




## NUBO: A Transparent Python Package for Bayesian Optimization

Mike Diessner   
Newcastle University

Kevin J. Wilson   
Newcastle University

Richard D. Whalley   
Queen's University Belfast

---

### Abstract

**NUBO**, short for Newcastle University Bayesian Optimisation, is a Bayesian optimization framework for the optimization of expensive-to-evaluate black-box functions, such as physical experiments and computer simulators. Bayesian optimization is a cost-efficient optimization strategy that uses surrogate modelling via Gaussian processes to represent an objective function and acquisition functions to guide the selection of candidate points to approximate the global optimum of the objective function. **NUBO** itself focuses on transparency and user experience to make Bayesian optimization easily accessible to researchers from all disciplines. Clean and understandable code, precise references, and thorough documentation ensure transparency, while user experience is ensured by a modular and flexible design, easy-to-write syntax, and careful selection of Bayesian optimization algorithms. **NUBO** allows users to tailor Bayesian optimization to their specific problem by writing the optimization loop themselves using the provided building blocks. It supports sequential single-point, parallel multi-point, and asynchronous optimization of bounded, constrained, and/or mixed (discrete and continuous) parameter input spaces. Only algorithms and methods that are extensively tested and validated to perform well are included in **NUBO**. This ensures that the package remains compact and does not overwhelm the user with an unnecessarily large number of options. The package is written in Python but does not require expert knowledge of Python to optimize your simulators and experiments. **NUBO** is distributed as open-source software under the BSD 3-Clause license.

*Keywords:* Bayesian optimization, black-box optimization, surrogate model, Gaussian process, Monte Carlo, design of experiments, Python.

---

## 1. Introduction

The optimization of expensive black-box functions is a common problem encountered by researchers in a wide range of disciplines, such as engineering, computing, and natural sciences.

Type	Modular	Sequential	Parallel	Asynchronous	Lines of code	Version
<b>NUBO</b>	Yes	Yes	Yes	Yes	1,322	1.2.1
<b>BoTorch</b>	Yes	Yes	Yes	Yes	38,419	0.8.4
<b>bayes_opt</b>	Yes	Yes	No	No	1,241	1.4.3
<b>SMAC3</b>	Yes	Yes	No	No	11,217	2.0.0
<b>pyGPGO</b>	Yes	Yes	No	No	2,029	0.5.0
<b>GPyOpt</b>	Yes	Yes	Yes	No	4,605	1.2.6
<b>Spearmint</b>	No	Yes	No	No	3,662	0.1

Table 1: Overview of available Bayesian optimization packages in Python. We compare whether individual packages have a modular design and support sequential single-point, parallel multi-point, and asynchronous optimization. We also list the number of lines of code of the core package (without comments, examples, tests, etc.) and the version number.

These functions are characterized by an unknown or not analytically solvable mathematical expression and high costs of evaluation. The principal way to gather information about a black-box function is to provide it with some inputs and observe its corresponding output. However, this process produces high costs, for example, material costs for physical experiments, computing costs for simulators, or time costs in general (Frazier 2018). Many optimization algorithms, such as Adam (Kingma and Ba 2014), L-BFGS-B (Zhu *et al.* 1997), and differential evolution (Storn and Price 1997), rely either on derivative information about the objective function or large numbers of function evaluations. Neither is typically feasible when working with an expensive black-box function, requiring us to search elsewhere for a cost-effective and sample-efficient alternative.

Bayesian optimization takes a surrogate model-based approach with the aim of optimising expensive black-box functions in a minimum number of function evaluations. Although the genesis of Bayesian optimization can be traced back to the middle of the 20th century (Kushner 1964; Žilinskas 1975; Moćkus 1975, 1989), it gained considerable popularity in the last two decades (Jones *et al.* 1998; Snoek *et al.* 2012; Shahriari *et al.* 2015; Frazier 2018). In recent years, it has been applied to simulators and experiments in various research areas. For example, Bayesian optimization was used in the field of computational fluid dynamics to maximize the drag reduction via the active control of blowing actuators (Diessner *et al.* 2022; O’Connor *et al.* 2023; Mahfoze *et al.* 2019), in chemical engineering for molecular design, drug discovery, molecular modelling, electrolyte design, and additive manufacturing (Wang and Dowling 2022), and in computer science to fine-tune hyper-parameters of machine learning models (Wu *et al.* 2019) and for architecture search of neural networks (White *et al.* 2021).

With **NUBO**, we provide an open-source implementation of Bayesian optimization aimed at researchers with expertise in disciplines other than statistics and computer science. To ensure that our target audience can understand and use Bayesian optimization to its full potential, **NUBO** focuses particularly on (a) transparency through clean and understandable code, precise references, and thorough documentation and (b) user experience through a modular and flexible design, easy syntax, and a careful selection of implemented algorithms. Various Python packages for Bayesian optimization exist as listed in Table 1. Most of them only support sequential single-point optimization, i.e., every point suggested by the algorithm has to be evaluated by the objective function before moving on to the next iteration. However, in many cases, parallelism can be exploited to speed up the optimization process. For exam-

ple, consider a simulator that can be run in parallel. Evaluating all points in parallel would save time as it would only take as long as evaluating a single point sequentially. **pyGPGO** (Jiménez and Ginebra 2017), **bayes\_opt**<sup>1</sup> (Nogueira 2014), **Spearmin** (Harvard University *et al.* 2014), and **SMAC3** (Lindauer *et al.* 2022) do not allow parallel multi-point optimization. Furthermore, **Spearmin** is not modular, resulting in less flexible implementations and giving the user less control when tailoring Bayesian optimization to unique research problems. To our knowledge, the closest available package to **NUBO** is **BoTorch** (Balandat *et al.* 2020) as it also supports parallel and asynchronous optimization through the use of Monte Carlo approximations of the acquisition functions. However, compared to the lightweight implementation of **NUBO**, **BoTorch** uses a very large code base that makes code comprehension difficult, as it often requires retracing various functions and objects through a large number of files. This can be quantified by the huge codebase represented in Table 1 as the total number of lines of code<sup>2</sup>: **NUBO** implements Bayesian optimization in only 1,322 lines of code over 20 files, while **BoTorch** uses 38,419 lines of code – roughly 29 times more than **NUBO** – and spreads them between 160 files. It also provides a large number of functions and methods that enforce decisions non-expert users do not have the knowledge and experience to make. **NUBO** lightens this burden of the user by limiting itself to the most important methods. Table 1 also includes **GPyOpt** (The **GPyOpt** authors 2016); however, it is no longer maintained and has recently been archived.

The number of code lines regards the underlying code bases of the packages and not the lines of code a user must write to apply Bayesian optimization. When talking about a transparent implementation, the former is a better proxy as it reflects the complexity of the whole package, that is, all functions and algorithms of the package. If a package has many thousands of lines – such as **BoTorch** – it is intuitive that it is more complex and thus harder to fully comprehend than a package with only a few hundred lines of code – such as **NUBO**. The number of code lines it takes to apply Bayesian optimization is less informative as it can easily be biased and distorted. Consider, for example, a very complex algorithm with many lines of code. It would be possible to wrap this algorithm into one function that can be called with one line of code. While this reduces the lines of code, it does not change the algorithm’s complexity. Thus, the comparison in this article focuses on the number of lines of the underlying code bases to give an idea of the size and complexity of the packages.

Although it is difficult to provide an exhaustive comparison of the relative efficiency of each of the packages, we have undertaken a limited comparison<sup>3</sup> of the following form. We compare **NUBO** to four of the packages mentioned above – **BoTorch**, **bayes\_opt**, **SMAC3** and **pyGPGO** – representing a reasonably wide range of complexity. All methods use Gaussian processes (introduced in Section 2.1) as the surrogate model and upper confidence bound (introduced in Section 2.2) as the acquisition function. Please see the replication materials published alongside this article for further details on the algorithms and the benchmarking. Two synthetic test functions from Surjanovic and Bingham (2013) were selected to benchmark the five packages. The first row of plots in Figure 1 compares the performance of sequential single-point optimization on A) the two-dimensional Levy function and B) the six-dimensional Hartmann function. The second row of plots compares the performance of

<sup>1</sup>The package is also known under the name **bayesian-optimization**.

<sup>2</sup>The total number of lines of code does not include comments, blank lines and files that are irrelevant to the actual algorithms, such as examples, tests and test functions.

<sup>3</sup>All comparisons were run on an Apple Mac mini with a M2 chip and 16 GB memory.

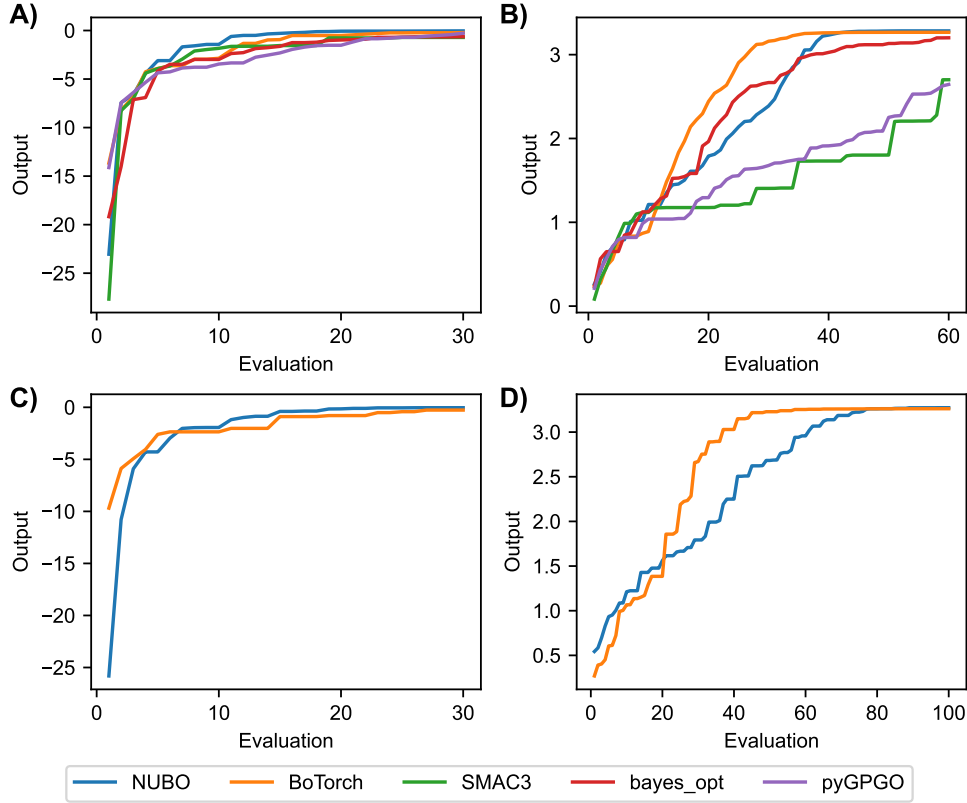


Figure 1: Comparison of different Python packages for Bayesian optimization. A) Sequential single-point optimization on the 2D Levy function; B) Sequential single-point optimization on the 6D Hartmann function; C) Parallel multi-point optimization with a batch size of four on the 2D Levy function; D) Parallel multi-point optimization with a batch size of four on the 6D Hartmann function.

parallel multi-point optimization with batches of four points on C) the two-dimensional Levy function and D) the six-dimensional Hartmann function. All functions are negated to transform them from their initial minimization problem into a maximization problem in line with the convention of Bayesian optimization. The plots provide the best observation at the current evaluation, averaging over ten replication runs. The results show that all packages converge towards the global optimum of 0.00 for the Levy and 3.32 for the Hartmann function. While **NUBO** requires more evaluations to find the global optimum for the Hartmann function (B) and D)) than the more complex packages such as **BoTorch**, it gets closest to the true optimum in all cases after all evaluations and shows low variance in these results (Table 2). These results show that the simplicity of **NUBO**'s implementation does not come at a cost in performance. **NUBO** can outperform packages with a similar level of complexity, such as **pyGPGO** and **bayes\_opt**, and compares well against more complex packages, such as **BoTorch** and **SMAC3**. This is not to say that **NUBO** is the superior package for any problem, but rather that **NUBO** performs competitively while focusing on a transparent and simple design. This makes **NUBO** a good candidate for the optimization of expensive black-box functions in the sciences – such as physical experiments and computer simulators – where transparency is vital.

	Sequential		Parallel	
	2D Levy	6D Hartmann	2D Levy	6D Hartmann
<b>NUBO</b>	−0.04 ( $\pm 0.06$ )	3.28 ( $\pm 0.06$ )	−0.04 ( $\pm 0.04$ )	3.27 ( $\pm 0.06$ )
<b>BoTorch</b>	−0.21 ( $\pm 0.20$ )	3.27 ( $\pm 0.07$ )	−0.27 ( $\pm 0.21$ )	3.26 ( $\pm 0.06$ )
<b>SMAC3</b>	−0.71 ( $\pm 0.58$ )	2.70 ( $\pm 0.38$ )	—	—
<b>bayes_opt</b>	−0.64 ( $\pm 0.74$ )	3.20 ( $\pm 0.13$ )	—	—
<b>pyGPGO</b>	−0.28 ( $\pm 0.31$ )	2.64 ( $\pm 1.05$ )	—	—

Table 2: Comparison of different Python packages for Bayesian optimization. The best observations averaged across the ten runs with corresponding standard errors are given for each package.

	Sequential		Parallel	
	2D Levy	6D Hartmann	2D Levy	6D Hartmann
<b>NUBO</b>	0.60s	1.88s	0.07s	2.20s
<b>BoTorch</b>	0.09s	0.22s	0.00s	0.19s
<b>SMAC3</b>	0.08s	0.25s	—	—
<b>bayes_opt</b>	0.14s	0.24s	—	—
<b>pyGPGO</b>	0.23s	0.65s	—	—

Table 3: Comparison of different Python packages for Bayesian optimization. The elapsed time per iteration averaged across the ten runs is given for each package.

However, the time **NUBO** requires to complete one iteration with a maximum of 2.20s for D) is, on average, higher than for the other packages (Table 3). While this might be important for some areas of optimization, it is negligible when it comes to the optimization of expensive black-box functions, as these functions are much more resource-intensive to evaluate. Thus, the small number of additional seconds that **NUBO** requires per iteration is insignificant compared to the resources required to conduct an experiment or a simulation.

Besides implementations in Python, there are also some implementations in other programming languages. For example, **rBayesianOptimization** (Yan 2024) and **ParBayesianOptimization** (Wilson 2022) implement basic Bayesian optimization algorithms for hyper-parameter tuning similar to **bayes\_opt** and **pyGPGO** in R. **ParBayesianOptimization** provides additional support for parallel optimization and follows Wilson *et al.* (2018).

The remainder of this paper is structured as follows. In Section 2, we introduce the Bayesian optimization algorithm, including Gaussian processes that form the surrogate model and acquisition functions that guide the optimization. The implementation of Bayesian optimization in **NUBO** is discussed in Section 3 before we illustrate how **NUBO** can be used to optimize expensive black-box functions through a non-trivial case study in Section 4. Finally, we draw conclusions and give an outlook on future work in Section 5.

## 2. Bayesian optimization

Bayesian optimization aims to solve the  $d$ -dimensional maximization problem

$$\mathbf{x}^* = \arg \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}), \quad (1)$$

where the input space is usually continuous and bounded by a hyper-rectangle  $\mathcal{X} \in [a, b]^d$  with  $a, b \in \mathbb{R}$ . The function  $f(\mathbf{x})$  is most commonly a derivative-free, expensive-to-evaluate black-box function that allows inputs  $\mathbf{x}_i$  to be queried and outputs  $y_i$  to be observed without gaining any further insights into the underlying system (Frazier 2018). We assume any noise  $\epsilon$  introduced when taking measurements to be independent and identically distributed Gaussian noise  $\epsilon \sim \mathcal{N}(0, \sigma^2)$  such that  $y_i = f(\mathbf{x}_i) + \epsilon$ . Hence, a set of  $n$  pairs of input data points and corresponding observations is defined as

$$\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$$

and we further define training inputs as the matrix  $\mathbf{X}_n = \{\mathbf{x}_i\}_{i=1}^n$  and their training outputs as the vector  $\mathbf{y}_n = \{y_i\}_{i=1}^n$ . Simulators and experiments in various disciplines can be formulated to fit this description, including but not limited to the examples given in the introduction.

Bayesian optimization (Frazier 2018; Gramacy 2020; Jones *et al.* 1998; Shahriari *et al.* 2015; Snoek *et al.* 2012) is a surrogate model-based optimization algorithm that aims to maximize the objective function  $f(\mathbf{x})$  in a minimum number of function evaluations. Typically, the objective function does not have a known or analytical solvable mathematical expression, and every function evaluation is expensive. Such problems require a cost-effective and sample-efficient optimization strategy. Bayesian optimization meets these criteria by representing the objective function through a surrogate model  $\mathcal{M}$ , often a Gaussian process. This representation can be used to find the input points to be evaluated sequentially by maximising a criterion specified through an acquisition function  $\alpha(\cdot)$ . A popular criterion is the upper confidence bound (UCB). This acquisition function can be classed as an optimistic acquisition function that considers the upper bound of the uncertainty around the surrogate model’s prediction to be true (Shahriari *et al.* 2015). Bayesian optimization is performed in a loop, where training data is used to fit the surrogate model before the next point suggested by the acquisition function is evaluated and added to the training data (see the Algorithm 1 below). The process then restarts and gathers more information about the objective function with each iteration. Bayesian optimization is run for as many iterations as the evaluation budget  $N$  allows, as shown in Algorithm 1, until a satisfactory solution is found or until a predefined stopping criterion is met.

Figure 2 illustrates how the Bayesian optimization algorithm works for an optimization loop that runs for eight iterations and starts with two initial training points. In this example, **NUBO** finds an approximation of the global optimum ( $x = 8$ ) for a simple 1-dimensional problem on iteration seven. The surrogate model uses the available observations to provide a prediction and associated uncertainty (here shown as 95% confidence intervals around the prediction). This is our best estimate of the underlying objective function. This estimate is then used in the acquisition function to evaluate which input value is likely to return a high output. Maximising the acquisition function provides the next candidate point to be observed from the objective function before it is added to the training data and the whole process is repeated. Figure 2 shows how the surrogate model converges to the true objective

**Algorithm 1** Bayesian optimization algorithm

---

**Require:** Evaluation budget  $N$ , number of initial points  $n_0$ , surrogate model  $\mathcal{M}$ , acquisition function  $\alpha$ .  
 Sample  $n_0$  initial training data points  $\mathbf{X}_0$  via a space-filling design (McKay *et al.* 1979) and gather observations  $\mathbf{y}_0$ .  
 Set  $n = 0$ .  
 Set  $\mathcal{D}_n = \{\mathbf{X}_0, \mathbf{y}_0\}$ .  
**while**  $n \leq N - n_0$  **do**  
   Fit surrogate model  $\mathcal{M}$  to training data  $\mathcal{D}_n$ .  
   Find  $\mathbf{x}_n^*$  that maximizes an acquisition criterion  $\alpha$  based on model  $\mathcal{M}$ .  
   Evaluate  $\mathbf{x}_n^*$  observing  $y_n^*$  and add to  $\mathcal{D}_n$ .  
   Increment  $n$ .  
**end while**  
**return** Point  $\mathbf{x}^*$  with highest observation  $y^*$ .

---

function with each iteration. The acquisition function covers the input space by exploring regions with high uncertainty and exploiting regions with a high prediction. This property, known as the exploration-exploitation trade-off, is a cornerstone of the acquisition functions provided in **NUBO**.

## 2.1. Gaussian processes

A popular choice for the surrogate model  $\mathcal{M}$  that acts as a representation of the objective function  $f(\mathbf{x})$  is a Gaussian process (Gramacy 2020; Rasmussen and Williams 2006), a flexible non-parametric regression model. A Gaussian process is a finite collection of random variables that has a joint Gaussian distribution and is defined by a prior mean function  $\mu_0(\mathbf{x}) : \mathcal{X} \mapsto \mathbb{R}$  and a prior covariance kernel  $\Sigma_0(\mathbf{x}, \mathbf{x}') : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$  resulting in the prior distribution

$$f(\mathbf{X}_n) \sim \mathcal{N}(m(\mathbf{X}_n), K(\mathbf{X}_n, \mathbf{X}_n)),$$

where  $m(\mathbf{X}_n) = \mu_0(\mathbf{X}_n)$  is the mean vector of length  $n$  over all training inputs and  $K(\mathbf{X}_n, \mathbf{X}_n) = \Sigma_0(\mathbf{X}_n, \mathbf{X}_n)$  is the  $n \times n$  covariance matrix between all training inputs.

Popular choices for the prior mean function  $\mu_0(\cdot)$  are the zero and constant mean functions given in Equations 2 and 3. For the prior covariance function  $\Sigma_0(\cdot, \cdot)$ , the squared exponential kernel, also called the radial basis function (RBF) kernel, and the Matérn  $\frac{5}{2}$  kernel are popular options. The covariance kernels in Equations 4 and 5 are based on the distance  $r = |\mathbf{x} - \mathbf{x}'|$  and have two parameters: the signal variance  $\sigma_f^2$ , sometimes also referred to as the output-scale, and the characteristic length-scale  $l$ . The former will scale the function with larger values resulting in a larger deviation from its mean, while the latter indicates for how long function values are correlated along the input axes, the smaller the length-scale  $l$  the shorter the correlation (Gramacy 2020; Rasmussen and Williams 2006).

$$\mu_{\text{zero}}(\mathbf{x}) = 0 \tag{2}$$

$$\mu_{\text{constant}}(\mathbf{x}) = c \tag{3}$$

$$\Sigma_{\text{RBF}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \exp\left(-\frac{r^2}{2l^2}\right) \tag{4}$$



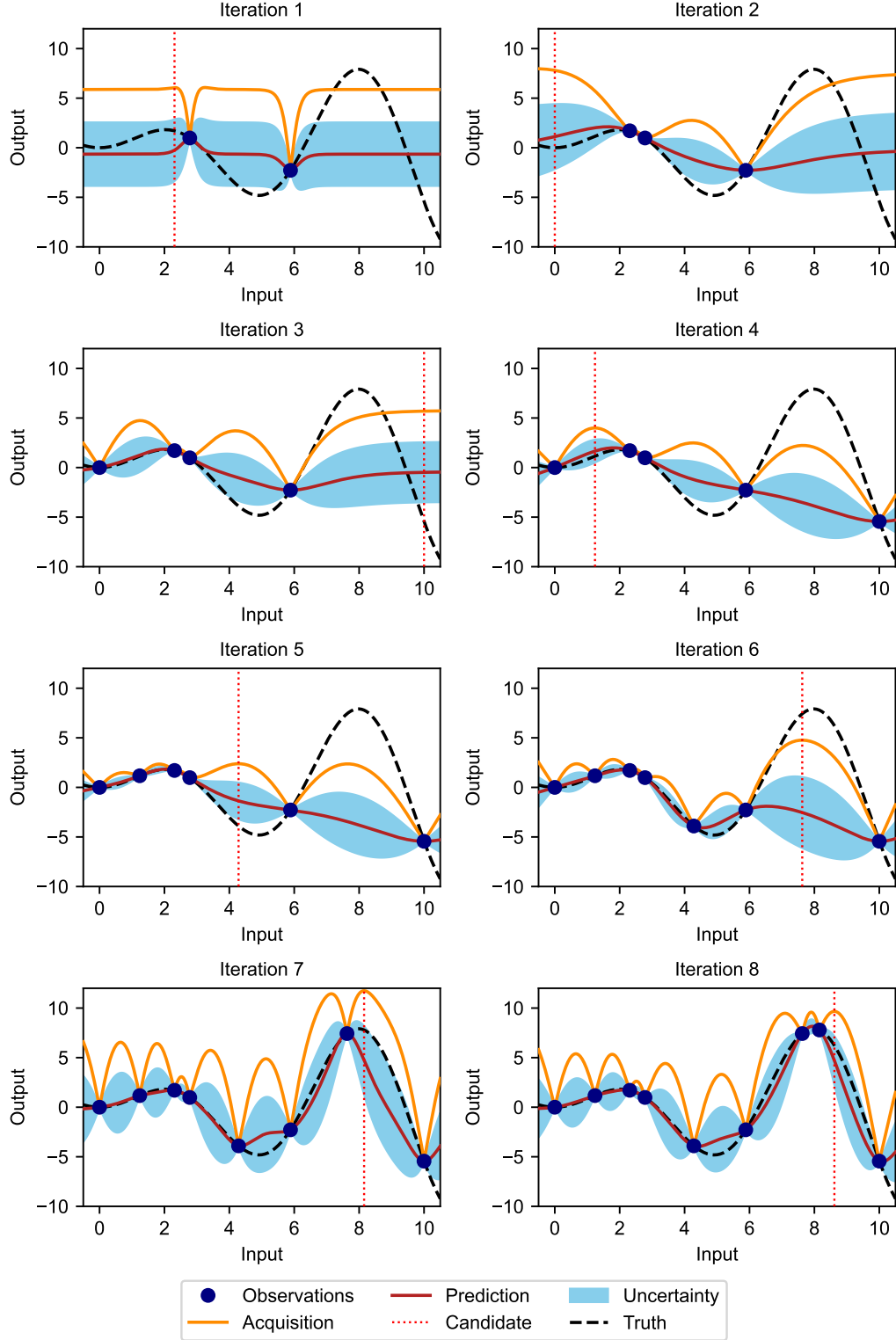


Figure 2: Bayesian optimization applied to a 1-dimensional function with one local and one global maximum. Upper confidence bound is used as the acquisition function. The input space is bounded by  $[0, 10]$ .



$$\Sigma_{\text{Matérn}}(\mathbf{x}, \mathbf{x}') = \sigma_f^2 \left( 1 + \frac{\sqrt{5}r}{l} + \frac{5r^2}{3l^2} \right) \exp \left( -\frac{\sqrt{5}r}{l} \right) \quad (5)$$

Covariance functions can be extended to include one characteristic length-scale  $l_d$  for each input dimension  $d$ . In this case, input dimensions with large length-scales are correlated for longer distances and are less relevant for changes in the prediction. This means that varying the values of the input dimension affects the prediction little. Input dimensions with small length-scales are correlated for shorter distances, and even small changes in the input values can affect the prediction significantly. Gaussian processes with covariance functions that include multiple length-scales are characterized by automatic relevance determination (ARD) of the input dimensions (Neal 1996). Here, the inverse of the length-scales can be interpreted as the relevance of the corresponding dimensions (Rasmussen and Williams 2006). The Gaussian process will estimate large length-scales for irrelevant dimensions, automatically assigning them less importance.

The posterior distribution for  $n_*$  test points  $\mathbf{X}_*$  can be computed as the multivariate Gaussian distribution conditional on training data  $\mathcal{D}_n$

$$f(\mathbf{X}_*) \mid \mathcal{D}_n, \mathbf{X}_* \sim \mathcal{N}(\mu_n(\mathbf{X}_*), \sigma_n^2(\mathbf{X}_*))$$

$$\mu_n(\mathbf{X}_*) = K(\mathbf{X}_*, \mathbf{X}_n) \left[ K(\mathbf{X}_n, \mathbf{X}_n) + \sigma_y^2 I \right]^{-1} (\mathbf{y} - m(\mathbf{X}_n)) + m(\mathbf{X}_*)$$

$$\sigma_n^2(\mathbf{X}_*) = K(\mathbf{X}_*, \mathbf{X}_*) - K(\mathbf{X}_*, \mathbf{X}_n) \left[ K(\mathbf{X}_n, \mathbf{X}_n) + \sigma_y^2 I \right]^{-1} K(\mathbf{X}_n, \mathbf{X}_*),$$

where  $m(\mathbf{X}_*)$  is the mean vector of length  $n_*$  over all test inputs,  $K(\mathbf{X}_*, \mathbf{X}_n)$  is the  $n_* \times n$  covariance matrix,  $K(\mathbf{X}_n, \mathbf{X}_*)$  is the  $n \times n_*$  covariance matrix,  $K(\mathbf{X}_*, \mathbf{X}_*)$  is the  $n_* \times n_*$  covariance matrix between training inputs  $\mathbf{X}_n$  and test inputs  $\mathbf{X}_*$  respectively, and  $\sigma_y^2$  is the noise variance of the Gaussian process.

The hyper-parameters  $\theta$  of the Gaussian process, for example, the constant  $c$  in the mean function, the signal variance  $\sigma_f^2$  and characteristic length-scales  $\mathbf{l}$  in the covariance kernel, and the noise variance  $\sigma_y^2$ , can be estimated by maximising the log-marginal likelihood in Equation 6 via maximum likelihood estimation (MLE, Rasmussen and Williams 2006).

$$\begin{aligned} \log P(\mathbf{y}_n \mid \mathbf{X}_n) = & -\frac{1}{2} (\mathbf{y}_n - m(\mathbf{X}_n))^\top [K(\mathbf{X}_n, \mathbf{X}_n) + \sigma_y^2 I]^{-1} (\mathbf{y}_n - m(\mathbf{X}_n)) \\ & - \frac{1}{2} \log |K(\mathbf{X}_n, \mathbf{X}_n) + \sigma_y^2 I| - \frac{n}{2} \log 2\pi \end{aligned} \quad (6)$$

## 2.2. Acquisition functions

Acquisition functions use the posterior distribution of the Gaussian process to compute a criterion that assesses if a test point is a good potential candidate point to evaluate via the objective function  $f(\mathbf{x})$ . Thus, maximising the acquisition function suggests the test point that, based on the current training data  $\mathcal{D}_n$ , has the highest potential and information gain to get closer to the global optimum while exploring the input space. To do this, an acquisition function  $\alpha(\cdot)$  balances exploration and exploitation. The former is characterized by areas with no or only a few observed data points where the uncertainty of the Gaussian process is high, and the latter by areas where the posterior mean of the Gaussian process is large.

This exploration-exploitation trade-off ensures that Bayesian optimization does not converge to the first (potentially local) maximum it encounters but efficiently explores the full input space.

### *Analytical acquisition functions*

**NUBO** supports two of the most popular acquisition functions whose performance has been demonstrated in both theoretical and empirical research. Expected improvement (EI, Jones *et al.* 1998) selects points with the biggest potential to improve on the current best observation, while upper confidence bound (UCB, Srinivas *et al.* 2010) takes an optimistic view of the posterior uncertainty and assumes it to be true to a user-defined level. Expected improvement (EI) is defined as

$$\alpha_{\text{EI}}(\mathbf{X}_*) = \left( \mu_n(\mathbf{X}_*) - y^{\text{best}} \right) \Phi(z) + \sigma_n(\mathbf{X}_*) \phi(z), \quad (7)$$

where  $z = \frac{\mu_n(\mathbf{X}_*) - y^{\text{best}}}{\sigma_n(\mathbf{X}_*)}$ ,  $\mu_n(\cdot)$  and  $\sigma_n(\cdot)$  are the mean and the standard deviation of the posterior distribution of the Gaussian process,  $y^{\text{best}}$  is the current best observation, and  $\Phi(\cdot)$  and  $\phi(\cdot)$  are the cumulative distribution function and probability density function of the standard normal distribution  $\mathcal{N}(0, 1)$ .

The upper confidence bound (UCB) acquisition function can be computed as

$$\alpha_{\text{UCB}}(\mathbf{X}_*) = \mu_n(\mathbf{X}_*) + \sqrt{\beta} \sigma_n(\mathbf{X}_*), \quad (8)$$

where  $\beta$  is a predefined trade-off parameter, and  $\mu_n(\cdot)$  and  $\sigma_n(\cdot)$  are the mean and the standard deviation of the posterior distribution of the Gaussian process. For guidance on the choice of  $\beta$ , consult the theoretical properties in Srinivas *et al.* (2010) or empirical conclusions in Diessner *et al.* (2022).

Both of these acquisition functions can be maximized with a deterministic optimizer, such as L-BFGS-B (Zhu *et al.* 1997) for bounded unconstrained problems or SLSQP (Kraft 1994) for bounded constrained problems. However, the use of analytical acquisition functions is restricted to sequential single-point problems for which every point suggested by Bayesian optimization is observed via the objective function  $f(\mathbf{x})$  immediately before the optimization loop is repeated.

### *Monte Carlo acquisition functions*

For parallel multi-point batches or asynchronous optimization, the analytical acquisition functions are, in general, intractable. To use Bayesian optimization in these cases, **NUBO** supports the approximation of the analytical acquisition function through Monte Carlo sampling (Snoek *et al.* 2012; Wilson *et al.* 2018).

The idea is to draw a large number of samples directly from the posterior distribution and then approximate the acquisition functions by averaging these Monte Carlo samples. This method is made viable by the reparameterization of the acquisition functions and the computation of samples from the posterior distribution via base samples randomly drawn from a standard normal distribution  $z \sim \mathcal{N}(0, 1)$ . Thus, the analytical acquisition functions from in Equations 7 and 8 can be approximated as

$$\alpha_{\text{EI}}^{\text{MC}}(\mathbf{X}_*) = \max \left( \text{ReLU}(\mu_n(\mathbf{X}_*) + \mathbf{L}z - y^{\text{best}}) \right)$$

$$\alpha_{\text{UCB}}^{\text{MC}}(\mathbf{X}_*) = \max \left( \mu_n(\mathbf{X}_*) + \sqrt{\frac{\beta\pi}{2}} |\mathbf{L}\mathbf{z}| \right),$$

where  $\mu_n(\cdot)$  is the mean of the posterior distribution of the Gaussian process,  $\mathbf{L}$  is the lower triangular matrix of the Cholesky decomposition of the covariance matrix  $\mathbf{L}\mathbf{L}^\top = \mathbf{K}(\mathbf{X}_n, \mathbf{X}_n)$ ,  $\mathbf{z}$  are samples from the multivariate standard normal distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I})$ ,  $y^{\text{best}}$  is the current best observation,  $\beta$  is the user-defined trade-off parameter, and  $\text{ReLU}(\cdot)$  is the rectified linear unit function that zeros all values below zero and leaves the rest unchanged.

Due to the randomness in the Monte Carlo samples, these acquisition functions can only be optimized by stochastic optimizers, such as Adam (Kingma and Ba 2014). However, there is some empirical evidence that fixing the base samples for individual Bayesian optimization loops does not affect the performance negatively (Balandat *et al.* 2020). This method would allow deterministic optimizers, such as L-BFGS-B (Zhu *et al.* 1997) and SLSQP (Kraft 1994), to be used but could potentially introduce bias due to sampling randomness.

Two optimization strategies for multi-point batches are proposed in the literature (Wilson *et al.* 2018): The first is a joint optimization approach, where the acquisition functions are optimized over all points of the batch simultaneously. The second option is a greedy sequential approach where one point after another is selected, holding all previous points fixed until the batch is full. Empirical evidence shows that both methods approximate the acquisition successfully. However, the greedy approach seems to have a slight edge over the joint strategy for some examples (Wilson *et al.* 2018). It is also faster to compute for larger batches.

Asynchronous optimization (Snoek *et al.* 2012) leverages the same property as sequential greedy optimization: the pending points that have not yet been evaluated can be added to the test points but are treated as fixed. In this way, they affect the joint multivariate normal distribution but are not considered directly in the optimization. Asynchronous optimization is particularly beneficial for objective functions for which the evaluation time varies. In these cases, the optimization can be continued while some points are still being evaluated.

### 3. NUBO

**NUBO** is a Bayesian optimization package in Python that focuses on transparency and user experience to make Bayesian optimization accessible to researchers from a wide range of disciplines whose area of expertise is not necessarily statistics or computer science. With this overall goal in mind, **NUBO** ensures transparency by implementing clean and comprehensible code, precise references and thorough documentation within this research article and on our website at [www.nubopy.com](http://www.nubopy.com). We avoid implementations of overly complex and convoluted functions and objects that require the retracing of individual elements through multiple files to be fully understood. We prioritize user experience defined by a modular and flexible design that can be intuitively tailored to unique problems, easy-to-read and write syntax, and a careful selection of Bayesian optimization algorithms. The latter is important as we try not to overwhelm the user with a larger number of options but rather focus on what is essential to optimize computer simulators and physical experiments successfully.

To create a powerful package with good longevity, it is important to start with a strong foundation. **NUBO** is built upon the **Torch**<sup>4</sup> ecosystem (Paszke *et al.* 2019) that provides

<sup>4</sup>See <https://pytorch.org/> for documentation and <https://pytorch.org/project/torch/> for package information on PyPI.

a strong scientific computation framework for working with tensors, a selection of powerful optimization algorithms, such as `torch.Adam` (Kingma and Ba 2014), automatic differentiation capabilities to compute gradients of acquisition functions via `torch.autograd`, and GPU acceleration. Furthermore, **GPyTorch**<sup>5</sup> (Gardner *et al.* 2018), the package we use to implement Gaussian processes for our surrogate modelling, is also based in **Torch** and combines seamlessly with **NUBO**. We borrow the L-BFGS-B (Zhu *et al.* 1997) and SLSQP (Kraft 1994) optimization algorithms from **SciPy**<sup>6</sup> (Virtanen *et al.* 2020) for the deterministic optimization of the acquisition functions and use **NumPy**<sup>7</sup> (Harris *et al.* 2020) to make data suitable for these optimizers.

**NUBO** and all its required dependencies can be installed from the Python Package Index (PyPI, Python Software Foundation 2003) with the packet installer **pip** (The **pip** Developers 2023) via the terminal:

```
pip install nubopy
```

### 3.1. Gaussian processes

**NUBO** uses the **GPyTorch** (Gardner *et al.* 2018) package to implement Gaussian processes for surrogate modelling. While **GPyTorch** allows the definition of many different Gaussian processes through its various mean functions, covariance kernels, and methods for hyper-parameter estimation, we provide a predefined Gaussian process in the `nubo.models` module that follows the work of Snoek *et al.* (2012). The `GaussianProcess` is specified by a constant mean function and the Matérn  $\frac{5}{2}$  ARD kernel that, due to its flexibility, is well suited for practical optimization as it can represent a wide variety of real-world objective functions. The code below implements a Gaussian process and estimates its hyper-parameters from some training inputs `x_train` and training outputs `y_train` by maximising the log-marginal likelihood in Equation 6 with the `fit_gp` function. The hyper-parameters include the constant in the mean function, the output-scale and length-scales in the covariance kernel, and the noise in the Gaussian likelihood. The training inputs and training outputs are specified as a `torch.Tensor` of size  $n \times d$  and length  $n$ , respectively, where  $n$  is the number of points and  $d$  is the number of input dimensions. Calling the function `fit_gp` results in a trained Gaussian process that can subsequently be used for Bayesian optimization.

```
>>> from nubopy.models import GaussianProcess, fit_gp
>>> from gpytorch.likelihoods import GaussianLikelihood
>>> likelihood = GaussianLikelihood()
>>> gp = GaussianProcess(x_train, y_train, likelihood = likelihood)
>>> fit_gp(x_train, y_train, gp = gp, likelihood = likelihood)
```

While Gaussian processes are capable of estimating noise, for example, observational noise occurring when taking measurements from the data, we might prefer specifying the noise

<sup>5</sup>See <https://gpytorch.ai/> for documentation and <https://pypi.org/project/gpytorch/> for package information on PyPI.

<sup>6</sup>See <https://scipy.org/> for documentation and <https://pypi.org/project/scipy/> for package information on PyPI.

<sup>7</sup>See <https://numpy.org/> for documentation and <https://pypi.org/project/numpy/> for package information on PyPI.

explicitly if it is known. In these cases, we can exchange the `GaussianLikelihood` for the `FixedNoiseGaussianLikelihood` and specify the noise for each training point. The `FixedNoiseGaussianLikelihood` allows us to decide if any additional noise should be estimated by setting the `learn_additional_noise` attribute to `True` or `False`. The snippet below fixes the observational noise of each training point at 2.5% and estimates any additional noise.

```
>>> from nubo.models import GaussianProcess, fit_gp
>>> from gpytorch.likelihoods import FixedNoiseGaussianLikelihood
>>> noise = torch.ones(x_train.size(0)) * 0.025
>>> likelihood = FixedNoiseGaussianLikelihood(noise = noise,
...                                           learn_additional_noise = True)
>>> gp = GaussianProcess(x_train, y_train, likelihood = likelihood)
>>> fit_gp(x_train, y_train, gp = gp, likelihood = likelihood)
```

### 3.2. Bayesian optimization

Before describing the individual optimization options in detail, we want to illustrate **NUBO**'s user experience, that is, its easy-to-read and write syntax, flexibility, and modularity, on a simple Bayesian optimization step that can be further divided into four substeps.

First, we define the input space. Here, we want to optimize a six-dimensional objective function that is bounded by the hyper-rectangle  $[0, 1]^6$  specified as `bounds`, a  $2 \times 6$  `torch.Tensor`, where the first row provides the lower bounds and the second row the upper bounds for all six input dimensions. Second, we load the training inputs `x_train` and the training outputs `y_train`. This training data can be selected manually or generated by using a space-filling design, such as Latin hypercube sampling introduced in Section 3.3. Third, we define and train the Gaussian process implemented in **NUBO** as discussed in Section 3.1, or set up a custom Gaussian process with **GPYTORCH**. Fourth, we specify an acquisition function that takes the fitted Gaussian process as an argument and chooses an optimization method. In this case, we use the upper confidence bound introduced in Equation 8 and optimize it with the L-BFGS-B algorithm (Zhu *et al.* 1997) using the `single` function.

```
>>> import torch
>>> from nubo.acquisition import UpperConfidenceBound
>>> from nubo.models import GaussianProcess, fit_gp
>>> from nubo.optimisation import single
>>> from gpytorch.likelihoods import GaussianLikelihood
>>> bounds = torch.tensor([[0., 0., 0., 0., 0., 0.],
...                        [1., 1., 1., 1., 1., 1.]])
>>> x_train = # load inputs as torch.Tensor
>>> y_train = # load outputs as torch.Tensor
>>> likelihood = GaussianLikelihood()
>>> gp = GaussianProcess(x_train, y_train, likelihood = likelihood)
>>> fit_gp(x_train, y_train, gp = gp, likelihood = likelihood)
>>> acq = UpperConfidenceBound(gp = gp, beta = 4)
>>> x_new, _ = single(func = acq, method = "L-BFGS-B", bounds = bounds)
```

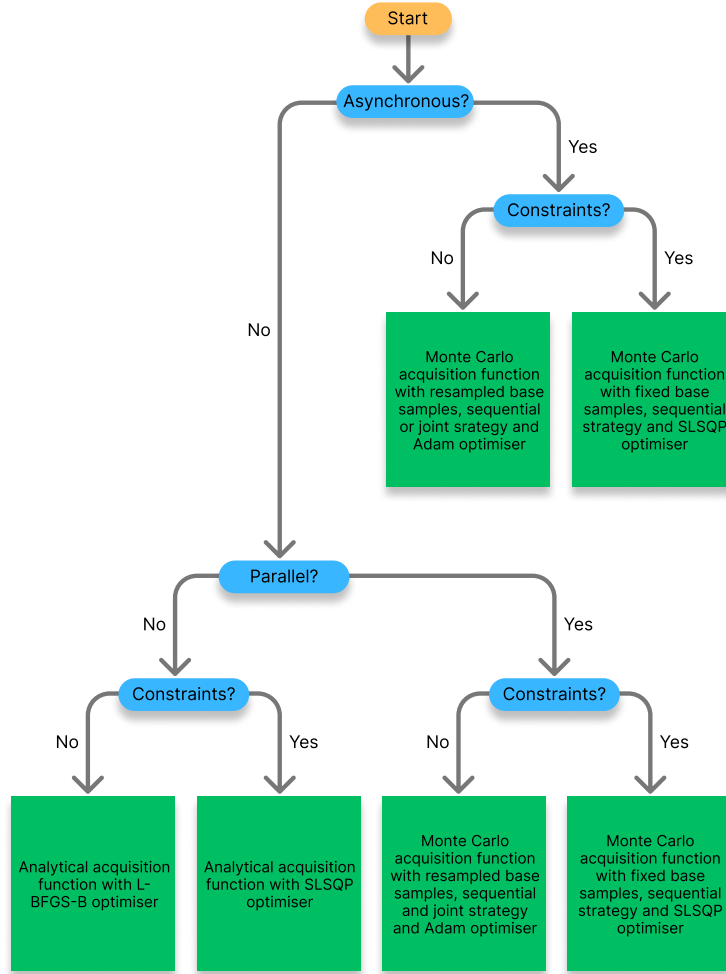


Figure 3: **NUBO** flowchart. Overview of the recommended algorithms for specific problems. Start in yellow, decisions in blue, and recommended algorithm in green.

**NUBO** is very flexible and allows the user to swap out individual elements for other options. For example, we can substitute the `UpperConfidenceBound` acquisition function or the `single` optimization strategy without changing any of the other lines of code. This makes it easy and fast to tailor Bayesian optimization to specific problems.

The remainder of this section introduces **NUBO**'s optimization strategies. Figure 3.2 shows a flowchart that helps users decide on the right acquisition function and optimizer for their specific problem.

### *Sequential single-point optimization*

In **NUBO** we differentiate between two optimization strategies: single-point and multi-point optimization. When using the single-point strategy via the `single` function, **NUBO** uses the analytical acquisition functions discussed in Section 2.2 to find the next point to be evaluated by the objective function. The corresponding observation must be gathered before the next iteration of the optimization loop can begin.

The code below shows how the analytical expected improvement (EI) and the analytical upper confidence bound (UCB) can be specified with **NUBO**. The former takes the best training output to date as the argument `y_best`, while the latter accepts the trade-off hyperparameter  $\beta$  as the `beta` argument. For bounded optimization problems with analytical acquisition functions, the optimization method of the `single` function should be set to `method = "L-BFGS-B"` and the arguments `num_starts` (default 10) and `num_samples` (default 100) can be set to enable multi-start optimization, where the selected optimization algorithm is run multiple times and each start is initialized at the best points from a large number of points sampled from a Latin hypercube. This reduces the risk of getting stuck in a local optimum. Section 3.3 introduces Latin hypercube sampling in more detail. The `single` function only returns the best start and its acquisition value. Sequential single-point optimization can be paired with constrained and mixed optimization, which are all detailed in this section.

```
>>> from nubo.acquisition import ExpectedImprovement, UpperConfidenceBound
>>> from nubo.optimisation import single
>>> acq = ExpectedImprovement(gp = gp, y_best = torch.max(y_train))
>>> acq = UpperConfidenceBound(gp = gp, beta = 4)
>>> x_new, _ = single(func = acq, method = "L-BFGS-B", bounds = bounds,
...                   num_starts = 5, num_samples = 50)
```

### *Parallel multi-point optimization*

The second optimization strategy is multi-point optimization. This strategy uses the Monte Carlo acquisition functions outlined in Section 2.2 to find multiple points, also called batches, in each iteration of the Bayesian optimization loop. This strategy is particularly beneficial for objective functions that support parallel evaluations as points can be queried simultaneously, speeding up the optimization process.

**NUBO** uses the Monte Carlo versions of expected improvement `MCEExpectedImprovement` and upper confidence bound `MCUpperConfidenceBound` in unison with either the `multi_joint` or `multi_sequential` function to compute batches. The two different options for the multi-point optimization strategy are discussed in Section 2.2. In addition to the arguments of the analytical acquisition functions, both Monte Carlo acquisition functions accept the number of Monte Carlo samples to be used to approximate the acquisition function as the `samples` argument (default 512). For the optimization functions, the number of points to be computed can be passed to the `batch_size` argument, while the `method` should be set to "Adam" to enable stochastic optimization via the Adam algorithm (Kingma and Ba 2014). The Adam algorithm can be fine-tuned by setting the learning rate `lr` (default 0.1) and the number of optimization steps `steps` (default 100). Parallel multi-point optimization can be paired with asynchronous, constrained, and mixed optimization, which are all detailed in this section.

```
>>> from nubo.acquisition import MCEExpectedImprovement,
...                               MCUpperConfidenceBound
>>> from nubo.optimisation import multi_joint, multi_sequential
>>> acq = MCEExpectedImprovement(gp = gp, y_best = torch.max(y_train),
...                               samples = 256)
>>> acq = MCUpperConfidenceBound(gp = gp, beta = 4, samples = 256)
```



```
>>> x_new, _ = multi_joint(func = acq, method = "Adam", lr = 0.1,
...                         steps = 100, batch_size = 4, bounds = bounds)
>>> x_new, _ = multi_sequential(func = acq, method = "Adam", lr = 0.1,
...                             steps = 100, batch_size = 4, bounds = bounds)
```

To enable the use of deterministic optimizers, such as L-BFGS-B (Zhu *et al.* 1997) and SLSQP (Kraft 1994), the base samples used to compute the Monte Carlo samples can be fixed by setting `fix_base_samples = True` (default `False`).

```
>>> from nubo.acquisition import MUpperConfidenceBound
>>> from nubo.optimisation import multi_joint, multi_sequential
>>> acq = MUpperConfidenceBound(gp = gp, beta = 4, fix_base_samples = True)
>>> x_new, _ = multi_joint(func = acq, method = "L-BFGS-B",
...                       batch_size = 4, bounds = bounds)
>>> acq = MUpperConfidenceBound(gp = gp, beta = 4, fix_base_samples = True)
>>> x_new, _ = multi_sequential(func = acq, method = "L-BFGS-B",
...                             batch_size = 4, bounds = bounds)
```

### *Asynchronous optimization*

NUBO supports asynchronous optimization, that is, the continuation of the optimization loop while some points are being evaluated by the objective function. In this case, the Monte Carlo acquisition functions `MExpectedImprovement` or `MUpperConfidenceBound` are used as outlined in Section 2.2. The code snippet below assumes that the two points `x_pend` are currently in the evaluation process. To continue the optimization, these points can be fed into the acquisition function by setting `x_pending = x_pend` and NUBO will take them into account for the subsequent iteration.

```
>>> import torch
>>> from nubo.acquisition import MUpperConfidenceBound
>>> from nubo.optimisation import multi_joint, multi_sequential
>>> x_pend = torch.tensor([[0.2, 0.9, 0.8, 0.4, 0.4, 0.1],
...                       [0.1, 0.3, 0.7, 0.2, 0.1, 0.2]])
>>> acq = MUpperConfidenceBound(gp = gp, beta = 4, x_pending = x_pend)
>>> x_new, _ = multi_joint(func = acq, method = "Adam",
...                       batch_size = 4, bounds = bounds)
>>> x_new, _ = multi_sequential(func = acq, method = "Adam",
...                             batch_size = 4, bounds = bounds)
```

While Monte Carlo acquisition functions are approximations of the analytical functions, they are mainly used for computing multiple points, where analytical functions are generally intractable. The Monte Carlo approach can also be used for single-point asynchronous optimization by setting `batch_size = 1`.

### *Constrained optimization*

The simple maximization problem in Equation 1 can be extended by including one or more

input constraints

$$\begin{aligned}
 \mathbf{x}^* &= \arg \max_{\mathbf{x} \in \mathcal{X}} f(\mathbf{x}), \\
 \text{subject to } g_i(\mathbf{x}) &= 0 & \forall i = 1, \dots, I & \text{ [Equality constraint]} \\
 h_j(\mathbf{x}) &\geq 0 & \forall j = 1, \dots, J & \text{ [Inequality constraint]}.
 \end{aligned} \tag{9}$$

In these instances, **NUBO** allows constrained Bayesian optimization by using the SLSQP algorithm to optimize the acquisition function. Implementing this method requires the additional step of specifying the constraints `cons` as a dictionary for one constraint or a list of dictionaries for multiple constraints. Each constraint requires two entries. The first is `"type"` and can either be set to `"ineq"` for inequality constraints or `"eq"` for equality constraints. The second is `"fun"`, which takes a function representing the constraint. The optimizer only selects points for which the constraint functions are greater than or equal to zero for inequality constraints and exactly zero for equality constraints. The code snippet below specifies two constraints: The first is an inequality constraint that requires the first two input dimensions to be smaller than or equal to 0.5. The second is an equality constraint that requires dimensions four, five, and six to sum to 1.2442.

```
>>> import torch
>>> bounds = torch.tensor([[0., 0., 0., 0., 0., 0.],
...                        [1., 1., 1., 1., 1., 1.]])
>>> cons = [{"type": "ineq", "fun": lambda x: 0.5 - x[0] - x[1]},
...         {"type": "eq", "fun": lambda x: 1.2442 - x[3] - x[4] - x[5]}]
```

After setting up the input space using the bounds and constraints, the Bayesian optimization loop is similar to before. We need to set the `method` argument of the optimization function to `"SLSQP"` and provide the function with the constraints `cons`.

```
>>> from nubo.acquisition import UpperConfidenceBound
>>> from nubo.optimisation import single
>>> acq = UpperConfidenceBound(gp = gp, beta = 4)
>>> x_new, _ = single(func = acq, method = "SLSQP",
...                   bounds = bounds, constraints = cons)
```

Constrained Bayesian optimization can be used with analytical and Monte Carlo acquisition functions as well as single-point, multi-point, asynchronous, and mixed optimization, all of which are detailed in this section.

### *Mixed optimization*

Bayesian optimization is predominantly focused on problems with continuous input parameters since the Gaussian process models all input dimensions as continuous variables. However, **NUBO** supports the optimization of mixed input parameter spaces via a workaround. To do this, **NUBO** first computes all possible combinations of the discrete parameters. Then, it maximizes the acquisition function for all continuous parameters while holding one combination of the discrete parameters fixed. Once the acquisition function is maximized for each of the possible discrete combinations, the best overall solution is returned. Note that this can be very time-consuming for many discrete dimensions or discrete values.

To implement mixed optimization in **NUBO**, bounds are specified as before, but the discrete dimensions are additionally defined in a dictionary where the keys are the dimensions (starting from zero) and the values are a list of all possible values for the discrete inputs. The code below specifies dimensions one and five as `disc`.

```
>>> import torch
>>> bounds = torch.tensor([[0., 0., 0., 0., 0., 0.],
...                        [1., 1., 1., 1., 1., 1.]])
>>> disc = {0: [0.2, 0.4, 0.6, 0.8],
...         4: [0.3, 0.6, 0.9]}
```

After setting up the input space specified by the bounds and discrete values, the Bayesian optimization loop is similar to before. We only need to provide the function with the dictionary specifying the discrete dimensions `discrete=disc`.

```
>>> from nubo.acquisition import UpperConfidenceBound
>>> from nubo.optimisation import single
>>> acq = UpperConfidenceBound(gp = gp, beta = 4)
>>> x_new, _ = single(func = acq, method = "L-BFGS-B",
...                   bounds = bounds, discrete = disc)
```

Mixed Bayesian optimization can be used in unison with analytical and Monte Carlo acquisition functions as well as single-point, multi-point, asynchronous, and constrained optimization, all of which are detailed in this section.

### 3.3. Test functions and utilities

**NUBO** provides a selection of test functions and utilities to make implementing and testing Bayesian optimization algorithms more convenient. The ten test functions were selected from the virtual library of [Surjanovic and Bingham \(2013\)](#) and represent a variety of challenges, such as bowl-shaped, plate-shaped, valley-shaped, uni-modal, and multi-modal functions. The functions can be imported from the `nubo.test_functions` module and instantiated by providing the number of dimensions (except for the Hartmann function that comes in 3D and 6D versions), the standard deviation of any noise that should be added, and whether the function should be minimized or maximized. These functions are equipped with the following attributes: the number of dimensions `dims`, the bounds `bounds`, and the inputs and outputs of the global optimum `optimum`.

```
>>> from nubo.test_functions import Ackley, Hartmann6D
>>> func = Ackley(dims = 5, noise_std = 0.1, minimise = False)
>>> func = Hartmann6D(minimise = False)
>>> dims = func.dims
>>> bounds = func.bounds
```

The `gen_inputs` function from the `nubo.utils` module allows us to generate input data that covers the input space efficiently by sampling a larger number of random Latin hypercube designs ([McKay et al. 1979](#)) and returning the design with the largest minimal distance between all points. Figure 3.3 compares Latin hypercube sampling to random sampling for

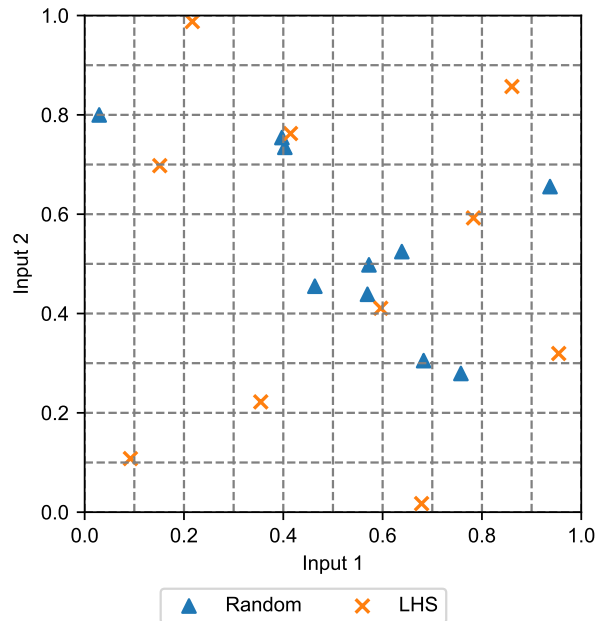


Figure 4: Latin hypercube sampling compared to random sampling.

two input dimensions. While many random points are in close proximity to each other, points from the Latin hypercube design cover the whole input space effectively by only placing one point in each row and column. The exact position of the point within the selected square is random. The code snippet below generates five points for each input dimension of the Hartmann function initiated above and uses `func` to compute the corresponding outputs.

```
>>> from nubo.utils import gen_inputs
>>> x_train = gen_inputs(num_points = dims * 5, num_dims = dims,
...                      bounds = bounds)
>>> y_train = func(x_train)
```

Finally, we discuss three convenience functions that can be used for data transformation. `normalise` and `unnormalise` can be used to scale input data to the unit cube  $[0, 1]^d$  and back to its original domain by providing the bounds of the input space. Furthermore, the outputs can be centred at zero with a standard deviation of one with the `standardise` function.

```
>>> from nubo.utils import standardise, normalise, unnormalise
>>> x_norm = normalise(x_train, bounds = bounds)
>>> x_train = unnormalise(x_norm, bounds = bounds)
>>> y_stand = standardise(y_train)
```

## 4. Case study

We present the general workflow for optimising an expensive-to-evaluate black-box function with **NUBO** by providing a detailed case study in which a test function with six input di-

mensions is optimized. This case study demonstrates how the user can specify the parameter input space, generate initial training data, and define and run the Bayesian optimization loop. So that the case study is reproducible, we set the seed for the pseudo-number generator within **Torch** to 123. We set some format options for the `print` function such that values are rounded to the fourth decimal place and are not formatted in scientific notation to increase readability.

```
>>> import torch
>>> torch.manual_seed(123)
>>> torch.set_printoptions(precision = 4, sci_mode = False)
```

A typical objective function optimized with Bayesian optimization is expensive to evaluate and thus not feasible to use in a case study that aims to illustrate how **NUBO** can be applied. Hence, we will use one of the synthetic test functions provided by **NUBO** as a surrogate expensive-to-evaluate black-box function. We use the six-dimensional Hartmann function that possesses multiple local and one global minimum. Its input space is bounded by the hyper-rectangle  $[0, 1]^6$ . Observational noise, such as measurement error, is represented by adding a small amount of random Gaussian noise to the function output by setting `noise_std = 0.1`. `minimise` is set to `False` to transform the minimization into a maximization problem as required for Bayesian optimization with **NUBO**.

```
>>> from nubo.test_functions import Hartmann6D
>>> black_box = Hartmann6D(noise_std = 0.1, minimise = False)
```

With our objective function specified, we can focus on defining the input space. We know that our objective function has six inputs that are all bounded by  $[0, 1]$ . As introduced in Section 3.2, the bounds are defined as a  $2 \times d$  `torch.tensor`, where the first row specifies the lower bounds and the second row specifies the upper bounds. This case study also highlights the mixed parameter optimization capabilities of **NUBO** (see Section 3.2) by assuming that the first input is a discrete parameter restricted to 0.2, 0.4, 0.6, and 0.8. We can implement this by specifying a dictionary, where the key is the input dimension and the value is a list of possible values the input can take, that is `{0: [0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0]}`. Note that indexing starts at zero in Python.

```
>>> dims = 6
>>> bounds = torch.tensor([[0., 0., 0., 0., 0., 0.],
...                        [1., 1., 1., 1., 1., 1.]])
>>> discrete = {0: [0.0, 0.1, 0.2, 0.3, 0.4, 0.5,
...                0.6, 0.7, 0.8, 0.9, 1.0]}
```

The Bayesian optimization loop requires initial training data. This is important to train the Gaussian process that tries to emulate the objective function. This case study uses the `gen_inputs` function introduced in Section 3.3 to generate 30 initial data points from a Latin hypercube design. We round the first input dimension to fit the discrete values specified above as Latin hypercube designs return continuous values. These points are evaluated by the objective function to complete our training data pairs consisting of input parameters `x_train` and observations `y_train`.

```

>>> from nubo.utils import gen_inputs
>>> x_train = gen_inputs(num_points = dims * 5, num_dims = dims,
...                      bounds = bounds)
>>> x_train[:, 0] = torch.round(x_train[:, 0], decimals = 1)
>>> y_train = black_box(x_train)

```

Next, we specify the Bayesian optimization algorithm we plan to use in our optimization loop. We define the `bo` function that takes our training pairs (`x_train`, `y_train`) and returns the next candidate point `x_new` which is evaluated by the objective function in four steps. First, we set up our surrogate model as the Gaussian process provided by **NUBO** with a Gaussian likelihood as discussed in Section 3.1. Second, we train the Gaussian process `gp` with our training data by maximising the likelihood with the Adam algorithm (Kingma and Ba 2014) via the `fit_gp` function. Here, we set a custom learning rate `lr` and the number of optimization steps `steps`. Third, we define an acquisition function that will guide our optimization. As we assume that our objective function allows parallel function evaluations, we decide to compute multi-point batches at each iteration and choose a Monte Carlo acquisition function, in this case `MCUpperConfidenceBound`. The acquisition function `acq` is instantiated by providing it with the fitted Gaussian process `gp`, a value for the trade-off hyper-parameter `beta`, and the number of Monte Carlo samples used to approximate the acquisition function. For further details, refer to Section 2.2. Fourth, we maximize the acquisition function `acq` with the `multi_sequential` function that uses the sequential strategy for computing multiple candidate points. We decide to compute four candidate points at each iteration by setting `batch_size=4` and providing the previously specified bounds and discrete values. The Adam optimizer is used as Monte Carlo acquisition functions require a stochastic optimizer due to their inherent randomness introduced by drawing the Monte Carlo samples. The optimizer is initialized at two different initial points chosen as the two points with the highest acquisition value out of 100 potential points sampled from a Latin hypercube design. We chose two initializations to keep the computational overhead within the replication script low. In practice, a higher number of initializations might be beneficial.

```

>>> from nubo.acquisition import MCUpperConfidenceBound
>>> from nubo.models import GaussianProcess, fit_gp
>>> from nubo.optimisation import multi_sequential
>>> from gpytorch.likelihoods import GaussianLikelihood
>>>
>>> def bo(x_train, y_train):
>>>
>>>     likelihood = GaussianLikelihood()
>>>     gp = GaussianProcess(x_train, y_train, likelihood = likelihood)
>>>     fit_gp(x_train, y_train, gp = gp, likelihood = likelihood,
...          lr = 0.1, steps = 200)
>>>     acq = MCUpperConfidenceBound(gp = gp, beta = 4, samples = 128)
>>>     x_new, _ = multi_sequential(func = acq, method = "Adam",
...                                batch_size = 4, bounds = bounds,
...                                discrete = discrete, lr = 0.1,
...                                steps = 200, num_starts = 2,
...                                num_samples = 100)

```

```
>>>
>>>     return x_new
```

Finally, we specify the entire optimization loop, that is, a simple `for`-loop that computes the next batch of candidate points using the defined Bayesian optimization algorithm `bo`, evaluates the candidate points by the objective function `black_box`, and adds the new data pairs `(x_new, y_new)` to the training data. We let the optimization loop run for ten iterations and print all evaluations, where the first six columns are the inputs and the final column is the output from the objective function. The first 30 rows give the initial training data generated by the Latin hypercube design, while the last 40 rows were chosen by the Bayesian optimization algorithm. The results clearly show that **NUBO** improves upon the initial space-filling design and produces points which are consistent with the bounds and discrete values that specify the parameter input space.

```
>>> iters = 10
>>> for iter in range(iters):
>>>
>>>     x_new = bo(x_train, y_train)
>>>
>>>     y_new = black_box(x_new)
>>>
>>>     x_train = torch.vstack((x_train, x_new))
>>>     y_train = torch.hstack((y_train, y_new))
>>>
>>> print(torch.hstack([x_train, y_train.reshape(-1, 1)]))

tensor([[0.2000, 0.6523, 0.1574, 0.7822, 0.3039, 0.8603, 0.1251],
        [0.5000, 0.9127, 0.8746, 0.4787, 0.6523, 0.1249, 2.2907],
        [0.4000, 0.5638, 0.0459, 0.6200, 0.7056, 0.2929, 0.6744],
        [0.2000, 0.3003, 0.2290, 0.8110, 0.9529, 0.2384, 0.0442],
        [0.1000, 0.7809, 0.5374, 0.1381, 0.5655, 0.5679, 0.6123],
        ...,
        [0.5000, 1.0000, 1.0000, 0.5839, 0.0000, 0.3425, 0.7659],
        [0.4000, 0.8899, 1.0000, 0.5622, 0.0000, 0.0421, 3.1330],
        [0.4000, 0.8026, 0.0000, 0.5356, 0.0000, 0.0000, 2.7828],
        [0.2000, 0.0556, 1.0000, 1.0000, 0.0000, 0.6695, 0.1134],
        [0.4000, 0.7124, 1.0000, 0.6842, 0.0000, 0.0000, 2.3028]],
        dtype=torch.float64)
```

**NUBO** explores the parameter space efficiently by switching between exploring areas with high uncertainty and areas with high predictions. This means that the algorithm does not monotonically converge to a single solution as conventional optimization algorithms would. Thus, the approximate solution to an objective function is the best value found during optimization. In this case study, the approximate solution, i.e., the solution with the highest output (last column in the Python output above), was found at iteration 53, and the inputs and outputs are printed below.



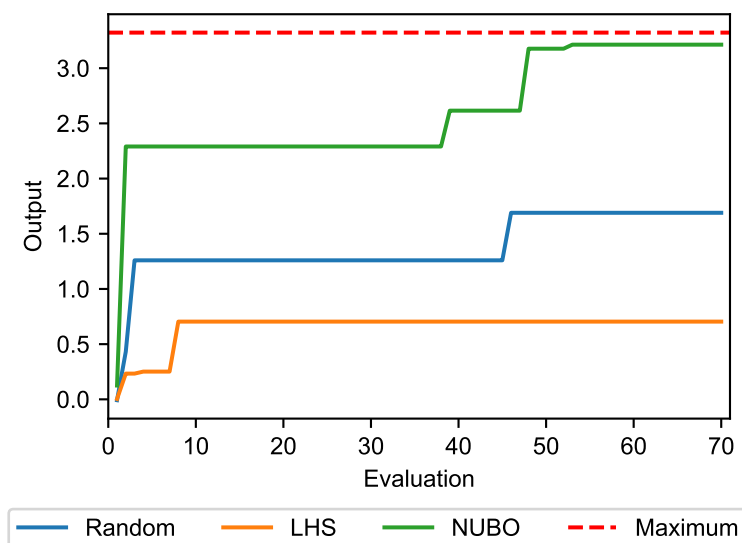


Figure 5: Results of the Bayesian optimization algorithm implemented with **NUBO** as defined in this case study compared to random sampling and Latin hypercube sampling.

```
>>> best_iter = int(torch.argmax(y_train))
>>> print("Approximate solution")
>>> print("-----")
>>> print(f"Evaluation: {best_iter + 1}")
>>> print(f"Inputs: {x_train[best_iter]}")
>>> print(f"Output: {y_train[best_iter]:.4f}")
```

Approximate solution

-----

Evaluation: 53

Inputs: tensor([0.4000, 0.9136, 1.0000, 0.5669, 0.0000, 0.0802],  
dtype=torch.float64)

Output: 3.2133

We compare the results provided by **NUBO** with the results from random sampling and using a space-filling design, in this case, Latin hypercube sampling (LHS). The code below generates results for the full budget of 70 evaluations for both sampling methods and plots the results against each other, with the number of evaluations on the x-axis and the accumulative best output for each method on the y-axis. Figure 4 shows that **NUBO** (green line) provides a better solution than either alternative approach and is very close to the true maximum of 3.32237. **NUBO** succeeds in accurately approximating the true optimum.

```
>>> import matplotlib.pyplot as plt
>>> import numpy as np
>>> torch.manual_seed(123)
>>> random = black_box(torch.rand((70, dims)))
```

```

>>> lhs = black_box(gen_inputs(num_points = 70, num_dims = dims,
...                             bounds = bounds))
>>> plt.plot(range(1, 71), np.maximum.accumulate(random), label = "Random")
>>> plt.plot(range(1, 71), np.maximum.accumulate(lhs), label = "LHS")
>>> plt.plot(range(1, 71), np.maximum.accumulate(y_train), label = "NUBO")
>>> plt.hlines(3.32237, 0, 71, colors = "red", linestyle = "dashed",
...            label = "Maximum")
>>> plt.title("Comparison against random designs")
>>> plt.xlabel("Evaluations")
>>> plt.ylabel("Output")
>>> plt.legend(loc = 'lower center', ncol = 4, bbox_to_anchor = (0.5, -0.275))
>>> plt.xlim(0, 71)
>>> plt.tight_layout()

```

## 5. Conclusion

This article introduces **NUBO**, a Python package for Bayesian optimization to optimize expensive-to-evaluate black-box functions, for example, computer simulators and physical experiments. The main objective of **NUBO** is to make Bayesian optimization accessible to researchers from all disciplines by providing a transparent and user-friendly implementation.

**NUBO** includes five sub-modules that implement Gaussian processes, acquisition functions, optimizers, test functions, and utilities. These modules provide all necessities for sequential single-point, parallel multi-point, and asynchronous optimization of expensive-to-evaluate black-box functions for bounded, constrained, and/or mixed (discrete and continuous) input parameter spaces. We have introduced and explained each of these functionalities with individual code snippets and illustrated **NUBO**'s general workflow using a detailed case study that takes a hypothetical six-dimensional expensive-to-evaluate black-box function and approximates its global optimum with a parallel multi-point Bayesian optimization algorithm.

A brief comparison with other Python packages for Bayesian optimization showed that **NUBO** has competitive performance while providing a transparent and simple implementation. This makes **NUBO** a good candidate for the optimization of expensive black-box functions when transparency is vital.

In the future, we plan to extend **NUBO** to include optimization strategies for multi-fidelity, multi-objective, and high-dimensional problems.

## Computational details

The results in this paper were obtained using Python 3.11.2 with the following packages: **NUBO** 1.2.1, **Torch** 2.0.0, **GPyTorch** 1.10, **SciPy** 1.10.1, **NumPy** 1.24.2, and **Matplotlib** 3.9.0 (Hunter 2007). For the package comparison **bayes\_opt** 1.4.3, **BoTorch** 0.8.4, **pyGPGO** 0.5.0 and **SMAC3** 2.0.0 were used. Python itself is available from the Python website at <https://www.python.org/> and all packages used are available from the Python Package Index (PyPI) at <https://www.pypi.org/>.

## Acknowledgments

The work has been supported by the Engineering and Physical Sciences Research Council (EPSRC) under grant number EP/T020946/1 and the EPSRC Centre for Doctoral Training in Cloud Computing for Big Data under grant number EP/L015358/1.

## References

- Balandat M, Karrer B, Jiang DR, Daulton S, Letham B, Wilson AG, Bakshy E (2020). “**BoTorch**: A Framework for Efficient Monte-Carlo Bayesian Optimization.” In *Advances in Neural Information Processing Systems 33*. Python package version 0.8.4, URL <https://proceedings.neurips.cc/paper/2020/hash/f5b1b89d98b7286673128a5fb112cb9a-Abstract.html>.
- Diessner M, O’Connor J, Wynn A, Laizet S, Guan Y, Wilson K, Whalley RD (2022). “Investigating Bayesian Optimization for Expensive-to-Evaluate Black Box Functions: Application in Fluid Dynamics.” *Frontiers in Applied Mathematics and Statistics*, **8**(1076296). doi:[10.3389/fams.2022.1076296](https://doi.org/10.3389/fams.2022.1076296).
- Frazier PI (2018). “A Tutorial on Bayesian Optimization.” *arXiv 1807.02811*, arXiv.org E-Print Archive. doi:[10.48550/arXiv.1807.02811](https://doi.org/10.48550/arXiv.1807.02811).
- Gardner JR, Pleiss G, Bindel D, Weinberger KQ, Wilson AG (2018). “**GPyTorch**: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration.” In *Advances in Neural Information Processing Systems*. Python package version 1.10.
- Gramacy RB (2020). *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. Chapman & Hall/CRC, Boca Raton, Florida. doi:[10.1201/9780367815493](https://doi.org/10.1201/9780367815493). URL <https://bobby.gramacy.com/surrogates/>.
- Harris CR, Millman KJ, Van Der Walt SJ, Gommers R, Virtanen P, Cournapeau D, Wieser E, Taylor J, Berg S, Smith NJ, *et al.* (2020). “Array Programming with **NumPy**.” *Nature*, **585**(7825), 357–362. doi:[10.1038/s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2). Python package version 1.24.2.
- Harvard University, University of Toronto, Université de Sherbrooke, Socpra Sciences et Génie SEC (2014). *Spearmint Software*. Python package version 0.1, URL <https://github.com/HIPS/Spearmint>.
- Hunter JD (2007). “**matplotlib**: A 2D Graphics Environment.” *Computing in Science & Engineering*, **9**(3), 90–95. doi:[10.1109/mcse.2007.55](https://doi.org/10.1109/mcse.2007.55). Python package version 3.9.0.
- Jiménez J, Ginebra J (2017). “**pyGPGO**: Bayesian Optimization for Python.” *The Journal of Open Source Software*, **2**(19), 431. doi:[10.21105/joss.00431](https://doi.org/10.21105/joss.00431). Python package version 0.5.0.
- Jones DR, Schonlau M, Welch WJ (1998). “Efficient Global Optimization of Expensive Black-Box Functions.” *Journal of Global Optimization*, **13**(4), 455. doi:[10.1023/a:1008306431147](https://doi.org/10.1023/a:1008306431147).

- Kingma DP, Ba J (2014). “Adam: A Method for Stochastic Optimization.” *arXiv 1412.6980*, arXiv.org E-Print Archive. doi:10.48550/arXiv.1412.6980.
- Kraft D (1994). “Algorithm 733: **TOMP** – Fortran Modules for Optimal Control Calculations.” *ACM Transactions on Mathematical Software*, **20**(3), 262–281. doi:10.1145/192115.192124.
- Kushner HJ (1964). “A New Method of Locating the Maximum Point of an Arbitrary Multipeak Curve in the Presence of Noise.” *Journal of Basic Engineering*, **86**(1), 97–106. doi:10.1115/1.3653121.
- Lindauer M, Eggensperger K, Feurer M, Biedenkapp A, Deng D, Benjamins C, Ruhkopf T, Sass R, Hutter F (2022). “**SMAC3**: A Versatile Bayesian Optimization Package for Hyperparameter Optimization.” *Journal of Machine Learning Research*, **23**(54), 1–9. Python package version 2.0.0.
- Mahfoze OA, Moody A, Wynn A, Whalley RD, Laizet S (2019). “Reducing the Skin-Friction Drag of a Turbulent Boundary-Layer Flow with Low-Amplitude Wall-Normal Blowing within a Bayesian Optimization Framework.” *Physical Review Fluids*, **4**(9), 094601. doi:10.1103/physrevfluids.4.094601.
- McKay MD, Beckman RJ, Conover WJ (1979). “A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code.” *Technometrics*, **21**(2), 239–245. doi:10.2307/1268522.
- Moćkus J (1975). “On Bayesian Methods for Seeking the Extremum.” In *Optimization Techniques IFIP Technical Conference: Novosibirsk, July 1–7, 1974*, pp. 400–404. Springer-Verlag. doi:10.1007/3-540-07165-2\_55.
- Moćkus J (1989). *Bayesian Approach to Global Optimization*, volume 37 of *Mathematics and Its Applications*. 1st edition. Springer-Verlag, Dordrecht. doi:10.1007/978-94-009-0909-0\_7.
- Neal RM (1996). *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. 1st edition. Springer-Verlag, New York. doi:10.1007/978-1-4612-0745-0.
- Nogueira F (2014). **bayes\_opt**: Open Source Constrained Global Optimization Tool for Python. Python package version 1.4.3, URL <https://github.com/fmfn/BayesianOptimization>.
- O’Connor J, Diessner M, Wilson K, Whalley RD, Wynn A, Laizet S (2023). “Optimisation and Analysis of Streamwise-Varying Wall-Normal Blowing in a Turbulent Boundary Layer.” *Flow, Turbulence and Combustion*, pp. 1–29. doi:10.1007/s10494-023-00408-3.
- Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L (2019). “**PyTorch**: An Imperative Style, High-Performance Deep Learning Library.” *Advances in Neural Information Processing Systems*, **32**. Python package version 2.0.0.
- Python Software Foundation (2003). “Python Package Index – PyPI.” Accessed 2023-05-09, URL <https://pypi.org/>.

- Rasmussen CE, Williams CKI (2006). *Gaussian Processes for Machine Learning*. The MIT Press, Cambridge.
- Shahriari B, Swersky K, Wang Z, Adams RP, De Freitas N (2015). “Taking the Human Out of the Loop: A Review of Bayesian Optimization.” *Proceedings of the IEEE*, **104**(1), 148–175. doi:10.1109/jproc.2015.2494218.
- Snoek J, Larochelle H, Adams RP (2012). “Practical Bayesian Optimization of Machine Learning Algorithms.” *Advances in Neural Information Processing Systems*, **25**.
- Srinivas N, Krause A, Kakade S, Seegre M (2010). “Gaussian Process Optimization in the Bandit Setting: No Regret and Experimental Design.” In *Proceedings of the International Conference on Machine Learning 2010*.
- Storn R, Price K (1997). “Differential Evolution – A Simple and Efficient Heuristic for Global Optimization over Continuous Spaces.” *Journal of Global Optimization*, **11**(4), 341–359. doi:10.1023/a:1008202821328.
- Surjanovic S, Bingham D (2013). “Virtual Library of Simulation Experiments: Test Functions and Datasets.” Accessed 2023-05-09, URL <https://www.sfu.ca/~ssurjano/>.
- The **GPyOpt** authors (2016). *GPyOpt: A Bayesian Optimization Framework in Python*. Python package version 1.2.6, URL <http://github.com/SheffieldML/GPyOpt>.
- The **pip** Developers (2023). *pip: Package Installer for Python*. Python package version 22.3.1, URL <https://pip.pypa.io/en/stable/>.
- Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, van der Walt SJ, Brett M, Wilson J, Millman KJ, Mayorov N, Nelson ARJ, Jones E, Kern R, Larson E, Carey CJ, Polat İ, Feng Y, Moore EW, VanderPlas J, Laxalde D, Perktold J, Cimrman R, Henriksen I, Quintero EA, Harris CR, Archibald AM, Ribeiro AH, Pedregosa F, van Mulbregt P, SciPy 10 Contributors (2020). “**SciPy** 1.0: Fundamental Algorithms for Scientific Computing in Python.” *Nature Methods*, **17**, 261–272. doi:10.1038/s41592-019-0686-2. Python package version 1.10.1.
- Wang K, Dowling AW (2022). “Bayesian Optimization for Chemical Products and Functional Materials.” *Current Opinion in Chemical Engineering*, **36**, 100728. doi:10.1016/j.coche.2021.100728.
- White C, Neiswanger W, Savani Y (2021). “BANANAS: Bayesian Optimization with Neural Architectures for Neural Architecture Search.” In *Proceedings of the AAAI Conference on Artificial Intelligence*.
- Wilson J, Hutter F, Deisenroth M (2018). “Maximizing Acquisition Functions for Bayesian Optimization.” *Advances in Neural Information Processing Systems*, **31**.
- Wilson S (2022). *ParBayesianOptimization: Parallel Bayesian Optimization of Hyperparameters*. doi:10.32614/CRAN.package.parbayesianoptimization. R package version 1.2.6.
- Wu J, Chen XY, Zhang H, Xiong LD, Lei H, Deng SH (2019). “Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimization.” *Journal of Electronic Science and Technology*, **17**(1), 26–40.

- Yan Y (2024). **rBayesianOptimization**: Bayesian Optimization of Hyperparameters. doi: [10.32614/CRAN.package.rbayesianoptimization](https://doi.org/10.32614/CRAN.package.rbayesianoptimization). R package version 1.2.1.
- Zhu C, Byrd RH, Lu P, Nocedal J (1997). “Algorithm 778: **L-BFGS-B**: Fortran Subroutines for Large-Scale Bound-Constrained Optimization.” *ACM Transactions on Mathematical Software*, **23**(4), 550–560. doi:[10.1145/279232.279236](https://doi.org/10.1145/279232.279236).
- Žilinskas AG (1975). “Single-Step Bayesian Search Method for an Extremum of Functions of a Single Variable.” *Cybernetics*, **11**(1), 160–166. doi:[10.1007/bf01069961](https://doi.org/10.1007/bf01069961).

**Affiliation:**

Mike Diessner  
Newcastle University  
School of Computing  
Urban Sciences Building  
Newcastle upon Tyne NE4 5TG, United Kingdom  
E-mail: [m.diessner2@ncl.ac.uk](mailto:m.diessner2@ncl.ac.uk)  
URL: <https://mikediessner.github.io/>

Kevin J. Wilson  
Newcastle University  
School of Mathematics, Statistics and Physics  
Herschel Building  
Newcastle upon Tyne NE1 7RU, United Kingdom  
E-mail: [kevin.wilson@ncl.ac.uk](mailto:kevin.wilson@ncl.ac.uk)

Richard D. Whalley  
Queen’s University Belfast  
School of Mechanical & Aerospace Engineering  
Ashby Building  
Belfast BT9 5AH, United Kingdom  
E-mail: [r.whalley@qub.ac.uk](mailto:r.whalley@qub.ac.uk)  
URL: <https://www.experimental-fluid-dynamics.com/>