ECAI 2024
U. Endriss et al. (Eds.)

doi:10.3233/FAIA240765

Global Optimisation of Black-Box Functions with Generative Models in the Wasserstein Space

Tigran Ramazyan[a,*], **Mikhail Hushchyn**[a] **and Denis Derkach**[a]

[a]HSE University, Moscow, Russia
ORCID (Tigran Ramazyan): https://orcid.org/0000-0002-8770-0637, ORCID (Mikhail Hushchyn): https://orcid.org/0000-0002-8894-6292, ORCID (Denis Derkach): https://orcid.org/0000-0001-5871-0628

Abstract. We propose a new uncertainty estimator for gradient-free optimisation of black-box simulators using deep generative surrogate models. Optimisation of these simulators is especially challenging for stochastic simulators and higher dimensions. To address these issues, we utilise a deep generative surrogate approach to model the black box response for the entire parameter space. We then leverage this knowledge to estimate the proposed uncertainty based on the Wasserstein distance - the Wasserstein uncertainty. This approach is employed in a posterior agnostic gradient-free optimisation algorithm that minimises regret over the entire parameter space. A series of tests were conducted to demonstrate that our method is more robust to the shape of both the black box function and the stochastic response of the black box than state-of-the-art methods, such as efficient global optimisation with a deep Gaussian process surrogate.

1 Introduction

Simulation of real-world experiments is key to scientific discoveries and engineering solutions. Such techniques use parameters describing configurations and architectures of the system. A common challenge is to find the optimal configuration for some objective, e.g., such that it maximises efficiency or minimises costs and overhead [21].

To generalise the problem, we examine the case of stochastic simulators. A single call of such simulators is often computationally expensive, which is the case for many domains, especially when the simulator is Monte Carlo-based. For instance simulators modelling particle scatterings [57], molecular dynamics [1], and radiation transport [6]. The problem is also often referred to as design optimisation [51].

In this case, the simulator is treated as a black box experiment. This means that observations come from an unknown and likely difficult-to-estimate function. Using a surrogate model for black-box optimisation (BBO) is an established technique [15], as BBO has a rich history of both gradient and gradient-free methods [5], most of which come from tackling problems that arise in physics, chemistry and engineering.

Although gradient-based optimisation has proven to work well for differentiable objective functions, e.g. [12, 19, 17], in practice nondifferentiable objective functions appear often in applications [2]. For instance, when one addresses parameter optimisation of a system represented by a Monte Carlo simulator which provides data from an intractable likelihood [34]. In addition to that, simulation samples used as input to train machine learning models usually consist of several samples representing disconnected regions of the parameter space. In this case, surrogate modelling and optimisation direction estimation are particularly challenging due to high uncertainty, thus making gradient-free optimisation much more preferable than gradient methods.

Common gradient-free optimisation methods, e.g., Bayesian optimisation, rely on uncertainty estimation over the parameter space to predict the potential minimum. We need to leverage the knowledge of the stochastic simulator response, hence an uncertainty estimator that would account for the complexity of both the response shape and the geometry of the objective function is required. We propose to minimise regret as in the lower confidence bound approach, but the variance of the predictive posterior is moved from $(\mathbb{R}^d, L_2)$ metric space to $(\mathcal{P}_2, \mathbb{W}_2)$ metric space. That is done by modelling the stochastic simulator response with a deep generative model (DGM) as the surrogate [50, 47].

A well-trained DGM surrogate can properly replicate the simulator response. The use of DGMs as surrogates provides the opportunity to work in agnostic predictive posterior settings [25]. In addition, a single run of a DGM surrogate is orders of magnitude faster than a simulator call, which is exactly the advantage of neural network surrogates over simulators.

We review related works in Section 2. Section 3 describes the proposed approach. In Section 4 we examine our approach on a set of experiments and compare our results with selected baselines. The results of the research are discussed and concluded in Section 5. The appendices can be found in Ref. [53].

2 Related Work

Black-box optimisation encompasses numerous methods. Common methods include conjugate gradient [55], Quasi-Newton methods, such as BFGS [49], trust-region methods [13, 14], and multi-armed bandit approaches [24]. In particular, Ref. [56, 10, 33, 38] explore deep generative models for black-box optimisation. [10, 33] focus on the inference of posterior parameters using given observations, and [56] uses direct gradient optimisation of the objective function using a local generative surrogate model. In addition, [30] and [38] exploit the architecture of generative models used for optimisation - [30] optimises in a latent space that a DGM learns, and [38] employ

* Corresponding Author. Email: tramazyan@hse.ru.
Original code: https://github.com/ramazyant/wu-go
Python package: https://github.com/hse-cs/waggon

the invertibility of diffusion models to solve the inverse optimisation problem. Further, [26] aims to find optimal parameters of a DGM that approximate a fixed black box metric the best.

In practice, nondifferentiable objective functions are often needed. In this case Bayesian optimisation (BO) [58, 54, 59], genetic algorithms [51, 8], or numerical differentiation [62] are preffered. BO is typically equipped with a Gaussian process (GP), which is defined by a mean function and a kernel function that describes the shape of the covariance. BO with a GP surrogate requires covariance matrix inversion with $O(n^3)$ cost in terms of the number of observations, which is challenging for a large number of responses and in higher dimensions. To make BO scalable [18, 20, 28, 66] consider a low-dimensional linear subspace and decompose it into subsets of dimensions, i.e., some structural assumptions are required that might not hold. In addition, GP requires a proper choice of kernel is crucial. BO may need the construction of new kernels [31]. [22] proposes a greedy approach for automatically building a kernel by combining basic kernels such as linear, exponential, periodic, etc., through kernel summation and multiplication.

BO is not bound to use GP. For example, deep Gaussian processes (DGP) [16, 35, 29] attempt to resolve the issue of finding the best kernel for a GP. That is done by stacking GPs in a hierarchical structure as Perceptrons in a multilayer perceptron, but the number of variational parameters to be learnt by DGPs increases linearly with the number of data points, which is infeasible for stochastic black-box optimisation, and they have the same matrix inverting issue as GPs, which limits their scalability.

Other surrogate models that are reasonable to consider for BO are Bayesian neural networks (BNN) [37, 60] and ensemble models [64]. BNN directly quantify uncertainty in their predictions and generalise better than other neural networks. However, their convergence may be slower as each weight is taken from a distribution. This also means that BNN might require much more data than GPs to yield good predictions.

On the other hand, an intuitive approach for uncertainty quantification is using an ensemble surrogate model, e.g., an adversarial deep ensemble (DE) [41, 64]. Single predictors of the ensemble are expected to agree on their predictions over observed regions of the feature space, i.e., where data are given and so the uncertainty is low and vice versa. The further these single predictors get from known regions of the feature space, the greater the discrepancy in their predictions.

We chose these surrogates as in this work, we focus on approaches that take advantage of the uncertainty incorporated in the predictive posteriors of the surrogate. GP, DGP, and BNN have a distribution as their outputs, and DE comprises one through its single predictors.

3 Method

3.1 Problem Statement

In design optimisation, one aims to find a configuration of the system θ in the search space $\Theta \subseteq \mathbb{R}^m$, which minimises the expected cost $\mathbb{E}^\mu[f(\theta, x)]$ of the black box. f is a real-valued function $f : \Theta \times \mathbb{R}^d \to \mathbb{R}$, where $x \sim \mu \in \mathcal{P}(\mathbb{R}^d)$, and $\mathcal{P}(\mathbb{R}^d)$ is the space of Borel probability measures in $\mathbb{R}^d$. This means that f is pointwise observable, defined over a set of parameters Θ. In this work, we take continuous Θ. We assume that f is bounded from below and Lipschitz continuous. The set Θ is convex, explicit, and non-relaxable, i.e., it does not rely on estimates and f cannot be computed outside the feasible region. The assumption of finite first and second moments of μ is reasonable. It is denoted as $\mu \in \mathcal{P}_2$, where $\mathcal{P}_2$ is the set of probability measures with finite second moment.

As mentioned earlier, in black box optimisation simulated samples typically represent disjoint regions of the parameter space. This means that there is little or no information about most of the search space. In addition, the probability measure μ is unknown. Thus, following [42], we consider the problem of optimisation under uncertainty - a problem tackled via distributionally robust optimisation. The problem of uncertainty estimation amounts to evaluating the worst-case risk of the objective function over an ambiguity set $\mathcal{P} \subset \mathcal{P}_2(\mathbb{R}^d)$:

$$\sup_{\mu \in \mathcal{P}} \mathbb{E}^\mu[f(\theta, x)]. \tag{1}$$

Generally, ambiguity sets are used to ensure statistical performance and computational feasibility [23]. In this work, we use the construction of considered ambiguity sets, Wasserstein balls to be precise, and the topology they induce to quantify uncertainty. The Wasserstein metric was studied for distributionally robust optimisation in Ref. [27, 9, 39].

Thus, we find the optimal configuration by solving the following optimisation problem:

$$\inf_{\theta \in \Theta} \sup_{\mu \in \mathcal{P}} \mathbb{E}^\mu[f(\theta, x)]. \tag{2}$$

The optimal design is found by selecting a point of potential minimum among a set of explored candidates and calling the black box for that point. Consequently, filling the search space with ground truths and exploiting the obtained information increases the confidence of the predicted optimal configuration.

3.2 Deep Generative Models for Surrogate Modelling

Having a generator model that would map a simple distribution, $\mathcal{Z}$, to a complex and unknown distribution, $p(\mu)$, is desired in many settings as it allows for the generation of samples from the intractable data space. For our purposes, conditional deep generative models (DGMs) are used to model the stochastic response of the simulator. This is particularly advantageous in exploring unknown regions of the parameter space. By parametrising the response space we exploit deep generative models in showing that they can be used for both approximating the objective function and modelling the simulator for the whole feature space and hence also computing the proposed uncertainty.

The goal of DGM is to obtain a generator $G : \mathbb{R}^s \to \mathbb{R}^d$ that maps a given distribution, e.g., a Gaussian, $\mathcal{Z}$ supported in $\mathbb{R}^s$ to a data distribution $p(\mu)$ supported in $\mathbb{R}^d$.

According to the problem statement, each sample is associated with a specific configuration, i.e., a vector of parameters θ, so we consider conditional DGMs. Then the goal is to obtain $G : \mathbb{R}^s \times \Theta \to \mathbb{R}^d$ such that distributions $G(\mathcal{Z}; \theta)$ and $p(\mu; \theta)$ match for each $\theta \in \Theta$. Since $\mathcal{Z}$ and $p(\mu)$ are independent, we get $G(\mathcal{Z}; \theta) \sim p(\mu; \theta)$. For simplicity of notation we denote $G(\mathcal{Z}; \theta)$ as $G(\theta)$.

Thus, we explore the search space by training a deep generative surrogate G on the set of ground truths $\mathcal{M}$. For each value $\theta \in \Theta$, we can generate the corresponding response of the simulator $\nu(\theta) = G(\theta)$, i.e., ν is the predictive distribution at θ.

3.3 Wasserstein Uncertainty

In the usual Bayesian optimisation formulation, the epistemic uncertainty estimator is the variance of the predictive posterior distribution. Since the simulator is stochastic, we aim to have an uncertainty estimator that would account for the intricacies of both sample density

and the geometry of the objective function. Hence, we propose the Wasserstein uncertainty - a Wasserstein distance-based uncertainty estimator that encapsulates the difference between known and predicted samples in both the distribution and the L_2 space.

Definition 1. *2-Wasserstein distance, denoted by $\mathbb{W}$, is the optimal transport cost between two probability distributions $\mu, \nu \in \mathcal{P}_2$ with L_2^2 cost function:*

$$\mathbb{W}(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int |x-y|^2 d\pi(x,y) \right)^{\frac{1}{2}}. \tag{3}$$

$\Pi(\mu,\nu)$ is the set of couplings of μ and ν. The Wasserstein distance defines a symmetric metric on $\mathcal{P}_2$.

Optimisation over an ambiguity set allows us to describe the regions of the parameter space with certain levels of uncertainty with respect to given ground truths. Ambiguity sets are usually defined to ensure statistical performance guarantees and computational tractability [48, 44]. A popular choice of ambiguity sets are Wasserstein balls [65, 52].

Definition 2. *Wasserstein ball of radius ε and centred at ν, denoted by $\overline{B}_\varepsilon(\nu)$, is the closure with respect to the topology induced by the Wasserstein distance as follows:*

$$\overline{B}_\varepsilon(\nu) := \{\mu \in \mathcal{P}_2(\mathbb{R}^d) : \mathbb{W}(\mu,\nu) \leq \varepsilon\}. \tag{4}$$

This significantly eases the search for a potential optimum, as for a candidate ν its uncertainty amounts to:

$$F(\mu) = \sup_{\mu \in \overline{B}_\varepsilon(\nu)} \mathbb{E}^\mu[f]. \tag{5}$$

The uncertainty of ν is associated with the closest known response, i.e., for any prediction ν we consider $\tilde{\mu} \in \mathcal{M}$ such that $\tilde{\mu} = \arg\inf_{\mu \in \mathcal{M}} \mathbb{W}(\mu,\nu)$. From the uncertainty estimation perspective, we can be as certain about ν as about the closest to it $\tilde{\mu} \in \mathcal{M}$. This also agrees with [39], where a decision under uncertainty is modelled with a real-valued loss function, and the risk of the optimal decision is the least risky admissible loss.

Consider $\overline{B}_\varepsilon(\nu)$ a Wasserstein ball centred at ν with radius $\varepsilon = \mathbb{W}(\nu,\tilde{\mu})$, i.e., such that $\tilde{\mu}$ is located at the boundary of the ambiguity set. Considering any other radius of $\overline{B}_\varepsilon(\nu)$ would yield either overestimation or underestimation of the uncertainty of ν.

In fact, for an approximation of f of the form $\hat{f} = \langle w, x \rangle$, e.g., a neural approximator, the supremum is attained at the boundary of the Wasserstein ball.

Lemma 1. *(Existence of a minimum at the boundary, [42]) There exists a worst-case probability measure μ^* attaining the supremum (5) such that $\mathbb{W}(\nu,\mu^*) = \varepsilon$.*

In general, finding μ^* is a task of its own, which would add computational costs to the optimisation problem. And ν would be equidistant from the closest known response, $\tilde{\mu}$, and the worst-case risk of the objective function, μ^*. Thus, for a predicted ν, we quantify its uncertainty as:

$$\sigma_{\mathbb{W}}(\theta) = \inf_{\mu \in \mathcal{M}} \mathbb{W}(\mu,\nu). \tag{6}$$

For $\sigma_{\mathbb{W}}$ to be a viable uncertainty estimator, it must be zero for known responses and vice versa. Lemma 2 follows from the fact that $\mathbb{W}$ is a distance. The proof is provided in Appendix A [53].

Lemma 2. *For a predicted random variable ν and a set of ground truths $\mathcal{M}$*

$$\sigma_{\mathbb{W}}(\theta) = 0 \Leftrightarrow \exists \mu \in \mathcal{M} : \mu = \nu. \tag{7}$$

Calculating the true Wasserstein distance would add computational costs that would not marginally benefit the algorithm's performance. It is typically approximated with the corresponding L_p norm. In all our experiments we consider only univariate real-valued random variables. According to Ref. [63], L_2 and Energy distances are equivalent in this case.

$$D^2(\mu,\nu) = 2\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\|, \tag{8}$$

where $X, X' \sim F_\mu$, $Y, Y' \sim F_\nu$, and F_μ, F_ν are the CDFs of μ and ν respectively. We estimate the Wasserstein uncertainty, Eq. 6, for a predictive posterior $G(\theta)$ via Energy distance as follows:

$$\hat{\sigma}_{\mathbb{W}}(\theta) = \min_{\mu \in \mathcal{M}} D(\mu, G(\theta)). \tag{9}$$

3.4 Optimisation

Bayesian optimisation (BO) [46] is perhaps the most common approach for global optimisation. It is based on acquisition functions. In BO the acquisition function is repeatedly applied until convergence.

The authors of the efficient global optimisation (EGO) algorithm [36] proposed the expected improvement (EI) heuristic to fully exploit predictive uncertainty. The potential for optimisation is linked via a measure of improvement - a random variable defined for an input $\theta \in \Theta$ as

$$I(\theta) = \max\{0, f_{min} - \nu(\theta)\}, \tag{10}$$

where f_{min} is the best objective value obtained so far, and $\nu(\theta)$ is the predictive distribution of the fitted model at θ. If $\nu(\theta)$ has a non-zero probability of taking any value on the real line, then $I(\theta)$ has a nonzero probability of being positive at any θ.

To target potentials for large improvements more precisely, the EGO algorithm aims to maximise expected improvement $EI(\theta) = \mathbb{E}[I(\theta)]$. In predictive posterior agnostic settings EI is calculated using the MC approximation.

$$EI(\theta) \approx \frac{1}{M} \sum_{j=1}^{M} \max\{0, f_{min} - x_j\}, \tag{11}$$

where $\{x_j\}_{j=1}^M, x_j \sim \nu(\theta)$. As $M \to \infty$ the approximation becomes exact. If $\nu(\theta)$ is Gaussian with mean $\mu(\theta)$ and variance $\sigma^2(\theta)$, e.g., as in Gaussian Processes, then EI accepts the following closed form.

$$EI(\theta) = \sigma(\theta) \left[z(\theta) \cdot \Phi\left(z(\theta)\right) + \phi\left(z(\theta)\right) \right], \tag{12}$$

where $z(\theta) = \frac{f_{min} - \mu(\theta)}{\sigma(\theta)}$ and Φ and ϕ are Gaussian CDF and PDF respectively.

Another common BO approach is lower confidence bound (LCB) [40]. Its acquisition function is regret defined as follows:

$$\mathcal{R}(\theta) = \mu(\theta) - \kappa \cdot \sigma(\theta). \tag{13}$$

It is a linear combination of exploitation, $\mu(\theta)$, and exploration, $\sigma(\theta)$. The trade-off between the two is controlled via the hyperparameter κ. Smaller κ yields more exploitation, and larger values of κ

yield more exploration of high-variance responses, where uncertainty is higher, i.e., where the black box is more unknown.

We omit any assumptions on the functional form of the posterior distribution and the computational and sampling challenges of Monte Carlo estimates. LCB can adjust exploration and exploitation, and EGO considers the predictive posterior distribution explicitly. We propose to combine both approaches and account for the stochasticity not explicitly in the objective as in EGO, but implicitly as in LCB. Thus, using the proposed in Eq. 6 Wasserstein uncertainty, we suggest the following formulation of regret:

$$\mathcal{R}_{\mathbb{W}}(\theta) = \mu(\theta) - \kappa \cdot \sigma_{\mathbb{W}}(\theta) \tag{14}$$

This also agrees with the optimal solution of the supremum (5) for a neural approximator (see Proposition 4.4 in Ref. [42]).

Optimisation of acquisition functions such as EI can be done numerically. The proposed acquisition function, Eq. 14, is non-convex, similarly to LCB, and non-differentiable, which complicates its optimisation. We do it via direct search over a large set of candidate points. It simplifies the search for the next best set of parameters but limits the precision of the optimisation algorithm. For consistency of comparison, we use the direct search approach for all methods. This issue requires further studies.

3.5 Convergence

Under some theoretical assumptions, convergence rates of Bayesian optimisation can be guaranteed. The authors of [11] provide convergence rates for expected improvement algorithms. However, since in practice priors are estimated from data, standard estimators may not hold up to these convergence rates, and thus they need to be altered.

One may search for a more efficient method with improved average-case performance that still ensures reliable convergence. The challenge in developing such a method lies in balancing exploration and exploitation, as with LCB or WU-GO. Exploiting the data specifically in regions of the search space where the black-box function is known to be sub-optimal may help us find the optimum quickly. However, without exploring all regions of the search space, we might never find the optimal design or be confident in the optimality of the proposed solution [45]. This is further discussed in the ablation study reported in Appendix C [53].

In general, to converge to the exact solution, as with any other gradient-free optimisation approach, it may take an infinite number of simulator calls. One might deduce a worst-case scenario, but it may require an unfeasibly large number of simulator calls. Since such scenarios are unlikely in practice and the number of simulations is limited, it makes sense to relax the guarantees to address practical constraints while still feasible solutions.

3.6 Algorithm

The proposed algorithm is summarised in Algorithm 1. Our algorithm requires a generative surrogate model and chooses the next point minimising $\mathcal{R}_{\mathbb{W}}$ from Eq. 14.

As stated in Algorithm 1, the black-box function is evaluated over a set of candidate values represented by a grid $\tilde{\Theta} \subset \Theta$ representing the search space. As discussed in Section 3.4, This is a numerical limitation, as the Wasserstein uncertainty is not differentiable. On the other hand, this allows us to easily consider configurations and subspaces of interest, and define stopping criteria, i.e., an ε-solution criterion.

Algorithm 1 Wasserstein Uncertainty Global Optimisation (WU-GO)

Input: Ground truths $\mathcal{M}$, generator G, candidates $\tilde{\Theta}$, parameter κ
Output: Optimal configuration $\hat{\theta}^*$

while stopping criteria are not met **do**
  Fit G on $\mathcal{M}$
  Approximate f: $\hat{f}(\theta) = \mathbb{E}[G(\theta)] \approx \frac{1}{n}\sum_{j=1}^{n} x_j, x_j \sim G(\theta)$
  Estimate $\sigma_{\mathbb{W}}$: $\hat{\sigma}_{\mathbb{W}}(\theta) = \min_{\mu\in\mathcal{M}} D(\mu, G(\theta))$
  Predict $\hat{\theta} = \arg\min_{\theta\in\tilde{\Theta}}\{\hat{f}(\theta) - \kappa \cdot \hat{\sigma}_{\mathbb{W}}(\theta)\}$
  Call simulator for $\hat{\theta} : \hat{\mu}$
  $\mathcal{M} = \mathcal{M} \cup \{\hat{\mu}\}$
end while

To select the value of κ, we conduct an ablation study. The value $\kappa = 2$ is the best among the considered ones (see Appendix C [53] for more details).

3.7 Surrogate Model and Baselines

WU-GO is not bound to any specific generative surrogate model. In this work, we take a conditional WGAN-GP [32]. Using GANs is advantageous for faster sampling, i.e., faster exploration of the search space. Gradient penalty provides a better objective function approximation, and the Wasserstein uncertainty could be directly estimated from the surrogate model, which would be highly efficient, especially for multidimensional outputs. However, [43] and [61] argue that for the true Wasserstein distance, one should alter the training of the WGAN.

As baselines, we take EGO and LCB with Gaussian posterior. As surrogates for these algorithms, we take Gaussian process regression, Deep Gaussian Processes, Bayesian neural networks, and adversarial deep ensembles. An effort was made to make all models as similar as possible to achieve more consistent results. Appendix B [53] contains model parameters and experiment settings for reproducibility.

4 Experiments

We evaluate WU-GO performance using eight experiments. A real working example of a high energy physics detector and seven experiments based on commonly used optimisation objective functions. In each synthetic test, we change the settings of the experiment to examine the models in various scenarios. A short description is provided below and for a full description see Appendix D [53].

Table 1. Experiments.

Black Box Function	N	n	Θ
Three Hump Camel	4	10^2	$[-5, 5]^2$
Ackley	4	10^2	$[-5, 5]^2$
Lévi	4	10^2	$[-4, 6]^2$
Himmelblau	4	10^1	$[-5, 5]^2$
Rosenbrock	121	10^2	$[-2, 2]^8$
Rosenbrock	25	10^2	$[-2, 2]^{20}$
Styblinki-Tang	25	10^2	$[-5, 5]^{20}$

Similarly to experiments of [7], all instances are stochastic. Each experiment is defined by the black-box response shape. Namely, $\mathcal{N}(f(\theta), 10^{-2})$, where $f(\theta)$ is the black-box function. The only exception is the Lévi experiment, where the variance response is a bimodal Gaussian changing the shape with the function.

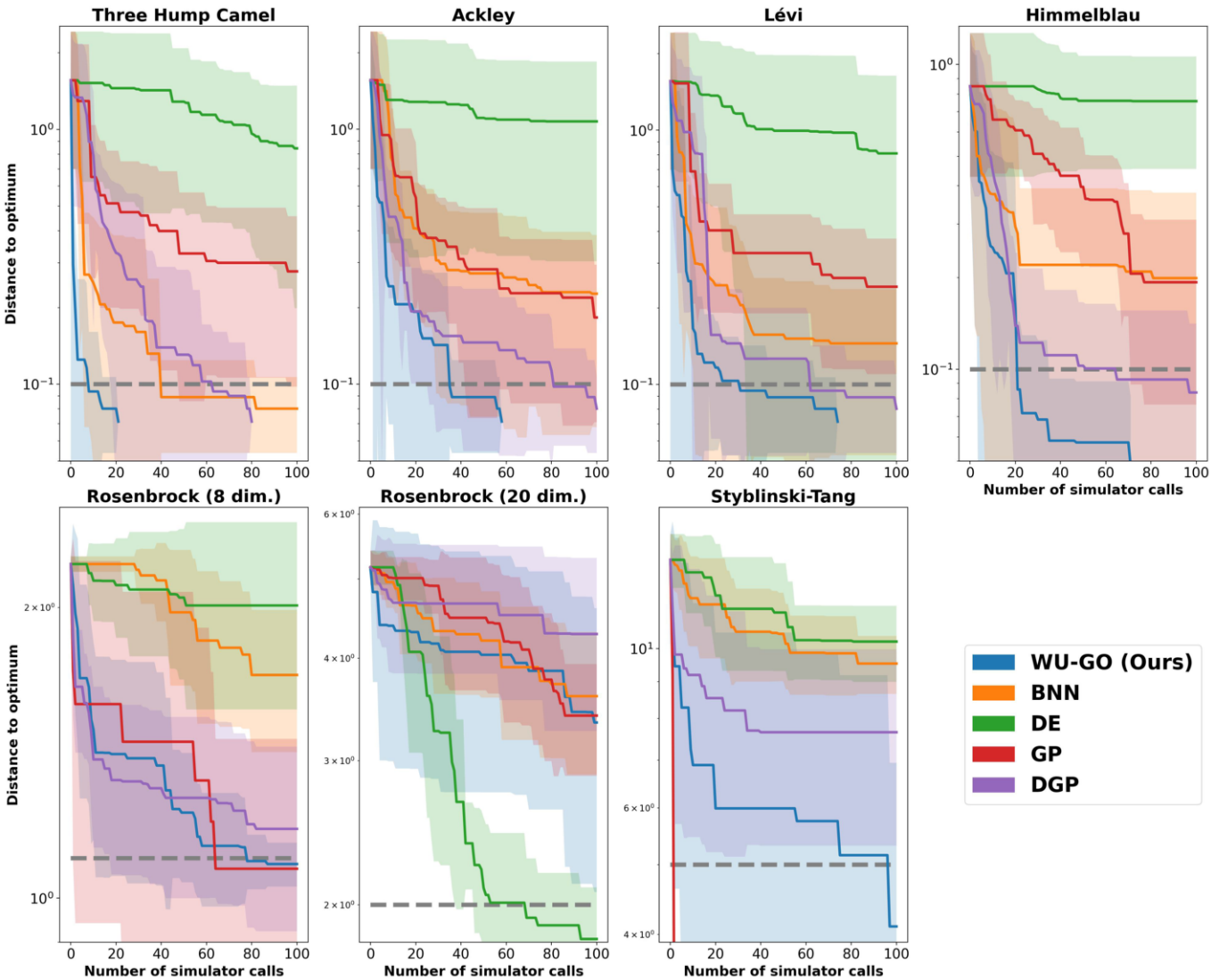

Figure 1. Experiment results - comparison of EGO and WU-GO. See Section 4.1 for more details.

Table 1 provides the parameters for each experiment. For each test function, we conduct 10 runs for N different starting configurations chosen via Latin hypercube sampling. Each response is represented by a sample of size n. Optimisation is performed in the search space Θ among a set of candidate points $\tilde{\Theta}$ consisting of 101^2 points.

4.1 Results

We evaluate all models in terms of distance to the optimum and speed of convergence. As each iteration of an optimisation algorithm would require a call of the simulator, we measure the speed of convergence as the number of simulator calls. In Fig. 1 we report the mean and the standard deviation of the distance to minimum w.r.t. the number of simulator calls.

Table 2 and Table 3 provide insights on results from Fig. 1 in a cross-sectional manner. Table 2 showcases the probability of a model to yield an ε-solution within the first 100 iterations. That is done by considering each model result as a Bernoulli random variable with the probability of success being the proportion of runs when the model met the stopping criteria. Table 3 is a numerical representation of Fig. 1 - it shows the distance to optima after 50 simulator calls. Top-3 results for each experiment are highlighted in green, yellow, and red respectively.

In the main body of this paper we compare WU-GO with EGO with baseline surrogates, as described in Section 3.7. The discrepancy between WU-GO and EGO results is less evident and should be discussed. The baselines with the LCB approach performed significantly worse. We place LCB results in Appendix E [53].

A faster decrease of mean values in Fig. 1 implies faster convergence of the algorithm. Smaller values of standard deviation mean higher confidence in predictions, i.e., consistency of the mean value. A combination of the two implies the efficiency of the model - not only does the model converge but it is also confident in its predictions.

In experiments with two-dimensional search spaces, WU-GO attains minima more efficiently than all selected baselines. That is not the case for, e.g., DE in the Lévi experiment or BNN in the Himmelblau experiment, in which standard deviation spreads over a large portion of the presented domain. Among the baselines, the performance of DGP is closest to that of WU-GO.

The Himmelblau test shows that despite responses being underrepresented - only 10 observations were simulated for each sample - WU-GO manages to outperform all baselines. It is likely that it would

Table 2. Probability to yield an ε-solution within the first 100 iterations.

Black Box Function	WU-GO	BNN	DE	GP	DGP
Three Hump Camel	1.0 ± 0.000	0.9 ± 0.095	0.1 ± 0.095	0.2 ± 0.126	1.0 ± 0.000
Ackley	1.0 ± 0.000	0.2 ± 0.126	0.1 ± 0.095	0.4 ± 0.155	0.9 ± 0.095
Lévi	1.0 ± 0.000	0.5 ± 0.158	0.0 ± 0.000	0.2 ± 0.126	0.9 ± 0.095
Himmelblau	1.0 ± 0.000	0.4 ± 0.155	0.0 ± 0.000	0.2 ± 0.126	0.7 ± 0.145
Rosenbrock (8 dim.)	0.6 ± 0.155	0.0 ± 0.000	0.1 ± 0.095	0.9 ± 0.095	0.5 ± 0.158
Rosenbrock (20 dim.)	0.3 ± 0.145	0.0 ± 0.000	0.8 ± 0.126	0.0 ± 0.000	0.0 ± 0.000
Styblinki-Tang	0.8 ± 0.126	0.0 ± 0.000	0.0 ± 0.000	1.00 ± 0.000	0.3 ± 0.145

Table 3. Distance to optimum after 50 simulator calls.

Black Box Function	WU-GO	BNN	DE	GP	DGP
Three Hump Camel	0.071 ± 0.000	0.089 ± 0.035	1.286 ± 0.796	0.326 ± 0.240	0.131 ± 0.151
Ackley	0.089 ± 0.035	0.272 ± 0.198	1.109 ± 0.779	0.283 ± 0.209	0.146 ± 0.091
Lévi	0.089 ± 0.035	0.157 ± 0.105	0.996 ± 0.978	0.329 ± 0.138	0.126 ± 0.074
Himmelblau	0.058 ± 0.034	0.220 ± 0.172	0.760 ± 0.307	0.397 ± 0.180	0.103 ± 0.063
Rosenbrock (8 dim.)	1.226 ± 0.188	1.977 ± 0.317	2.048 ± 0.363	1.452 ± 0.624	1.270 ± 0.270
Rosenbrock (20 dim.)	3.934 ± 1.460	4.205 ± 0.623	2.210 ± 0.398	4.483 ± 0.827	4.667 ± 0.837
Styblinki-Tang	5.993 ± 4.452	10.490 ± 1.479	11.233 ± 1.924	2.067 ± 0.028	7.651 ± 2.334

not be able to reproduce the original densities precisely. Still, the black-box function approximation is good enough for the optimisation task.

This is all evidence that WU-GO carries out optimisation efficiently of complex functions, e.g., Ackley and Lévi, as well as it does on simpler tests such as the Three Hump Camel. This signals that WU-GO should be more robust in general than EGO. It is worth noting that the results across all four two-dimensional experiments are similar. This shows the consistency of all models.

As the dimensionality of the search space increases, convergence is expected to decrease. We see this effect with WU-GO. However, the only time DE converges and shows a stable performance is in the 20-dimensional Rosenbrock experiment. GP behaves similarly. Although GP converges in the 8-dimensional Rosenbrock and the 20-dimensional Styblicky-Tang experiment, they fail with the 20-dimensional Rosenbrock. EGO may marginally outperform WU-GO, but the performance varies significantly depending on the surrogate model. This should make one question the stability and the consistency of EGO.

Although WU-GO's performance worsens, it converges in over half of the runs and is the only other model to show such results except for the previously mentioned DE and GP on specific tests. WU-GO's performance can be improved by tuning κ. It does require additional computations, but the resulting performance may be worth it as in the two-dimensional experiments (see ablation study in Appendix C [53] for more details).

4.2 *Physics experiment*

The muon shield is an essential element of the Search for Hidden Particles (SHiP) experiment, deflecting the abundant muon flux generated within the target away from the detector, otherwise a substantial background for particle searches. Muon deflection out of the spectrometer's acceptance by a magnetic field is straightforward; however, the challenge lies in the wide distribution of muons in phase space. To address this, SHiP uses both magnetic and passive shielding, as detailed in Ref. [4], to safeguard the emulsion target from muons.

To refine the muon shielding and optimize its configuration, a GEANT4 [3] simulation was developed to track muons' trajectories through the magnets. Each magnet's specifications encompass seven parameters: length, widths at both ends, heights, and air-gap widths between the field and return field. Overall we consider 22 parameters describing the architecture and the geometry of the detector. Thus we aim to improve the muon shield cost, $\mathbb{E}[f]$, over the space of experiment responses, $\mu \in \mathcal{P}_2$, by altering the configuration of the muon shield, $\theta \in \Theta$.

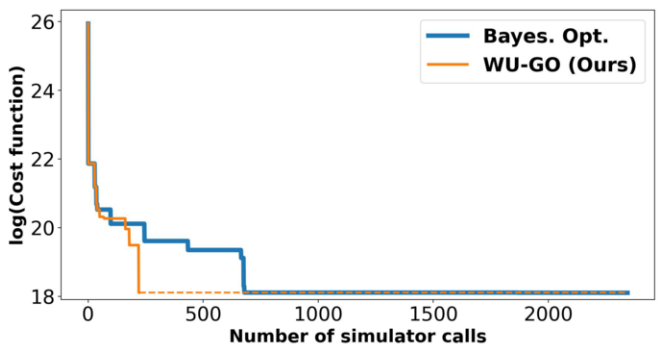

Figure 2. SHiP Muon Shield optimisation results.

The objective function of the muon shield optimisation is a combination of shielding efficiency and a mass cost function. We compare the result of WU-GO with a previously obtained result by Bayesian optimisation with a Gaussian process surrogate. The design suggested via Bayesian optimisation was approved for the construction of the experiment.

Within the first 300 iterations, WU-GO suggested a configuration of the SHiP experiment with almost identical performance as that suggested by Bayesian optimisation. In contrast with Bayesian optimisation, WU-GO also suggested some lightweight and compact designs. Although the shielding efficiency is suboptimal, the smaller dimensions of the muon shield may allow us to combine the muon shield with the rest of the experiment in a more efficient manner, i.e., minimise the background noise. This will be the subject of a future work of ours.

The starting configuration is described in Ref. [4] and all above-mentioned muon shield configurations can be found in Appendix F [53].

It is worth noting that the Bayesian optimisation stops after 2500

iterations because it runs out of memory. WU-GO is constant in memory. Hence we are not as limited in the number of optimisation loop iterations and we may gather as much statistics as we would like. In fields such as high energy physics, this feature is significant.

Bayesian optimisation with a GP surrogate model may be the most common approach to a design optimisation problem. With a GP surrogate, the following two issues may also arise.

The first challenge occurs when ground-truth observations cover the search space with large gaps in between. As previously mentioned, this is a general occurrence in black-box optimisation. Especially, when the search space is of higher dimensionality, e.g., the 22-dimensional search space of SHiP experiment parameters. In this case, GP tends to underestimate the uncertainty within the gaps of the search space.

The second issue arises when the variance of ground-truth observations depends on the derivative of the black-box function. Then the GP variance estimates tend to the variance of all ground truth samples, i.e., some average value. Thus the variance is generally mispredicted.

Both are illustrated on toy examples in Appendix G [53]. The GP is fitted well in the sense that the black-box function approximation is close. However, the aforementioned issues of mispredicted variance may yield mispredictions of both EGO and LCB optimisation algorithms.

5 Conclusion

We propose a novel approach for gradient-free optimisation of stochastic non-differentiable simulators. The algorithm is based on the concept of Wasserstein balls as ambiguity sets for uncertainty quantification. It uses a deep generative surrogate model, which allows us to model and optimise entities from simple Gaussian random variables to more complicated cases. We perform experiments on a real high energy physics detector simulator and a series of toy problems covering a wide range of possible real experiment settings. Our method, WU-GO, is compared with efficient global optimisation and lower confidence bound with Bayesian neural network, adversarial deep ensemble, Gaussian process, and deep Gaussian process as surrogate models. WU-GO attains minima more efficiently in comparison with all selected baselines, with DGP results being the closest - typically worse by a margin of one standard deviation. Experiments also suggest that our approach is robust to the shape of both the response and the objective function. WU-GO is easily extendible to higher dimensions of the parameters space but might require tuning of the exploration v. exploitation hyperparameter κ. Our method may not just mitigate issues of commonly used approaches but resolve them - a deep generative surrogate model can handle high-dimensional inputs, yield highly accurate reconstruction for various required configurations, and generate large samples with low computing time. WU-GO may be a universal approach for optimisation problems, especially for non-differentiable objectives and when reconstruction plays a key role in the optimisation pipeline.

Acknowledgements

The research leading to these results has received funding from the Basic Research Programme at the National Research University Higher School of Economics. This research was supported in part through computational resources of HPC facilities at HSE University.

References

[1] M. J. Abraham and J. E. Gready. Optimization of parameters for molecular dynamics simulation using smooth particle-mesh Ewald in GROMACS 4.5. *Journal of computational chemistry*, 32(9):2031–2040, 2011.

[2] M. Aehle et al. Exploration of differentiability in a proton computed tomography simulation framework. *Physics in Medicine & Biology*, 68 (24):244002, dec 2023. doi: 10.1088/1361-6560/ad0bdd. URL https://dx.doi.org/10.1088/1361-6560/ad0bdd.

[3] S. Agostinelli et al. Geant4—a simulation toolkit. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 506(3):250–303, 2003. ISSN 0168-9002. doi: https://doi.org/10.1016/S0168-9002(03)01368-8. URL https://www.sciencedirect.com/science/article/pii/S0168900203013688.

[4] A. Akmete et al. The active muon shield in the SHiP experiment. *Journal of Instrumentation*, 12(05):P05011, may 2017. doi: 10.1088/1748-0221/12/05/P05011. URL https://dx.doi.org/10.1088/1748-0221/12/05/P05011.

[5] S. Alarie, C. Audet, A. E. Gheribi, M. Kokkolaras, and S. Le Digabel. Two decades of blackbox optimization applications. *EURO Journal on Computational Optimization*, 9:100011, 2021.

[6] G. Amadio, J. Apostolakis, M. Bandieramonte, S. Behera, R. Brun, P. Canal, F. Carminati, G. Cosmo, L. Duhem, D. Elvira, et al. Stochastic optimization of GeantV code by use of genetic algorithms. In *Journal of Physics: Conference Series*, volume 898(4), page 042026. IOP Publishing, 2017.

[7] C. Audet, J. Bigeon, R. Couderc, and M. Kokkolaras. Sequential stochastic blackbox optimization with zeroth-order gradient estimators, 2023.

[8] W. Banzhaf, P. Nordin, R. E. Keller, and F. D. Francone. *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers Inc., 1998.

[9] J. Blanchet and K. R. A. Murthy. Quantifying Distributional Model Risk via Optimal Transport, 2017.

[10] D. Brookes, H. Park, and J. Listgarten. Conditioning by adaptive sampling for robust design. In *International conference on machine learning*, pages 773–782. PMLR, 2019.

[11] A. D. Bull. Convergence Rates of Efficient Global Optimization Algorithms. *Journal of Machine Learning Research*, 12(88):2879–2904, 2011. URL http://jmlr.org/papers/v12/bull11a.html.

[12] M. B. Chang, T. D. Ullman, A. Torralba, and J. B. Tenenbaum. A Compositional Object-Based Approach to Learning Physical Dynamics. *CoRR*, abs/1612.00341, 2016. URL http://arxiv.org/abs/1612.00341.

[13] R. Chen, M. Menickelly, and K. Scheinberg. Stochastic optimization using a trust-region method and random models. *Mathematical Programming*, 169:447–487, 2018.

[14] A. R. Conn, N. I. Gould, and P. L. Toint. *Trust region methods*. SIAM, 2000.

[15] A. Cozad, N. V. Sahinidis, and D. C. Miller. Learning surrogate models for simulation-based optimization. *AIChE Journal*, 60(6): 2211–2227, 2014. doi: https://doi.org/10.1002/aic.14418. URL https://aiche.onlinelibrary.wiley.com/doi/abs/10.1002/aic.14418.

[16] A. Damianou and N. D. Lawrence. Deep Gaussian Processes. In C. M. Carvalho and P. Ravikumar, editors, *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics*, volume 31 of *Proceedings of Machine Learning Research*, pages 207–215, Scottsdale, Arizona, USA, 29 Apr–01 May 2013. PMLR. URL https://proceedings.mlr.press/v31/damianou13a.html.

[17] F. de Avila Belbute-Peres, K. Smith, K. Allen, J. Tenenbaum, and J. Z. Kolter. End-to-End Differentiable Physics for Learning and Control. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/842424a1d0595b76ec4fa03c46e8d755-Paper.pdf.

[18] N. de Freitas and Z. Wang. Bayesian Optimization in High Dimensions via Random Embeddings. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013.

[19] J. Degrave, M. Hermans, J. Dambre, and F. Wyffels. A Differentiable Physics Engine for Deep Learning in Robotics. *CoRR*, abs/1611.01652, 2016. URL http://arxiv.org/abs/1611.01652.

[20] J. Djolonga, A. Krause, and V. Cevher. High-dimensional gaussian process bandits. *Advances in neural information processing systems*, 26, 2013.

[21] T. Dorigo et al. Toward the end-to-end optimization of particle physics instruments with differentiable programming. *Reviews in Physics*, 10: 100085, 2023. ISSN 2405-4283. doi: https://doi.org/10.1016/j.revip.2023.100085. URL https://www.sciencedirect.com/science/article/pii/S2405428323000047.

[22] D. Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014.

[23] P. M. Esfahani and D. Kuhn. Data-driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations, 2017.

[24] A. D. Flaxman, A. T. Kalai, and H. B. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient, 2004.

[25] P. I. Frazier. A Tutorial on Bayesian Optimization, 2018.

[26] S.-W. Fu, C.-F. Liao, Y. Tsao, and S.-D. Lin. Metricgan: Generative adversarial networks based black-box metric scores optimization for speech enhancement. In *International Conference on Machine Learning*, pages 2031–2041. PMLR, 2019.

[27] R. Gao and A. J. Kleywegt. Distributionally Robust Stochastic Optimization with Wasserstein Distance, 2022.

[28] R. Garnett, M. A. Osborne, and P. Hennig. Active learning of linear embeddings for Gaussian processes. *arXiv preprint arXiv:1310.6740*, 2013.

[29] R. Gnanasambandam, B. Shen, A. C. C. Law, C. Dou, and Z. Kong. Deep Gaussian Process for Enhanced Bayesian Optimization and its Application in Additive Manufacturing. 6 2023. doi: 10.36227/techrxiv.23548143.v1. URL https://www.techrxiv.org/articles/preprint/Deep_Gaussian_Process_for_Enhanced_Bayesian_Optimization_and_its_Application_in_Additive_Manufacturing/23548143.

[30] R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams, and A. Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[31] N. S. Gorbach, A. A. Bian, B. Fischer, S. Bauer, and J. M. Buhmann. Model selection for gaussian process regression. In *Pattern Recognition: 39th German Conference, GCPR 2017, Basel, Switzerland, September 12–15, 2017, Proceedings 39*, pages 306–318. Springer, 2017.

[32] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. Improved training of Wasserstein GANs. *Advances in neural information processing systems*, 30, 2017.

[33] A. Gupta and J. Zou. Feedback GAN for DNA optimizes protein functions. *Nature Machine Intelligence*, 1(2):105–111, 2019.

[34] M. U. Gutmann, J. Cor, and er. Bayesian Optimization for Likelihood-Free Inference of Simulator-Based Statistical Models. *Journal of Machine Learning Research*, 17(125):1–47, 2016. URL http://jmlr.org/papers/v17/15-017.html.

[35] A. Hebbal, L. Brevault, M. Balesdent, T. El-Ghazali, and N. Melab. Bayesian optimization using deep Gaussian processes with applications to aerospace system design - Optimization and Engineering — link.springer.com. https://link.springer.com/article/10.1007/s11081-020-09517-8#citeas, 2020. [Accessed 01-12-2023].

[36] D. R. Jones, M. Schonlau, and W. J. Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13: 455–492, 1998.

[37] I. Kononenko. Bayesian neural networks. *Biological Cybernetics*, 61(5): 361–370, 1989.

[38] S. Krishnamoorthy, S. M. Mashkaria, and A. Grover. Diffusion Models for Black-Box Optimization, 2023.

[39] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics*, pages 130–166. Informs, 2019.

[40] T. Lai and H. Robbins. Asymptotically efficient adaptive allocation rules. *Advances in Applied Mathematics*, 6(1):4–22, 1985. ISSN 0196-8858. doi: https://doi.org/10.1016/0196-8858(85)90002-8. URL https://www.sciencedirect.com/science/article/pii/0196885885900028.

[41] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[42] N. Lanzetti, S. Bolognani, and F. Dörfler. First-order Conditions for Optimization in the Wasserstein Space, 2022.

[43] H. Liu, X. GU, and D. Samaras. A Two-Step Computation of the Exact GAN Wasserstein Distance. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3159–3168. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/liu18d.html.

[44] F. Luo and S. Mehrotra. Distributionally robust optimization with decision dependent ambiguity sets. *Optimization Letters*, 14(8):2565–2594, Nov. 2020. ISSN 1862-4480. doi: 10.1007/s11590-020-01574-3.

[45] W. Macready and D. Wolpert. Bandit problems and the exploration/exploitation tradeoff. *IEEE Transactions on Evolutionary Computation*, 2(1):2–22, 1998. doi: 10.1109/4235.728210.

[46] J. Mockus. *Bayesian approach to global optimization: Theory and applications*. Kluwer Academic Publishers, 1989.

[47] M. Mustafa, D. Bard, W. Bhimji, Z. Lukić, R. Al-Rfou, and J. M. Kratochvil. CosmoGAN: creating high-fidelity weak lensing convergence maps using Generative Adversarial Networks. *Computational Astrophysics and Cosmology*, 6(1):1, May 2019. ISSN 2197-7909. doi: 10.1186/s40668-019-0029-9.

[48] J. Nie, L. Yang, S. Zhong, and G. Zhou. Distributionally Robust Optimization with Moment Ambiguity Sets. *Journal of Scientific Computing*, 94(1):12, Dec. 2022. ISSN 1573-7691. doi: 10.1007/s10915-022-02063-8.

[49] J. Nocedal and S. J. Wright. *Numerical optimization*. Springer, 1999.

[50] M. Paganini, L. de Oliveira, and B. Nachman. Accelerating Science with Generative Adversarial Networks: An Application to 3D Particle Showers in Multilayer Calorimeters. *Phys. Rev. Lett.*, 120:042003, Jan 2018. doi: 10.1103/PhysRevLett.120.042003. URL https://link.aps.org/doi/10.1103/PhysRevLett.120.042003.

[51] P. Y. Papalambros and D. J. Wilde. *Principles of optimal design: modeling and computation*. Cambridge university press, 2000.

[52] S. M. Pesenti and S. Jaimungal. Portfolio Optimization within a Wasserstein Ball. *SIAM Journal on Financial Mathematics*, 14(4):1175–1214, 2023. doi: 10.1137/22M1496803. URL https://doi.org/10.1137/22M1496803.

[53] T. Ramazyan, M. Hushchyn, and D. Derkach. Global Optimisation of Black-Box Functions with Generative Models in the Wasserstein Space. 2024. URL https://arxiv.org/abs/2407.11917.

[54] B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas. Taking the Human Out of the Loop: A Review of Bayesian Optimization. *Proceedings of the IEEE*, 104(1):148–175, 2016. doi: 10.1109/JPROC.2015.2494218.

[55] J. R. Shewchuk. An Introduction to the Conjugate Gradient Method Without the Agonizing Pain. Technical report, USA, 1994.

[56] S. Shirobokov, V. Belavin, M. Kagan, A. Ustyuzhanin, and A. G. Baydin. Black-Box Optimization with Local Generative Surrogates. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14650–14662. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/a878dbebc902328b41dbf02aa87abb58-Paper.pdf.

[57] T. Sjöstrand, S. Mrenna, and P. Skands. A brief introduction to PYTHIA 8.1. *Computer Physics Communications*, 178(11):852–867, 2008.

[58] J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. In F. Pereira, C. Burges, L. Bottou, and K. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL https://proceedings.neurips.cc/paper_files/paper/2012/file/05311655a15b75fab86956663e1819cd-Paper.pdf.

[59] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, and R. Adams. Scalable Bayesian optimization using deep neural networks. In *International conference on machine learning*, pages 2171–2180. PMLR, 2015.

[60] J. T. Springenberg, A. Klein, S. Falkner, and F. Hutter. Bayesian Optimization with Robust Bayesian Neural Networks. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/a96d3afec184766bfeca7a9f989fc7e7-Paper.pdf.

[61] J. Stanczuk, C. Etmann, L. M. Kreusser, and C.-B. Schönlieb. Wasserstein GANs work because they fail (to approximate the Wasserstein distance). *arXiv preprint arXiv:2103.01678*, 2021.

[62] K. Svanberg. The method of moving asymptotes—a new method for structural optimization. *International Journal for Numerical Methods in Engineering*, 24(2):359–373, 1987. doi: https://doi.org/10.1002/nme.1620240207. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/nme.1620240207.

[63] G. Szekely. E-Statistics: The energy of statistical samples. *Skandinavisk Aktuarietidskrift*, 01 2003.

[64] J. Yin and N. Li. Ensemble learning models with a Bayesian optimization algorithm for mineral prospectivity mapping. *Ore Geology Reviews*, 145: 104916, 2022. ISSN 0169-1368. doi: https://doi.org/10.1016/j.oregeorev.2022.104916. URL https://www.sciencedirect.com/science/article/pii/S0169136822002244.

[65] M.-C. Yue, D. Kuhn, and W. Wiesemann. On Linear Optimization over Wasserstein Balls, 2021.

[66] M. Zhang, H. Li, and S. Su. High dimensional Bayesian optimization via supervised dimension reduction. *arXiv preprint arXiv:1907.08953*, 2019.