Wasserstein-enabled characterization of designs and myopic decisions in Bayesian Optimization

[Submitted to Special session 8: Bayesian optimization]

Antonio Candelieri[1 [0000−0003−1431−576X]] and Francesco Archetti[1 [0000−0003−1131−3830]]

University of Milano-Bicocca, Milan, 20126, Italy
{antonio.candelieri,francesco.archetti}@unimib.it

Abstract. Impractical assumptions, an inherently myopic nature, and the crucial role of the initial design, all together contribute to making theoretical convergence proofs of little value in real-life Bayesian Optimization applications. In this paper, we propose a novel characterization of the design depending on its distributional properties, separately measured with respect to the coverage of the search space and the concentration around the best observed function value. These measures are based on the Wasserstein distance and enable a model-free evaluation of the information value of the design before deciding the next query. Then, embracing the myopic nature of Bayesian Optimization, we take an empirical approach to analyze the relation between the proposed characterization of the design and the quality of the next query. Ultimately, we provide important and useful insights that might inspire the definition of a new generation of acquisition functions in Bayesian Optimization.

Keywords: Bayesian Optimization · Wasserstein distance · myopic acsuisition functions

1 Introduction

Bayesian Optimization (BO) [2, 14] is a sample-efficient sequential model-based algorithm for global optimization of black-box, expensive-to-evaluate, non-convex objective functions. The reference problem is

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

where $\mathcal{X} \subset \mathbb{R}^d$ is the so-called search space, typically box-bounded.

BO iterates by (i) fitting a probabilistic surrogate model of $f(x)$ to an available set of observations $\mathcal{D} = \{(x_i, y_i)\}_{i=1:n}$, usually called *design*, and (ii) selecting the next *query*, namely x', by optimizing an acquisition function that balances between exploration and exploitation, according to the surrogate.

BO is known to be *sample efficient*, meaning that it finds a better solution than other global optimization strategies, within a few queries. Thanks to this

property, it has been and is still widely adopted to address industrial-relevant applications [21, 27] and also Machine Learning tasks [1, 11, 17].

Although theoretical convergence proofs exist for specific BO implementations, such as the widely known [28], actual convergence in real-life settings is a critical issue, mostly due to impractical assumptions of convergence theorems.

The most relevant and well-known obstacle to global convergence of BO is the mispecification of the probabilistic surrogate model, typically a Gaussian Process (GP) [15, 32]. Theoretical convergence proofs hold only in the standard setting [5], meaning that the objective function is a member of a specific Reproducing Kernel Hilbert Space (RKHS) with bounded norm, known a-priori. In other terms, convergence to the optimum is proved under the assumption that the GP kernel's hyperparameters are known in advance, which is not the case in practice. In BO, the GP kernel's hyperparameters are tuned depending on the available design $\mathcal{D}$ via Maximum Likelihood Estimation (MLE) or Maximum-A-Posteriori (MAP), leading common "no regret" acquisition functions to get stuck in local optima [7, 9]. In simpler terms, sequential queries are model-based and thus hyperparameter-sensitive; the acquisition function can introduce a feedback loop where sequential queries reinforce bad hyperparameters. Heuristic methods [30, 31, 29, 4] have been proposed over time, basically using Lemma 4 of [7], stating that decreasing the *length-scale* hyperparameter of the kernel leads to consider a larger RKHS, increasing exploration at the expense of a higher regret. Other methods to increase the chance of global convergence propose random exploration steps [8, 20] or inexact optimization of the acquisition function [19].

Another issue, strictly related to model mispecification, is that the random design generated to initialize BO can definitely prevent from reaching the global optimum, regardless of the acquisition function and its specific exploitation-exploration trade-off mechanism. Indeed, the typical experimental setting for BO consists in averaging on multiple independent runs from different initial designs, all with the same number of random observations, aiming at mitigating the impact of random initialization. However, this way of proceeding is impractical for many real-life problems with highly expensive objective functions.

Indeed, many recent papers propose strategies to dynamically select the behavior of the acquisition function [8, 20], the model's hyperparameters [30, 31, 29, 4], or both [23], to possibly revert from being stack into local optima. The long-term perspective proposed in [6], consists of developing approaches that dynamically decide whether to take the next sample from a (quasi-)random distribution or whether to derive it from the surrogate model, with an expected substantial improvement of performance, especially when the maximum allowed number of queries is known in advance.

A specific analysis on the difficulty of generating a good design for GP fitting – not necessarily for BO – is given in [33]. While *maximin* designs can be awful, Latin Hypercube Sampling (LHS) methods offer better performance. However, neither method is well suited for GP's hyperparameters inference much so that random design can surprisingly result in the most effective strategy. Specifically for (few-shot) BO, [18] has recently proposed a novel acquisition function that

balances predictive uncertainty reduction with hyperparameter learning using information-theoretic principles.

Finally, another – often overlooked – issue affecting the practical convergence of BO is that (almost all) the acquisition functions are *myopic*. Knowledge Gradient (KG) [12] and Predictive Entropy Search (PES) [16] are exceptions, which have been shown to be effective but entailing a relevant computational overhead for nested integrations or Monte Carlo sampling.

Motivated by the issues discussed so far, this paper aims to contribute to this challenge as follows:

- proposing a novel and statistically principled mechanism, based on the Wasserstein distance, to characterize – aka *featurize* – a design depending on its distributional properties;
- analyzing, according to the proposed characterization, the differences between designs suitably generated for representing typical situation in BO (i.e., LHS, queries converging to the global optimum, and queries converging to different a solution);
- investigating possible relations between the characterization of a design, the risk of model mispecification, and the possible improvement provided by different myopic acquisition functions.

For the empirical analysis, a set of well-diversified test problems has been considered. The related works are the references quoted so far.

2 Background

2.1 Bayesian Optimization in brief

Given a set of available observations $\mathcal{D} = \{(x_i, y_i)\}$, we separately denote the set of queried locations by $X = \{x_i\}_{i=1:n}$ and the set of associated values by $Y = \{y_i = f(x_i)\}_{i=1:n}$, with $f(x)$ assumed *noise-free* in this paper.

As a probabilistic surrogate model, we consider a GP, with prediction and associated (model-based) uncertainty denoted by $\mu(x|\mathcal{D})$ and $\sigma(x|\mathcal{D})$, respectively. For the sake of readability, we omit their well-known equations.

As far as the acquisition function is concerned, here, we introduce a slightly different notation from usual, and consider that the next location to query x' is given by a parametric *policy* $\pi(\mathcal{D}|\theta)$, consisting of GP fitting followed by the optimization of a specific acquisition function. Formally,

$$x' \leftarrow \pi(\mathcal{D}|\theta) \tag{2}$$

where θ denotes all the hyperparameters involved, both the GP's and the acquisition function's ones.

Embracing the myopic nature of common acquisition functions, the goal of $\pi(\mathcal{D}|\theta)$ should be to obtain $x' = x^*$ or, more generally, to provide x' such that $x^* \in \mathcal{D} \cup \{(x', y')\}$, with $y' = f(x')$, in order to consider both the case in which the optimum is already in $\mathcal{D}$ or it is exactly x'.

2.2 Wasserstein distance in brief

The Wasserstein distance comes from the Optimal Transport (OT) theory [24, 25] and – contrary to divergence measures – it is a *distance metric* over the space of probability distributions. Moreover, it can be used to compare two discrete or two continuous probability distributions, as well as one discrete and one continuous. Due to these properties, it has recently gained a central role in Machine Learning [22] and Generative AI [10, 13]. For a detailed overview on OT and the Wasserstein distance, the reader can refer to [24, 25]; here, we summarize the background needed for the scope of this paper.

Our setting consists only of empirical discrete probability distributions. Let α and β be two of that kind, we can denote them by:

$$\alpha = \frac{1}{m_\alpha}\sum_{j=1}^{m_\alpha}\delta_{a_j} \quad , \quad \beta = \frac{1}{m_\beta}\sum_{l=1}^{m_\beta}\delta_{b_l} \tag{3}$$

with $a_j, b_l \in \mathbb{R}^d$ and δ_w denoting the Dirac's delta function centered at w.

Another way to look at two empirical discrete probability distributions is to consider them as point clouds, represented as matrices: $A \in \mathbb{R}^{m_\alpha \times d}$, with $A_{j*} = a_j$, and $B \in \mathbb{R}^{m_\beta \times d}$, with $B_{l*} = b_j$. In other terms, A and B are the discrete supports of the two uniform distributions α and β, respectively.

Computing the 2-Wasserstein distance between α and β means solving the following linear programming problem (Kantorovich's formulation):

$$\begin{aligned} \mathcal{W}_2(\alpha,\beta) = \min_{T\in\mathbb{R}^{m_\alpha \times m_\beta}} & \left[\sum_{\substack{j=1:m_\alpha \\ l=1:m_\beta}} \|a_j - b_l\|^2 T_{jl}\right]^{\frac{1}{2}} \\ s.t.\ & T^\top \mathbf{1}_{m_\alpha} = B \\ & T\mathbf{1}_{m_\beta} = A \\ & T_{ij} \geq 0,\ \forall i = 1:m_\alpha, j = 1:m_\beta \end{aligned} \tag{4}$$

where $\mathbf{1}_q$ denotes a vector of all-ones of q dimensions.

In the rest of the paper, we use $\mathcal{W}_2(A,B)$ or $\mathcal{W}_2(\alpha,\beta)$ interchangeably, depending on which notation is more convenient for the discussion.

3 Information value of $\mathcal{D}$ *vs* quality of myopic decisions

Our research hypothesis is that the information value of the current design $\mathcal{D}$ can be described by the distributional properties of its two sets X and Y, which can be jointly analyzed to predict the quality of the next query x', **without relying on any specific model**.

First, we assume that there exists a vector-valued function $\mathcal{S} : [\mathcal{X} \times \mathbb{R}]^n \to \mathbb{R}^h$ characterizing the information value of $\mathcal{D} = (X, Y)$ with respect to h different features. In this study, we propose $h = 2$, with $\mathcal{S}_1(\mathcal{D})$ and $\mathcal{S}_2(\mathcal{D})$ denoting the information value of $\mathcal{D}$ with respect to the coverage of the search space $\mathcal{X}$ and the observed variability of $f(x)$, respectively.

3.1 $\mathcal{S}_1(\mathcal{D})$: information value of $\mathcal{D}$ with respect to $\mathcal{X}$

$\mathcal{S}_1(\mathcal{D})$ is computed as the 2-Wasserstein distance between the available set of queried locations, namely X, and a regular $m \times m$ grid $G_{\mathcal{X}} = \{g_j\}_{j=1:m^2}$ over the search space $\mathcal{X}$. Formally,

$$\mathcal{S}_1(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(X, G_{\mathcal{X}}) \tag{5}$$

Thus, $\mathcal{S}_1(\mathcal{D})$ quantifies how far the set X is from the uniform distribution approximated by $G_{\mathcal{X}}$. A value of $\mathcal{S}_1(\mathcal{D})$ close to zero means that X is spread almost uniformly over $\mathcal{X}$.

Obviously, any discrepancy measure can be alternatively used to quantify $\mathcal{S}_1(\mathcal{D})$, but we suggest to choose among those which are distance metrics (e.g., Star Discrepancy, Maximum-Mean-Discrepancy, Total Variation) and to avoid divergences, such as the Kullback-Leibler divergence.

We use the 2-Wasserstein distance to ensure coherence throughout the proposed framework, since it is the only suitable choice to measure the information value of Y, as explained in the next section.

3.2 $\mathcal{S}_2(\mathcal{D})$: information value of $\mathcal{D}$ with respect to $f(x)$

Since the objective function $f(x)$ is black-box, its co-domain is unknown, and, contrary to $\mathcal{S}_1(\mathcal{D})$, it is not possible to pre-define a regular grid of values to be compared against the available set Y.

On the other hand, an obvious property of a good Y is to show some variability, since the trivial assumption is that $f(x)$ is not flat. Although the standard deviation of Y is an intuitive choice, it is not suitable to quantify the difference between various distributions of Y and their concentration near the best value observed so far, namely the *best seen* $y^+ = \min\{Y\}$. Thus, we use the 2-Wasserstein distance between Y and a Dirac's delta over y^+:

$$\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+}) \tag{6}$$

It is trivial to prove that $\mathcal{W}_2^2(Y, \delta_{y^+}) = \sum_{i=1}^{n}(y_i - y^+) = \left(\sum_{i=1}^{n} y_i\right) - ny^+$.

Finally, any $\mathcal{D}$ can be represented as a point in the space spanned by $\mathcal{S}_1(\mathcal{D})$ and $\mathcal{S}_2(\mathcal{D})$, which are the distributional properties of the sets X and Y.

Figure 1 provides a graphical representation of how the location of a given $\mathcal{D}$ within this space can provide useful insights about the information value of $\mathcal{D}$ itself. A small $\mathcal{W}_2^2(X, G_{\mathcal{X}})$ indicates that X covers the search space $\mathcal{X}$

almost uniformly. A small $\mathcal{W}_2^2(Y, \delta_{y^+})$ suggests that $f(x)$ "looks like" flat: we remark that it "looks like" flat because Y depends on the locations in X. More important, a large $\mathcal{W}_2^2(Y, \delta_{y^+})$ and a small $\mathcal{W}_2^2(X, G_{\mathcal{X}})$ suggest that the best solution of the design, denoted by (x^+, y^+), could lie within a narrow region.

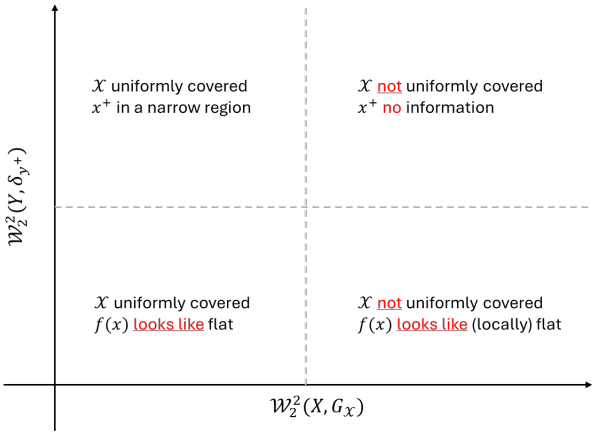

Fig. 1. A conceptual representation of how a design $\mathcal{D}$ is characterized by its distributional properties $\mathcal{S}_1(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(X, G_{\mathcal{X}})$ and $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$.

3.3 Model mispecification and quality of myopic decisions

While the previously defined $\mathcal{S}_1(\mathcal{D})$ and $\mathcal{S}_2(\mathcal{D})$ are fully specified before the myopic decision x', here we define two measures aimed at evaluating a possible model mispecificaton implied by $\mathcal{D}$ – under the adopted model fitting procedure (e.g., MLE) – and the quality of the resulting myopic decision.

The first measure is simply the root mean squared error (RMSE) between $f(x)$ and the GP's prediction for the points of the grid $G_{\mathcal{X}}$, that is:

$$RMSE = \sqrt{\frac{1}{|G_{\mathcal{X}}|} \sum_{x_i \in G_{\mathcal{X}}} \Big(f(x_i) - \mu(x_i|\mathcal{D}, \theta)\Big)^2} \tag{7}$$

Clearly, this measure cannot be computed in a real-life setting; it is used in this study just to quantify GP's mispecification in our experiments.

The second measure, denoted by Δ_y, is the improvement/worsening implied by $y' = f(x')$ with respect to y^+, formally:

$$\Delta_y = y^+ - y' \tag{8}$$

where a value of Δ_y lower/higher than zero indicates that y' is worse/better than y^+, while Δ_y is equal to zero when $y' = y^+$. Unlike RMSE, Δ_y can be computed in a real-life BO task, but only after x' is evaluated. Furthermore, also the immediate regret $y' - y^*$ is computed but, similarly to RMSE, only for analytical purposes.

4 Experiments and results

4.1 Setting

In this section, we summarize the relevant details of our experimental setting:

- eight test problems with $d = 1$ and six with $d = 2$ (all search spaces have been preliminary rescaled to $[0, 1]^d$). A graphical representation of the test problems is provided in Appendix;
- three types of design $\mathcal{D}$, generated as follows:
  - (1) from LHS;
  - (2) 50% from LHS and 50% from a neighborhood $N(x^*)$ of the true optimizer to simulate a BO process converging to x^*;
  - (3) 50% from LHS and 50% from the neighborhood $N(\tilde{x})$ of a random point $\tilde{x} \neq x^*$ to simulate a BO process converging to a wrong solution;
- number of observations in $\mathcal{D}$: $n \in \{5d, \lfloor 12.5d \rfloor, 20d\}$;
- GP model: Matérn$_{3/2}$ kernel; fitting via MLE and nugget estimation only in the case of an ill-conditioned kernel matrix; no standardization/scaling of Y;
- Acquisition function: to avoid randomness implied by the method used to optimize the acquisition function, we just pick the best solution on the finite grid $G_{\mathcal{X}}$ (consisting of 10.000 points). Four acquisition functions, all sharing the same fitted GP, are considered:
  - surface-response (SR): $\pi(x|\mathcal{D}, \theta) = \underset{x \in G_{\mathcal{X}}}{\arg\min}\ \mu(x|\mathcal{D})$
  - uncertainty reduction – i.e., maximization of the GP's standard deviation (SD): $\pi(x|\mathcal{D}, \theta) = \underset{x \in G_{\mathcal{X}}}{\arg\max}\ \sigma(x|\mathcal{D})$
  - Expected Improvement (EI): $\pi(x|\mathcal{D}, \theta) = \underset{x \in G_{\mathcal{X}}}{\arg\max}\ EI(x|\mathcal{D})$
  - Lower Confidence Bound (LCB): $\pi(x|\mathcal{D}, \theta) = \underset{x \in G_{\mathcal{X}}}{\arg\min}\ \mu(x|\mathcal{D}) - \beta\sigma(x|\mathcal{D})$, with $\beta = 1$;

To guarantee reproducibility of the experiments, the code is available at the following Github repository:
https://github.com/acandelieri/LION20_Candelieri_Archetti.git

4.2 Results on characterizing designs through $\mathcal{S}_1(\mathcal{D})$ and $\mathcal{S}_2(\mathcal{D})$

In this section, we summarize the most relevant results on the characterization of different designs (i.e., type and size) on a set of diversified test problems. The characterization is given by the two proposed functions $\mathcal{S}_1(\mathcal{D})$ and $\mathcal{S}_2(\mathcal{D})$.

Figure 2 and Figure 3 refer to the 1-dimensional and the 2-dimensional test problems, respectively. Each figure is organized into three charts, one for each type of design, while different shapes and colors of the points refer to the size of $\mathcal{D}$ and the specific test problem, respectively. Since LHS leads to significantly lower values of $\mathcal{W}_2^2(X.G_{\mathcal{X}})$ than the other two design types, the axis ranges of the three charts are different; the dashed gray rectangle helps for the comparison.

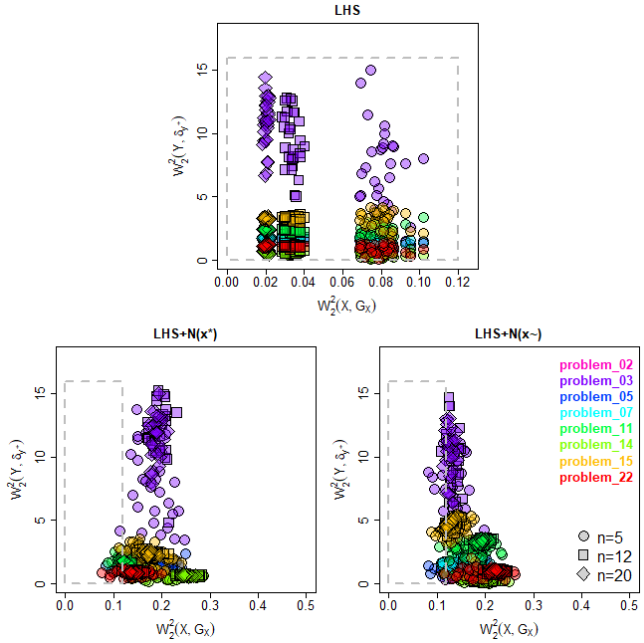

Fig. 2. Thirty designs, with size 5, 12, and 20, on eight 1-dimensional test problems, separately for three types of design: LHS, LHS+$N(x^*)$, and LHS+$N(\tilde{x})$.

As far $\mathcal{S}_1(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(X, G_{\mathcal{X}})$ is concerned, we observe that it clearly depends on n: the higher the value of n, the lower the 2-Wasserstein distance between X and $G_{\mathcal{X}}$. Furthermore, the deviation in the values of $\mathcal{W}_2^2(X, G_{\mathcal{X}})$ also decreases as n increases. Finally, the values of $\mathcal{W}_2^2(X, G_{\mathcal{X}})$ are significantly lower than those obtained with $\mathcal{D}$ from LHS+$N(x^*)$ and LHS+$N(\tilde{x})$.

This "well-shaped" organization of the designs in the $\mathcal{S}_1(\mathcal{D}) \times \mathcal{S}_2(\mathcal{D})$ space – with respect to the x-axis – has been observed in both the 1-dimensional and 2-dimensional test problems, and is in line with the expectation about the LHS technique: the design becomes more uniformly distributed in $\mathcal{X}$ with n increasing.

As far as $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ is concerned, there is no clear relation with n. Indeed, $\mathcal{S}_2(\mathcal{D})$ depends on $f(x)$ together with $\mathcal{D}$ and is uncorrelated with $\mathcal{W}_2^2(X, G_{\mathcal{X}})$, regardless of the type of design. On the other hand, an important insight is that when $f(x)$ is highly "oscillating", with many local optima, the associated $\mathcal{W}_2^2(Y, \delta_{y^+})$ is quite high.

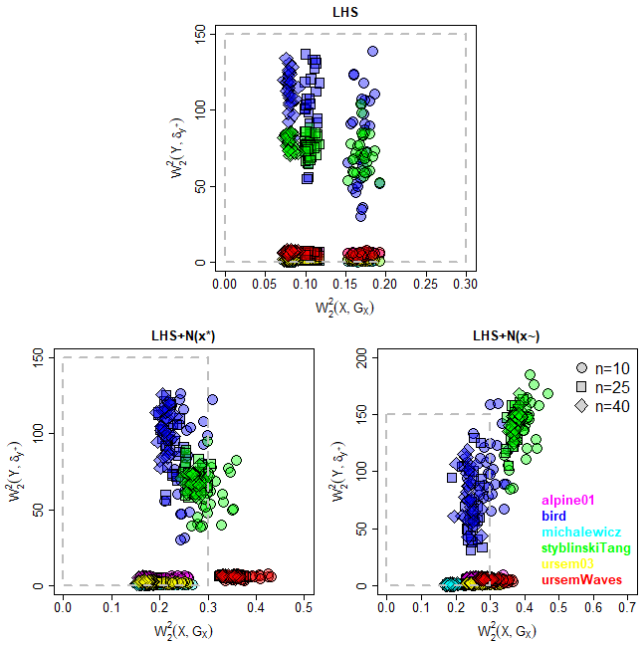

Fig. 3. Thirty designs, with size 10, 25, and 40, on six 2-dimensional test problems, separately for three types of design: LHS, LHS+$N(x^*)$, and LHS+$N(\tilde{x})$.

4.3 Results on the quality of x'

This section summarizes the results related to the quality of the decision x'. Figure 4 shows the differences in terms of Δ_y for the 1-dimensional test problems: a boxplot for each type of design and a box for each of the four acquisition functions. It is important to recall that the acquisition functions share the same GP, so that a possible mispecification would affect all of them. All 1-dimensional problems and all sizes $n = \{5, 12, 20\}$ of the design are considered. Surprisingly, the simple SR is the most robust acquisition function for all three types of design, with LCB and EI providing a more negative Δ_y than SD in some cases.

Figure 5 reports the same kind of analysis, but for the 2-dimensional test problems. The previous considerations are confirmed: SR is the most robust acquisition function, SD does not lead to any improvement, and bot LCB and EI provide more negative values of Δ_y than SR.

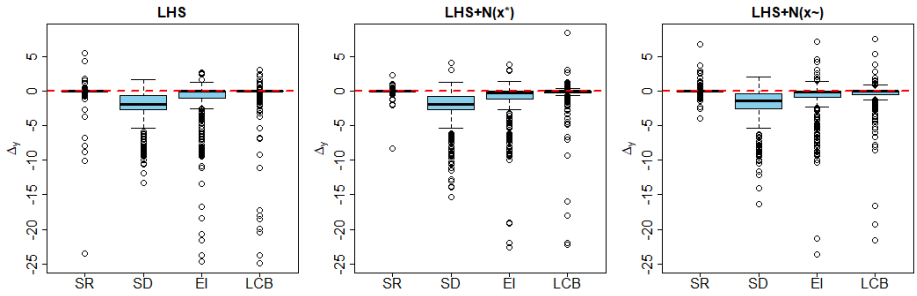

Fig. 4. Boxplot of the Δ_y values on the **1-dimensional problems**: one chart for each type of design. Boxes refer to acquisition functions (i.e., SR: surface response, SD: maximization of the GP's standard deviation, EI: Expected Improvement, and LCB: Lower Confidence Bound). All the acquisition functions share the same GP model.

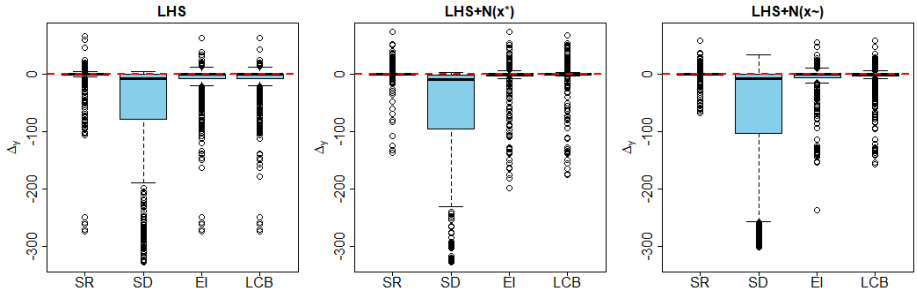

Fig. 5. Boxplot of the Δ_y values on the **2-dimensional problems**: one chart for each type of design. Boxes refer to acquisition functions (i.e., SR: surface response, SD: maximization of the GP's standard deviation, EI: Expected Improvement, and LCB: Lower Confidence Bound). All the acquisition functions share the same GP model.

Even more interesting, SR results in the most robust choice also in terms of immediate regret $y' - y^*$, as depicted in Figure 6 and Figure 7.

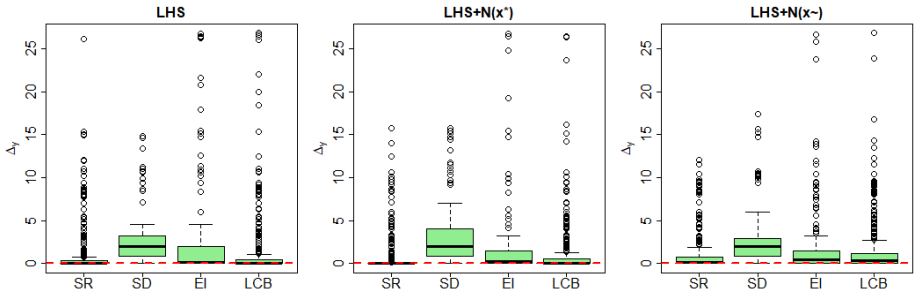

Fig. 6. Boxplot of the immediate regret $y' - y^*$ on the **1-dimensional problems**.

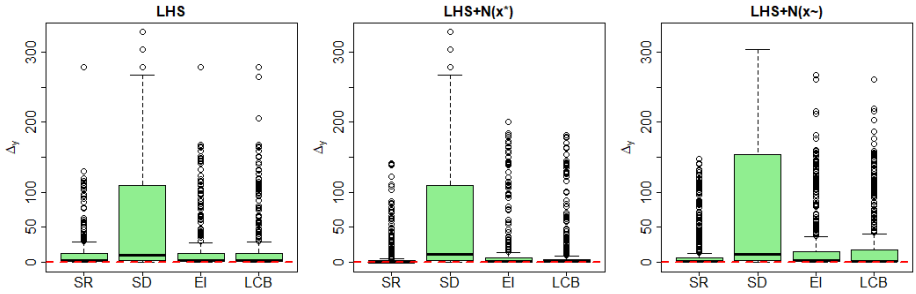

Fig. 7. Boxplot of the immediate regret $y' - y^*$ on the **2-dimensional problems**.

It is interesting to note that this result is consistent with the empirical evidence reported in [3], showing that switching to an exploration-biased acquisition function such as Probability of Improvement, close to the end of the BO process, can outperform BO based on a single no-regret acquisition function.

4.4 Relations between $\mathcal{S}_1(\mathcal{D}), \mathcal{S}_2(\mathcal{D})$, RMSE, and Δ_y

Here, we restrict our attention to the case $n = 20d$, given the empirical evidence suggesting that BO can find an optimal solution within 15-20 time the number of dimensions. From our experiments, we observed that there exists a (linear) correlation between $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ and GP's mispecification as measured by RMSE. Table 1 reports the Pearson's correlation coefficient – and the associated p-value – separately for the type of designs (by row) and the dimensionality of the test problems (by column).

In simpler terms, the fewer/more values in Y are concentrated close to y^+ (i.e., large/small values of $\mathcal{W}_2^2(Y, \delta_{y^+})$) the higher/lower the risk of model mispecification. This holds for the three types of design, with high correlation values, even if slightly different from a design type to another. The most important insight is that LHS cannot avoid the risk of model mispecification [18, 33].

design type	1-dimensional test problems	2-dimensional test problems
LHS	0.9377 (p-value<0.0001)	0.8763 (p-value<0.0001)
LHS+$N(x^*)$	0.9644 (p-value<0.0001)	0.9331 (p-value<0.0001)
LHS+$N(\tilde{x})$	0.8631 (p-value<0.0001)	0.6891 (p-value<0.0001)

Table 1. Pearson's correlation coefficient between $\mathcal{W}_2^2(Y, \delta_{y^+})$ and RMSE, $n = 20$.

Based on the results on the quality of x' for the different acquisition functions, we consider only the SR acquisition function and report the correlations between RMSE and Δ_y^{SR} (Table 2). The most interesting result refers to LHS, with a (statistically significant) negative correlation – even if weak – which has been observed on both the 1-dimensional and 2-dimensional test problems. This holds for LHS, but not for the other two design types, especially for LHS+$N(x^*)$ because, even if the GP can be globally mispecified, it is hardly mispecified close to x^* and thus SR has more chances to improve with respect to y^+.

design type	1-dimensional test problems	2-dimensional test problems
LHS	-0.2526 (p-value<0.0001)	-0.4192 (p-value<0.0001)
LHS+$N(x^*)$	0.2521 (p-value<0.0001)	0.1285 (p-value=0.086)
LHS+$N(\tilde{x})$	0.2538 (p-value<0.0001)	-0.1616 (p-value=0.010)

Table 2. Pearson's correlation coefficient between RMSE and Δ_y^{SR}, $n = 20d$.

The most important result comes from an obvious question: since $\mathcal{W}_2^2(Y, \delta_{y^+})$ can be calculated before x' and is correlated with RMSE, and since RMSE is inversely correlated with Δ_y^{SR}, is there a correlation between $\mathcal{W}_2^2(Y, \delta_{y^+})$ and Δ_y^{SR}? Luckily, the answer to this question is *"yes, for LHS and SR"*: Table 3 reports the correlation coefficient between $\mathcal{W}_2^2(Y, \delta_{y^+})$ and Δ_y^{SR}. In the case of an LHS design, we again observe a (statistically significant) negative correlation, this time between $\mathcal{W}_2^2(Y, \delta_{y^+})$ and Δ_y^{SR}, on both 1-dimensional and 2-dimensional test problems. For the other two types of design, the correlation is positive but statistically significant only for the 1-dimensional problems.

design type	1-dimensional test problems	2-dimensional test problems
LHS	-0.2087 (p-value<0.0001)	-0.2569 (p-value<0.0001)
LHS+$N(x^*)$	0.3009 (p-value<0.0001)	0.1280 (p-value=0.087)
LHS+$N(\tilde{x})$	0.2295 (p-value<0.0001)	0.0556 (p-value=0.458)

Table 3. Pearson's correlation coefficient between $\mathcal{W}_2^2(Y, \delta_{y^+})$ and Δ_y^{SR}, $n = 20d$.

Finally, we have always observed a statistically significant linear positive correlation between $\mathcal{W}_2^2(Y, \delta_{y^+})$ and immediate regret, with a smaller correlation coefficient for SR and LCB and a larger one for EI and SD. This further remarks that high values of $\mathcal{W}_2^2(Y, \delta_{y^+})$ indicate a more complicated to be optimized $f(x)$.

5 Discussions on results and opportunities

We summarize the most important insights from our empirical study:

1. the risk of model mispecification increases with $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ and regardless of $\mathcal{S}_1(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(X, G_{\mathcal{X}})$. Thus, even if the design $\mathcal{D}$ in a generic BO iteration is close to LHS, this does not prevent model mispecification.

2. if $\mathcal{D}$ is close to LHS and $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ is "small", then a simple SR acquisition function can provide a high Δ_y^{SR}, which is hopefully positive. The main issue – at the moment – is to quantify what "small" means.
3. if $\mathcal{D}$ is a design associated with a BO task that is converging towards a solution – be it x^* or not – and $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ is "large", again a simple SR acquisition function can offer the chance of high Δ_y^{SR}.
4. the remaining two cases are the most critical: (a) $\mathcal{D}$ is close to LHS and $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ is "large", or (b) $\mathcal{D}$ is a design converging to a solution and $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ is "small". In these cases, no one of the acquisition functions considered has a relevant chance of leading to $\Delta_y \geq 0$, but due to two completely different reasons: $f(x)$ is "complicated" (a) and the BO is (or is going to be) stuck in a local/global optimum (b).

The last point of this list represents the first relevant opportunity that can follow from this paper. For the two cases described in 4., a new policy $\hat{\pi}(\mathcal{D}|\theta)$ should be defined with the aim of resorting $\mathcal{D}$ to one of the two more favorable situations in 2. and 3. — i.e., "moving" $\mathcal{D}$ towards LHS if $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$ is "small" or forcing convergence towards a solution if it is "large".

Thus, a new generation acquisition function should switch between $\hat{\pi}(\mathcal{D}|\theta)$ and $\pi(\mathcal{D}|\theta)$ (i.e., SR) jointly depending on the values of $\mathcal{S}_1(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(X, G_{\mathcal{X}})$ and $\mathcal{S}_2(\mathcal{D}) \stackrel{\text{def}}{=} \mathcal{W}_2^2(Y, \delta_{y^+})$, **before any model fitting**.

6 Conclusions

This paper presents a new characterization of a design based on its distributional properties measured through the Wasserstein distance. Empirical analysis of myopic decisions induced by common basic acquisition functions led to the observation of a relationship between the proposed characterization of the design, immediate regret, and improvement with respect to the best observed value in the design. This relation suggests a way for dynamically changing the behavior of the acquisition step, more principled with respect to a simple switch towards more/less exploration. Indeed, instead of just increasing exploration, queries should be selected to improve distributional properties of the design up to an optimal myopic decision can be made. It would also be interesting to investigate whether the distributional properties of the design could be related to stopping criteria, such as the Gittin's index proposed in [26].

Although the empirical results of this study are interesting, limitations regard a suitable quantification of what "small" and "large" $\mathcal{W}_2^2(Y, \delta_{y^+})$ really mean. Currently, this is the main challenge that our research is now focusing on in order to design a new generation of acquisition functions in BO.

Acknowledgments. This work has benefitted from the first author's participation in Dagstuhl Seminar 25451 Bayesian Optimisation (Nov 02 – Nov 07, 2025).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alcobaça, E., de Carvalho, A.C.: A literature review on automated machine learning. Artificial Intelligence Review **59**(1), 1–39 (2026)
2. Archetti, F., Candelieri, A.: Bayesian optimization and data science, vol. 849. Springer (2019)
3. Benjamins, C., Raponi, E., Jankovic, A., van der Blom, K., Santoni, M.L., Lindauer, M., Doerr, C.: Pi is back! switching acquisition functions in bayesian optimization. arXiv preprint arXiv:2211.01455 (2022)
4. Berkenkamp, F., Schoellig, A.P., Krause, A.: No-regret bayesian optimization with unknown hyperparameters. arXiv:1901.03357 (2019)
5. Bogunovic, I., Krause, A.: Misspecified gaussian process bandit optimization. Advances in Neural Information Processing Systems **34**, 3004–3015 (2021)
6. Bossek, J., Doerr, C., Kerschke, P.: Initial design strategies and their effects on sequential model-based optimization: an exploratory case study based on bbob. In: Proceedings of the 2020 genetic and evolutionary computation conference. pp. 778–786 (2020)
7. Bull, A.D.: Convergence rates of efficient global optimization algorithms. Journal of Machine Learning Research **12**(10) (2011)
8. Candelieri, A.: Mastering the exploration-exploitation trade-off in bayesian optimization. arXiv preprint arXiv:2305.08624 (2023)
9. Candelieri, A., Signori, E.: Mle-free gaussian process based bayesian optimization. In: International Conference on Learning and Intelligent Optimization. pp. 81–95. Springer (2024)
10. Cheng, X., Lu, J., Tan, Y., Xie, Y.: Convergence of flow-based generative models via proximal gradient descent in wasserstein space. IEEE Transactions on Information Theory (2024)
11. Elshewey, A.M., El-Bakry, H.M., Osman, A.M.: Enhancing water potability classification using a hybrid gru with lstm deep learning models based on relief algorithm and bayesian optimization. Evolving Systems **17**(1), 10 (2026)
12. Frazier, P.I., Powell, W.B., Dayanik, S.: A knowledge-gradient policy for sequential information collection. SIAM Journal on Control and Optimization **47**(5), 2410–2439 (2008)
13. Gao, X., Nguyen, H.M., Zhu, L.: Wasserstein convergence guarantees for a general class of score-based generative models. Journal of Machine Learning Research **26**(43), 1–54 (2025)
14. Garnett, R.: Bayesian optimization. Cambridge University Press (2023)
15. Gramacy, R.B.: Surrogates: Gaussian process modeling, design, and optimization for the applied sciences. Chapman and Hall/CRC (2020)
16. Hernández-Lobato, J.M., Hoffman, M.W., Ghahramani, Z.: Predictive entropy search for efficient global optimization of black-box functions. Advances in neural information processing systems **27** (2014)
17. Hiçyılmaz, M.: Bayesian optimization-enhanced vision transformer for damage detection in steel truss structures using continuous wavelet transform analysis. In: Structures. vol. 85, p. 111059. Elsevier (2026)
18. Hvarfner, C., Eriksson, D., Bakshy, E., Balandat, M.: Informed initialization for bayesian optimization and active learning. arXiv preprint arXiv:2510.23681 (2025)
19. Kim, H., Liu, C., Chen, Y.: Bayesian optimization with inexact acquisition: Is random grid search sufficient? arXiv preprint arXiv:2506.11831 (2025)

20. Kim, H., Sanz-Alonso, D.: Enhancing gaussian process surrogates for optimization and posterior approximation via random exploration. SIAM/ASA Journal on Uncertainty Quantification **13**(3), 1054–1084 (2025)
21. Li, G., Wang, Y., Kar, S., Jin, X.: Bayesian optimization with active constraint learning for advanced manufacturing process design. IISE Transactions **58**(3), 257–271 (2026)
22. Montesuma, E.F., Mboula, F.M.N., Souloumiac, A.: Recent advances in optimal transport for machine learning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
23. Park, J.H., Cheon, M., Wi, J., Koh, D.Y.: Boost: Bayesian optimization with optimal kernel and acquisition function selection technique. arXiv preprint arXiv:2508.02332 (2025)
24. Peyré, G., Cuturi, M.: Computational optimal transport: With applications to data science. Now Foundations and Trends (2019)
25. Santambrogio, F.: Optimal transport for applied mathematicians (2015)
26. Scully, Z., Terenin, A.: The gittins index: A design principle for decision making under uncertainty. In: Tutorials in Operations Research: Advances in Analytics and Operations Research: Improving Decisions to Secure the Future, pp. 28–70. INFORMS (2025)
27. Siska, M., Pajak, E., Rosenthal, K., del Rio Chanona, A., von Lieres, E., Helleckes, L.M.: A guide to bayesian optimization in bioprocess engineering. Biotechnology and Bioengineering (2026)
28. Srinivas, N., Krause, A., Kakade, S.M., Seeger, M.W.: Information-theoretic regret bounds for gaussian process optimization in the bandit setting. IEEE transactions on information theory **58**(5), 3250–3265 (2012)
29. Wabersich, K.P., Toussaint, M.: Advancing bayesian optimization: The mixed-global-local (mgl) kernel and length-scale cool down. arXiv:1612.03117 (2016)
30. Wang, Z., de Freitas, N.: Theoretical analysis of bayesian optimisation with unknown gaussian process hyper-parameters. arXiv preprint arXiv:1406.7758 (2014)
31. Wang, Z., Hutter, F., Zoghi, M., Matheson, D., De Feitas, N.: Bayesian optimization in a billion dimensions via random embeddings. Journal of Artificial Intelligence Research **55**, 361–387 (2016)
32. Williams, C.K., Rasmussen, C.E.: Gaussian processes for machine learning, vol. 2. MIT press Cambridge, MA (2006)
33. Zhang, B., Cole, D.A., Gramacy, R.B.: Distance-distributed design for gaussian process surrogates. Technometrics **63**(1), 40–52 (2021)

7 Appendix

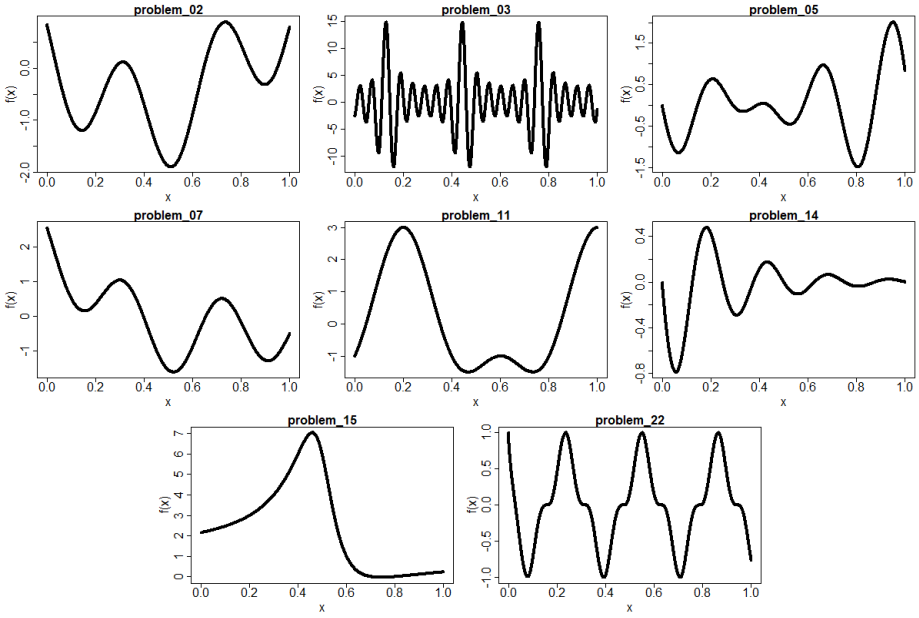

Fig. 8. 1-dimensional test problems. Original search spaces rescaled in $[0, 1]$

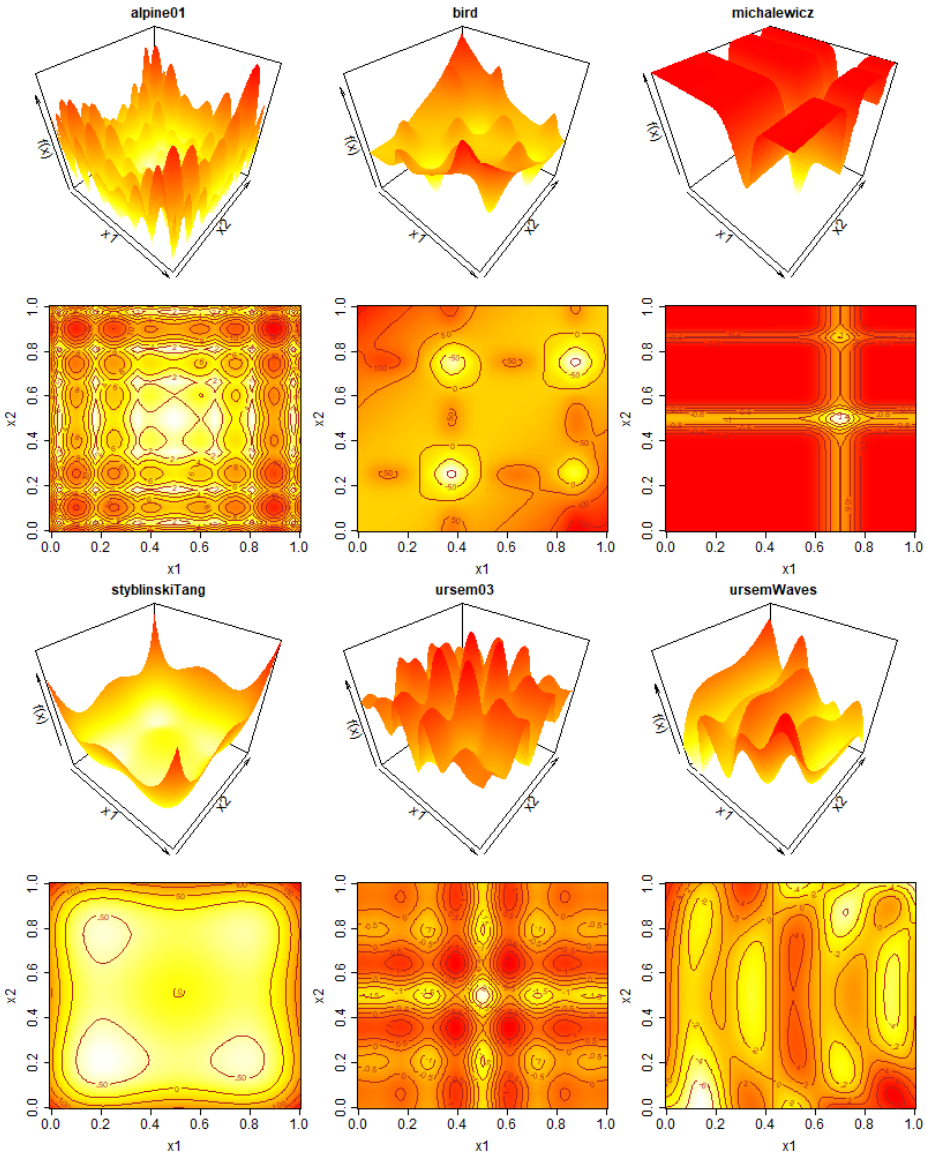

Fig. 9. 2-dimensional test problems (3D representation and countourplot for each test problem). Original search spaces rescaled in $[0, 1] \times [0, 1]$