DIRECT Optimisation with Bayesian Insights: Assessing Reliability Under Fixed Computational Budgets

Fu Wang FW377@EXETER.AC.UK
Zeyu Fu Z.FU@EXETER.AC.UK
University of Exeter, Exeter, EX4 4QF, UK

Xiaowei Huang XIAOWEI.HUANG@LIVERPOOL.AC.UK
University of Liverpool, Liverpool, L69 3BX, UK

Wenjie Ruan* W.RUAN@TRUSTAI.UK
University of Exeter, Exeter, EX4 4QF, UK

Abstract

We introduce a method for probabilistically evaluating the reliability of Lipschitzian global optimisation under a constrained computational budget, a context frequently encountered in various applications. By interpreting the slope data gathered during the optimisation process as samples from the objective function's derivative, we utilise Bayesian posterior prediction to derive a confidence score for the optimisation outcomes. We validate our approach using numerical experiments on four multi-dimensional test functions, and the results highlight the practicality and efficacy of our approach.

1. Introduction

The DIRECT [4] algorithm is a gradient-free global optimisation method that does not rely on the Lipschitz constant. This method is renowned for its efficiency in zeroing in on the vicinity of a global optimum and can yield a robust approximation of the global optimum given ample function evaluations [1]. In practice, however, function evaluation could be a computationally expensive operation or have limited access during optimisation. Given these constraints, our focus shifts to determining the quality of the approximation under a finite number of function evaluations, which aligns more closely with real-world scenarios.

Building on the foundation of DIRECT optimisation, our study posits that the task of predicting whether the actual value exceeds a predetermined threshold can be re-envisioned as estimating if the local Lipschitz constant within each subspace goes beyond a certain value. In light of this, we propose an approach in which the recorded slope data is considered as random samples from the first-order derivative. Subsequently, we perform Bayesian posterior prediction for each discrete subspace, and collating this probabilistic information allows us to predict the confidence of the optimisation outcome within a defined computational constraint. To validate our approach, we conducted numerical experiments on four widely used multi-dimensional test functions and compared the estimated distribution based on DIRECT sampling to the results obtained via grid and random sampling. Besides, we also compared the confidence obtained via maximum likelihood estimation

*Corresponding Author

and Bayesian posterior prediction on different setups, where the latter gave more precise predictions in our experiments.

2. A Brief Overview of the DIRECT Optimisation

Suppose that there is a Lipschitz-continuous objective function f. We want to find the optimal solution, $x^* = \underset{x \in \mathcal{X}}{\arg\min} f(x)$, to achieve the global minimum of f within a bound parameter space $\mathcal{X} \in \mathbb{R}^n$. In DIRECT optimisation [4], $\mathcal{X}$ is firstly projected into a unit n-dimensional search space $\mathcal{C}$, which is viewed as the initial Potential Optimal (PO) subspace. DIRECT evaluates the centre point of the PO subspace and the points sampled at a distance of one-third the side length of search space from the centre point in each coordinate and then divides the PO subspace into smaller subspaces accordingly, where each newly sampled point becomes the centre of a hyperrectangle. New PO subspaces are identified for the following divisions, completing one optimisation round. The loop of space divisions and PO space selection is carried out until termination or convergence.

Wang et al. [7] adopted DIRECT-l [1] to verify the robustness of deep neural network models against adversarial geometric transformations. Although DIRECT does not require the Lipschitz constant to proceed, this information is extremely critical to enable the obtained lower or upper bounds with provable guarantees. Therefore, they proposed computing and recording the absolute values of the slopes between queried points and utilising the largest value to estimate the lower bound of the objective function.

3. Quantifying the Uncertainty Before Convergence

In robustness and reachability analysis [2, 5, 6, 9], a common scenario is to determine whether the value of a target function can be smaller than a given threshold z in a specific domain. Typically, the target function can measure a system's performance, and its values smaller than z could indicate system faults or unsafe behaviours [3, 7, 8]. Performing DIRECT could allow us to see if $\min_{c \in \mathcal{C}} f(c) \geq z$ holds after convergence, but it may take a large number of function evaluations, requiring unrealistic computational costs. Therefore, parallel to comparing the minimum found and the threshold z, we propose to estimate the possibility of $\min_{c \in \mathcal{C}} f(c) \geq z$ at each iteration.

Notation Given a function f, recall that we are interested in whether the true minimum $\min_{c \in \mathcal{C}} f(c)$ is smaller than a threshold z in each iteration of DIRECT . We denote c_i as the centre of hyperrectangle indexed by $i \in \mathcal{I}$ and write the query result at c_i as $f(c_i)$. For a hyperrectangle i, we denote $d_{i,j}$ and $K_{i,j}$ as half of its side length and the local Lipschitz constant at dimension j, respectively. Although the local Lipschitz constants are unknown, we follow the approach proposed by Wang et al. [7] and record $k_{i,j}$, the absolute values of slopes between the centre of the hyperrectangle i and newly sampled points along dimension j, during optimisation.

3.1. Gain Confidence from Queries

Suppose the DIRECT optimisation has been carried out for some iterations, and all query results are greater than the threshold of interest. In each subspace, the probability of the ground truth minimum

$f(\hat{c}_i)$ smaller than z can be expressed as

$$\mathrm{P}(f(\hat{c}_i) < z) \leq \mathrm{P}\Big(\bigcup_{j \in \mathbb{N}^+_{\leq n}} f(c_i) - K_{i,j} d_{i,j} < z \Big) = 1 - \mathrm{P}\Big(\bigcap_{j \in \mathbb{N}^+_{\leq n}} K_{i,j} \leq \frac{f(c_i) - z}{d_{i,j}} \Big). \tag{1}$$

The above inequality holds because of $f(\hat{c}_i) \geq \min(f(c_i) - K_{i,j} d_{i,j})$. Then, globally, we can write the confidence of the ground truth minimum greater than z as

$$\mathrm{P}(\min_{c \in \mathcal{C}} f(c) \geq z) = 1 - \mathrm{P}(\exists i \in \mathcal{I} : f(\hat{c}_i) < z) \geq \mathrm{P}\Big(\bigcap_{i \in \mathcal{I}} \bigcap_{j \in \mathbb{N}^+_{\leq n}} K_{i,j} \leq \frac{f(c_i) - z}{d_{i,j}} \Big). \tag{2}$$

We can see the right-hand side of Eq. (2) $\rightarrow 0$ as the number of subspaces increases. Therefore, we define the confidence of $\min_{c \in \mathcal{C}} f(c) \geq z$ as follows to avoid the numerical issue,

$$1 - \mathop{\mathbb{E}}_{i \in \mathcal{I}, j \in \mathbb{N}^+_{\leq n}} \Big[\mathrm{P}\big(K_{i,j} > \frac{f(c_i) - z}{d_{i,j}} \big) \Big]. \tag{3}$$

3.2. Slopes are Randomly Sampled Derivative that Subject to Exponential Distribution

After any iteration of the DIRECT optimisation, the algorithm makes observations of the objective function, allowing us to obtain values of $f(c_i)$, d_i, c_i, and $k_{i,j}$ so that every $\frac{f(c_i)-z}{d_{i,j}}$ in Eq. (3) is a known constant. Let $\hat{z}_{i,j} = f(c_i) - z/d_{i,j}$, the problem can then be translated into measuring the probability of whether the local Lipschitz constant $K_{i,j}$ is greater than $\hat{z}_{i,j}$. As stated in Remark 1, every observed slope can be the partial derivative of a point within $\mathcal{C}$. But it would be difficult to obtain the location of these points. Such uncertainty, therefore, motivates us to view these slopes as randomly sampled first-order derivatives subject to certain distributions. As the upper bound of the derivative in a subspace, the local Lipschitz constant should also be subject to the same distribution. To compute $\mathrm{P}(K_{i,j} > \hat{z}_{i,j})$, we would need to estimate the distribution of derivatives.

Remark 1 (Mean value theorem) *Let k^j be the slope between two points c_a and c_b along the dimension j, i.e., $k^j = |f(c^a) - f(c^b)|/||c^a - c^b||$. Assuming that the objective function is differentiable, according to the Mean Value Theorem, a point c' exists between c^a and c^b such that $\partial f(c')/\partial c'^j = k^j$.*

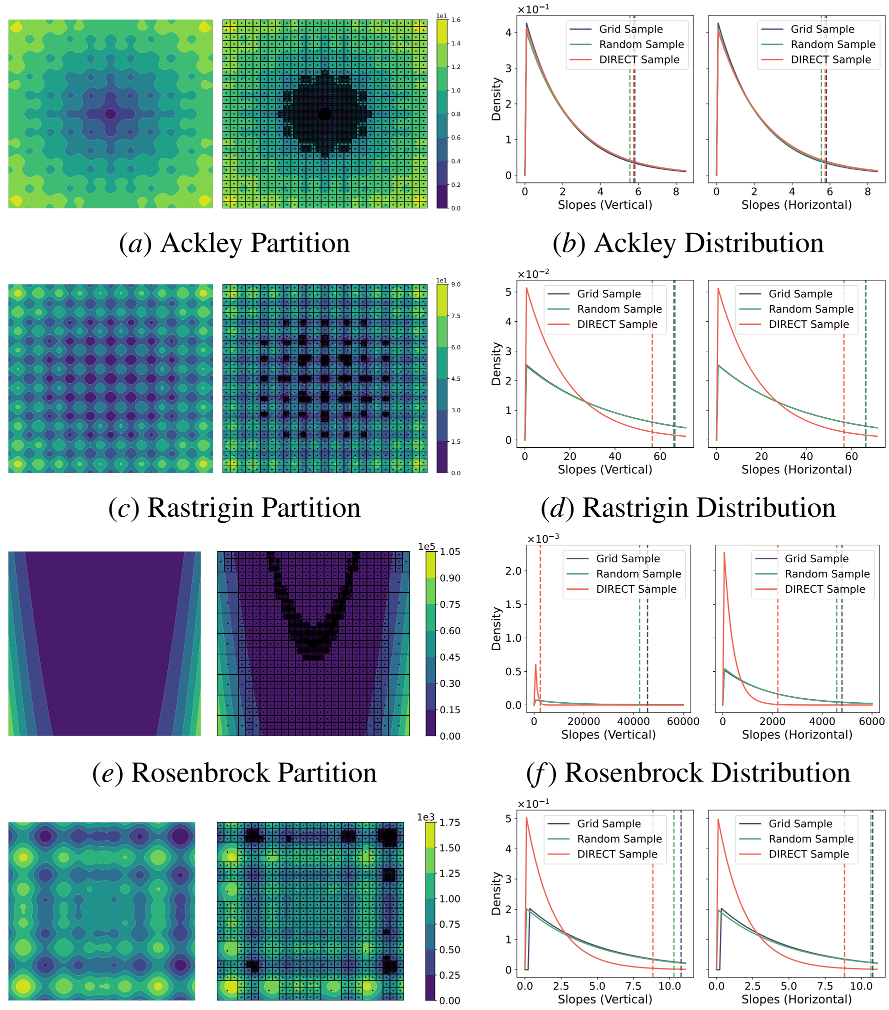
Straightforwardly, we consider the exponential distribution as a good fit because $k_{i,j}$ is non-negative. Assuming all dimensions are independent and $K_j \sim exp(\lambda_j)$, for $\forall i \in \mathcal{I}$, we have $\mathrm{P}(K_{i,j} > \hat{z}_{i,j}) = \exp(-\lambda_j \hat{z}_{i,j})$, and then the confidence defined in Eq. (3) is given by

$$\mathop{\mathbb{E}}_{i \in \mathcal{I}} \Big[\prod_{j \in \mathbb{N}^+_{\leq n}} \big(1 - \exp(-\lambda_j \hat{z}_{i,j})\big) \Big] \tag{4}$$

Through Maximum Likelihood Estimation (MLE), we can use $\hat{\lambda}_j = 1/\mathbb{E}_{i \in \mathcal{I}}[k_{i,j}]$ as an estimation of λ_j, which allows us to calculate the Eq. (4).

3.3. Applying Bayesian Posterior Prediction

In addition to using the MLE to fit the exponential distributions, we can take a step forward and treat the λs inside exponential distributions as latent variables and perform posterior prediction. Inside a



(*a*) Ackley Partition (*b*) Ackley Distribution

(*c*) Rastrigin Partition (*d*) Rastrigin Distribution

(*e*) Rosenbrock Partition (*f*) Rosenbrock Distribution

(*g*) Schwefel Partition (*h*) Schwefel Distribution

Figure 1: Visualisations of the partition results on four test functions with the estimated distribution of derivatives and their 95-th percentiles. The distributions are estimated via MLE based on different sampling strategies with the same amount of function evaluations.

subspace i, we are interested in the probability of the local Lipschitz constant greater than a certain value, which can be given as

$$\mathrm{P}(K_{i,j} > \hat{z} | \{k_{i,j}, i \in \mathcal{I}\}) = \int \mathrm{P}(K_{i,j} > \hat{z} | \lambda) \cdot \mathrm{P}(\lambda | \{k_{i,j}, i \in \mathcal{I}\}) \, d\lambda, \tag{5}$$

where $\{k_{i,j}, i \in \mathcal{I}\}$ represents a set that contains all the sampled slopes along j-th dimension. As we assume that the local Lipschitz constant is subject to the exponential distribution, it is natural to choose a gamma distribution $\mathrm{Gamma}(\lambda; \alpha, \beta)$ as the prior distribution. Accordingly, the posterior

Table 1: Computing the confidences corresponds to different thresholds. 'min' and 'LB' stand for the found minimum and approximated lower bound. 'GT' represents the ground truth minimum, which is 0 for all test functions.

Functions	# Dim.	# Query	min	LB	< GT		< min		< LB	
					MLE	PP	MLE	PP	MLE	PP
Schwefel	2	5e3	2.9e-5	-1.68	0.0164%	0.0155%	0.02%	0.02%	0	0
	3	1e4	4.3e-5	-2.03	0.007%	0.006%	0.01%	0.01%	0	0
	4	2e4	5.8e-5	-2.7	0.005%	0.003%	0.006%	0.006%	0	0
Rosenbrock	2	5e2	4e-6	-44.51	41.23%	37.46%	41.23%	37.46%	0.03%	0.01%
	3	1e3	1.4e-3	-179.01	24.98%	18.36%	25.41%	18.82%	0	0
	4	2e3	2.1e-3	-59.76	16.16%	9.87%	17.1%	11.02%	0	0
Rastrigin	2	5e2	1.9e-4	-0.083	2.97%	1.12%	3.2%	1.12%	0	0
	3	1e3	2.8e-4	-0.13	0.17%	0.08%	0.26%	0.16%	0	0
	4	2e3	3.7e-4	-0.16	0.25%	0.06%	0.35%	0.12%	0	0
Ackley	2	2e2	2.7e-3	-0.01	0	0	0.54%	0.54%	0	0
	3	5e2	2.8e-3	-0.01	0	0	0.23%	0.23%	0	0
	4	1e3	7.4e-3	-0.03	0	0	0.05%	0.05%	0	0

distribution is also a gamma distribution given by

$$\begin{aligned} \mathrm{P}(\lambda|\{k_{i,j}, i \in \mathcal{I}\}) &= \mathrm{Gamma}(\lambda; Q+\alpha, \beta + \sum_{i \in \mathcal{I}} k_{i,j}) \\ &= \frac{(\beta + \sum_{i \in \mathcal{I}} k_{i,j})^{Q+\alpha}}{\Gamma(Q+\alpha)} \lambda^{Q+\alpha-1} \exp\big(-(\beta + \sum_{i \in \mathcal{I}} k_{i,j})\lambda\big). \end{aligned} \tag{6}$$

By substituting $\mathrm{P}(K_{i,j} > \hat{z}|\lambda) = \exp(-\lambda\hat{z})$ and Eq. (6) into Eq. (5), we can get

$$\mathrm{P}(K_{i,j} > \hat{z}|\{k_{i,j}, i \in \mathcal{I}\}) = \left(\frac{\beta + \sum_{i \in \mathcal{I}} k_{i,j}}{\beta + \sum_{i \in \mathcal{I}} k_{i,j} + \hat{z}}\right)^{Q+\alpha}. \tag{7}$$

4. Empirical Study

We evaluate the confidence estimation on different multi-dimensional test functions, including the Rosenbrock function, Rastrigin function, Ackley function and Schwefel function. The domain range for the Rosenbrock, Rastrigin, and Ackley functions is [-5,5], while the Schwefel function is evaluated within [-500,500]. Following the implementation given by Wang et al. [7], we adopt the DIRECT-l to find the minimum and approximate the lower bound. In addition, we utilise the recorded slope values to estimate the distribution of the first-order derivative and compute the confidence with respect to different thresholds.

In Fig. 1, we provide visualisations of the 2-D case of each test function. The function landscapes and the partitions are given in the left column, and the distributions corresponding to each dimension obtained via MLE and DIRECT are plotted on the right. Here, DIRECT-l is carried out with up to 2500 function evaluations, and we also performed grid and random sampling on the first-order derivative to estimate the distributions for comparison. Compared to the grid and random sampling, using DIRECT-l results in a long-tailed distribution with higher density on the left. This distinction could potentially lead to an underestimation of the likelihood of a larger Lipschitz constant. In Tab. 1, we compare the confidence based on MLE given in Eq. (4) with the posterior prediction given in Eq. (7). To prevent the global minimum from being located in the initial query via DIRECT-l, we adjust the domain of the Ackley and Rastrigin functions from [-5,5] to [-5,4]. Based on the same records, we can see that posterior prediction could provide more accurate confidence than MLP.

5. Conclusion

In this paper, we propose to incorporate a probabilistic assessment of the reliability of DIRECT optimisation with a fixed computational budget. During the optimisation, we propose to view the recorded slopes as a sampling on the first-order derivative and perform Bayesian posterior prediction to compute a confidence score about whether the ground truth would exceed a threshold. Our empirical study demonstrates the feasibility and effectiveness of the proposed method, but it also reveals that DIRECT sampling could introduce bias when fitting the distribution of the derivative. In future work, we aim to refine the distribution estimation and provide a more precise confidence score prediction.

References

[1] Joerg M. Gablonsky and Carl T. Kelley. A locally-biased form of the DIRECT algorithm. *J. Glob. Optim.*, 21(1):27–37, 2001.

[2] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, et al. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270., 2020.

[3] Xiaowei Huang, Wenjie Ruan, Qiyi Tang, and Xingyu Zhao. Bridging formal methods and machine learning with global optimisation. In *ICFEM*, 2022.

[4] Donald R. Jones et al. Lipschitzian optimization without the lipschitz constant. *Journal of Optimization Theory and Applications*, 79:157–181, 1993.

[5] Wenjie Ruan, Xiaowei Huang, and Marta Kwiatkowska. Reachability analysis of deep neural networks with provable guarantees. In *IJCAI*, 2018.

[6] Wenjie Ruan, Min Wu, et al. Global robustness evaluation of deep neural networks with provable guarantees for the hamming distance. In *IJCAI*, 2019.

[7] Fu Wang, Peipei Xu, Wenjie Ruan, and Xiaowei Huang. Towards verifying the geometric robustness of large-scale neural networks. In *AAAI*, 2023.

[8] Peipei Xu, Fu Wang, Wenjie Ruan, Chi Zhang, and Xiaowei Huang. Sora: Scalable black-box reachability analyser on neural networks. In *ICASSP*, 2023.

[9] Chi Zhang, Wenjie Ruan, and Peipei Xu. Reachability analysis of neural network control systems. In *AAAI*, 2023.