition using evolutionary startegies

- Define a Gaussian Process and Kernel:

$$\boldsymbol{\kappa}^{HEBO}_{\theta_1,\theta_2}(\mathbf{x},\mathbf{x}') := \boldsymbol{\kappa}^{linear}_{\theta_1}(\mathbf{x},\mathbf{x}') + \boldsymbol{\kappa}^{Matern32}_{\theta_1}(\mathbf{x},\mathbf{x}').$$

- Construct a multi-objective function

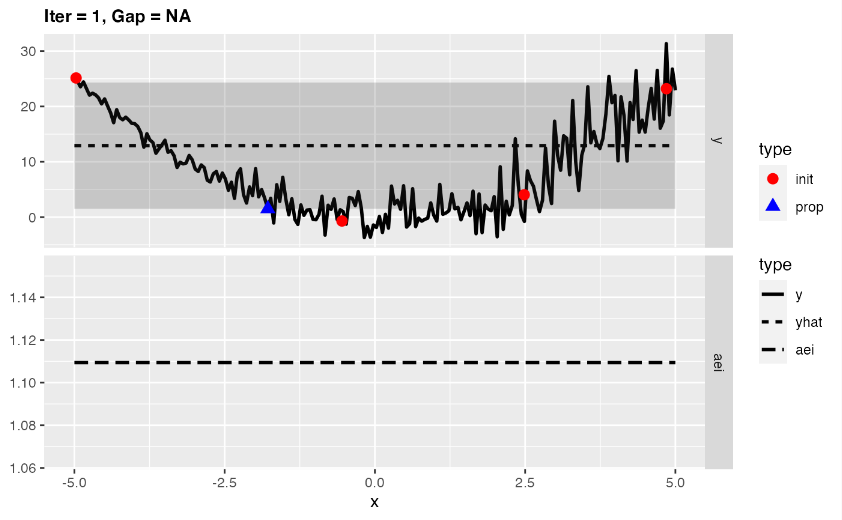
$$\min_{\mathbf{x}\in\mathcal{X}} \left[\alpha_{\text{q-UCB}}(\mathbf{x}|\mathcal{D}), -\alpha_{\text{q-EI}}(\mathbf{x}|\mathcal{D}), -\alpha_{\text{q-PI}}(\mathbf{x}|\mathcal{D})\right]^T$$

- Stochastic mean function: Locate where the likelihood noise are high gives better exploration

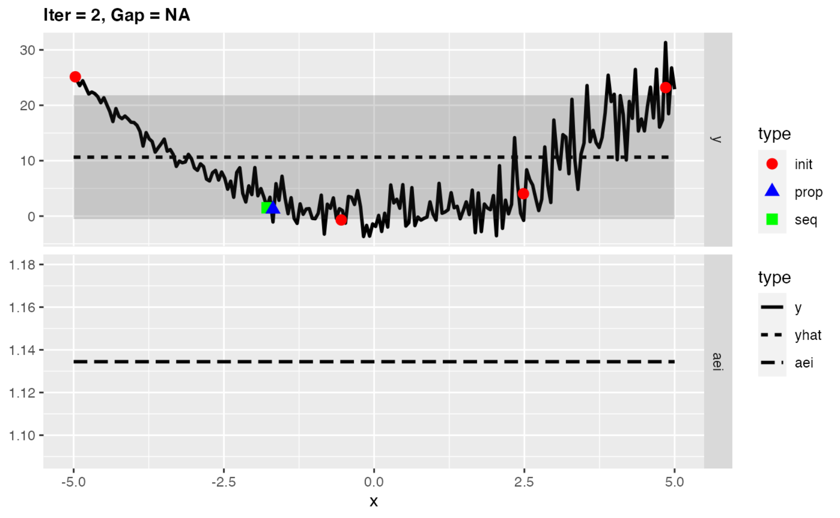
$$\boldsymbol{\mu}(\hat{\boldsymbol{x}}) = \boldsymbol{\mu}(\boldsymbol{x}) + \epsilon\boldsymbol{\sigma}(\mathbf{f}(\boldsymbol{x}))$$

- As the noise likelihood increases we get increasingly random search in the desired region (more exploration)

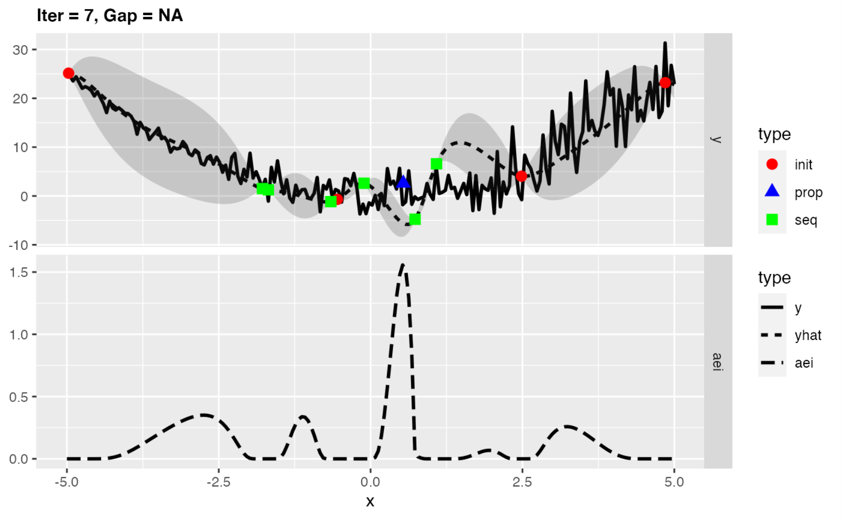
Step 3: Evaluate the Black BOX



Step 3: Evaluate the Black BOX



Step 3: Evaluate the Black BOX



MACE

Key Contribution of HEBO

- Input (init points) handles non-linear transformations
- Output (output of the GP) handles heteroscedasticity
- Multi-objectivity avoids conflicts.
- Enables a consensus among various acquisition functions through a Pareto-frontier.

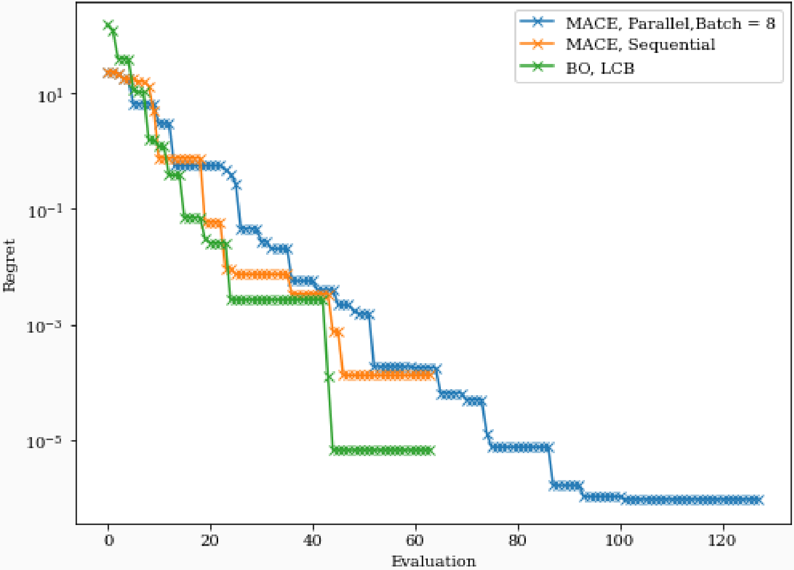
The multi-objective acquisition ensemble algorithm (MACE)

- MACE searches for a Pareto from across multiple acquisitions functions.
- The solution retains the higher score across all tested acquisition.
- For numerical stability we need to approximate $\alpha_{\text{q-EI}}(\mathbf{x}_i|\mathcal{D}_i)$ such that

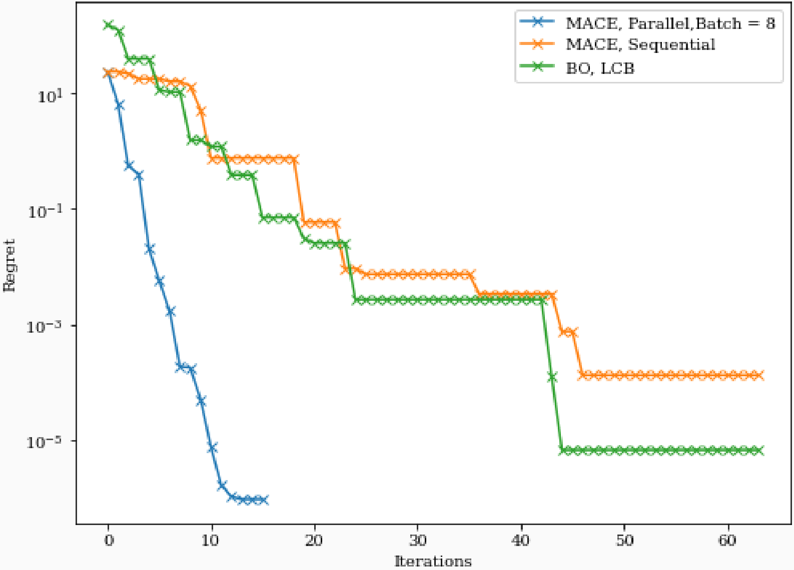
$$\lim_{z_i \to -\infty} \log \alpha_{\text{q-EI}}(\mathbf{x}_i|\mathcal{D}_i) = \log \sigma(\mathbf{x}_i; \theta) - \frac{1}{2}\mathbf{z}_i^2 - \frac{1}{2}\log(2\pi) - \log(z_i^2 - 1),$$

- where $z = \frac{\tau - \boldsymbol{\mu}(\mathbf{x}_i;\theta)}{\sigma(\mathbf{x}_i;\theta)}$ and τ is the best function value observed so far.
- This result enables parallel optimisation as the multi-objective optimisation.
- Returns multiple Pareto-optimal recommendations.

The multi-objective acquisition ensemble algorithm (MACE)



The multi-objective acquisition ensemble algorithm (MACE)



Final Results Comparaison

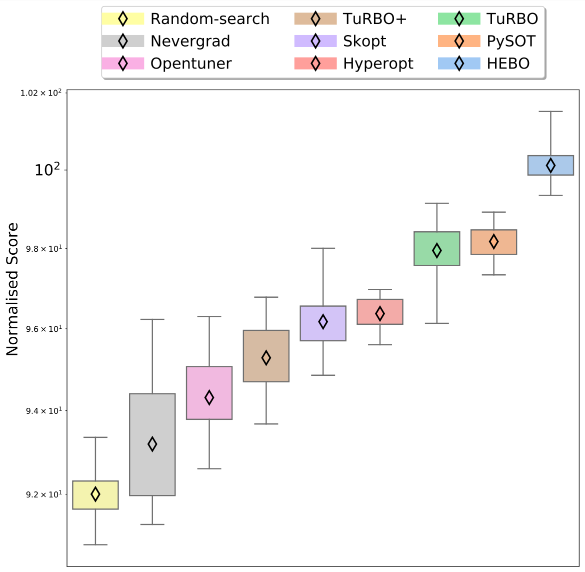


Figure 3: Source: HEBO-github

Questions?

References i

G. E. P. Box and D. R. Cox.
An analysis of transformations.
Journal of the Royal Statistical Society. Series B (Methodological, pages 211–252, 1964.

A. I. Cowen-Rivers, W. Lyu, R. Tutunov, Z. Wang, A. Grosnit, R. R. Griffiths, H. Jianye, J. Wang, and H. B. Ammar.
An empirical study of assumptions in bayesian optimisation, 2021.

A. I. Cowen-Rivers, W. Lyu, Z. Wang, R. Tutunov, H. Jianye, J. Wang, and H. B. Ammar.
Hebo: Heteroscedastic evolutionary bayesian optimisation.
arXiv preprint arXiv:2012.03826, 2020.
winning submission to the NeurIPS 2020 Black Box Optimisation Challenge.

E. Snelson, Z. Ghahramani, and C. Rasmussen.
Warped gaussian processes.
In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press, 2004.