Exploring Nonlinear Distance Metrics for Lipschitz Constant Estimation in Lower Bound Construction for Global Optimization

MOHAMMADSINA ALMASI, University of Illinois at Chicago, USA

HADIS ANAHIDEH, University of Illinois at Chicago, USA

JAY M. ROSENBERGER, University of Texas at Arlington, USA

Bounds play a crucial role in guiding optimization algorithms, improving their speed and quality, and providing optimality gaps. While Lipschitz constant-based lower bound construction is an effective technique, the quality of the linear bounds depends on the function's topological properties. In this research, we improve upon this by incorporating nonlinear distance metrics and surrogate approximations to generate higher-quality bounds. We emphasize the importance of using a flexible distance metric that can adapt to any function. We examine the characteristics and properties of different incorporated distance metrics. While the linear distance metric is popular in the literature due to its simplicity and intuitive interpretation, we discuss the potential benefits of alternative distance metrics, such as sublinear and superlinear distance metrics, which may be used for Hölder continuous functions. Sublinear distance metrics are advantageous for sparse data settings, while superlinear distance metrics can capture nonlinear relationships between data points. Combining surrogate models and nonlinear distance metrics for Lipschitz constant estimations results in high-quality lower bounds that can contribute to more effective exploration and exploitation and more accurate optimality gap estimation.

1 INTRODUCTION

Global optimization problems are ubiquitous in various fields of science and engineering, ranging from aerospace and mechanical engineering to finance and logistics [13, 40]. These problems often involve complex objective functions, nonlinear constraints, and high-dimensional search spaces, making them particularly challenging to solve [10, 36]. Consequently, developing efficient optimization algorithms that can find a global optimum is crucial for many practical applications.

One approach to solving global optimization problems is to leverage lower [1] bounds to guide the optimization process and narrow down the search space. Various techniques have been developed for constructing lower bounds on the

---

[1]While the focus of this paper is on minimization problems and the use of lower bounds, the same discussion can be extended to maximization problems and the use of upper bounds.

optimal objective value, such as convex relaxation [e.g., 50], interval analysis [e.g., 7], statistical inference [e.g., 56], and Lipschitz continuity [e.g., 33].

The research developed here is motivated by surrogate optimization, sometimes referred to as Bayesian optimization. The premise of surrogate optimization is the cost to evaluate a point $\mathbf{x}$ in the objective function f is expensive [e.g., 44]. Consequently, surrogate optimization methods usually employ the following general approach. First, a design of experiments technique determines an initial set of points to evaluate. Costly objective evaluations are performed, and a surrogate model $\hat{f}$ to predict f at unevaluated points is trained. Then based on the surrogate (exploitation) and the spatial representation of the evaluated points (exploration), new points to evaluate are determined. Except for the initial design of experiments, these steps iterate until a stopping criterion is met.

An estimated lower bound on f would be valuable within this surrogate optimization framework. It could be used to estimate an optimality gap, balance exploitation and exploration, and provide a stopping criterion. Moreover, while the lower bounding approach developed in this research is motivated by surrogate optimization, it could be used within other black box optimization or simulation optimization methods. The only assumption in this research is that f satisfies the Hölder condition [e.g., 63].

Using the Lipschitz constant for lower-bound construction is a powerful technique for improving the efficiency and effectiveness in a variety of contexts, such as branch and bound algorithms [24, 66], Lipschitzian optimization [e.g., 49], and black-box optimization [e.g., 38]. It allows for more informed and focused search strategies that can lead to faster convergence and better solutions. The Lipschitz constant is a mathematical concept that characterizes the local smoothness of a function [43, 62]. Specifically, it measures the maximum rate of change of a function over any two points in its domain.

The Lipschitz bound for a function f at a point $\mathbf{x}^i$ is given by:

$$f(\mathbf{x}^i) \geq f(\mathbf{x}^j) - L\|\mathbf{x}^i - \mathbf{x}^j\| \tag{1}$$

In addition to the constant L, the other two terms are an evaluated point $\mathbf{x}^j$ and a distance metric $\|\mathbf{x}^i - \mathbf{x}^j\|$, which is an L-2 norm. This bound is well-known for any Lipschitz continuous function [e.g., 51], and is used in data-driven spatial Branch and Bound [e.g., 42].

Adding to this, in statistical terms, particularly in the context of confidence and prediction bounds, a fundamental aspect is the assessment of how far an estimated function, $\hat{f}(\mathbf{x})$, deviates from the expected true function, $E[f(\mathbf{x})]$. This can be expressed in the form of a probability statement:

$$\Pr\left(\hat{f}(\mathbf{x}) - \varphi(\mathbf{x}) \leq E[f(\mathbf{x})]\right) = 1 - \alpha \tag{2}$$

Here, $\varphi(\mathbf{x})$ is an estimation of the prediction error of $\hat{f}$ [9, 16]. Once the significance level α is determined, the structure of the bound can be established alternatively as follows:

$$E[f(\mathbf{x})] \geq \hat{f}(\mathbf{x}) - L\varphi(\mathbf{x}, \bar{\mathbf{x}}) \tag{3}$$

In this case, $\bar{\mathbf{x}}$ is usually the center of the data, $\varphi(\mathbf{x}, \bar{\mathbf{x}})$ is a function of the distance between $\mathbf{x}$ and $\bar{\mathbf{x}}$ and the uncertainty of the surrogate, and the constant L is usually based upon a statistic. For example, in Working-Hotelling [e.g., 64], $\bar{\mathbf{x}}$ is the mean of the evaluated points instead of the nearest evaluated point, as in the Lipschitz framework. Similar bounds are used in Probabilistic Branch and Bound [e.g., 24].

Estimating the Lipschitz constant can be challenging, particularly when the function is not known explicitly and can only be accessed through a surrogate model [19, 53]. In recent years, several techniques have been proposed for estimating the Lipschitz constant using surrogate models and linear distance metric functions [30, 38]. These techniques aim to improve the accuracy and applicability of Lipschitz estimation in various domains and scenarios.

In this research, we investigate methods in which we use the nearest evaluated point, similar to that of the Lipschitz bound, for the distance term, but we use a surrogate model instead of the value of the nearest evaluated point. In addition, the mathematical literature considers Hölder bounds as alternative distance metrics within the Lipschitz framework [e.g., 57]. Specifically, the bounds in this research are generally given by:

$$f(\mathbf{x}) \geq \hat{f}(\mathbf{x}^i) - L\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i) \tag{4}$$

in which $\bar{\mathbf{x}}^i$ is the nearest evaluated point. The proposed bound is intended to outperform the Lipschitz bound. First, instead of considering a single evaluated point, we consider a surrogate model $\hat{f}$ that is trained over a set of evaluated points. Moreover, because surrogates are used in some global optimization procedures, such as surrogate optimization [e.g., 14], $\hat{f}$ may be determined and available within an optimization algorithm. In addition, the L-2 norm distance metric may be either too conservative or too aggressive to find a tight bound. Consequently, Hölder-style nonlinear distance metrics are considered. We explain why such an alternative distance metric may provide tighter bounds depending on the underlying function f.

Moreover, the proposed bound is likely to outperform the statistical bound because the reference concerning the center of the data may not be appropriate for an optimization framework. Sampled data within an optimization algorithm often has poorly distributed evaluated points, rendering the center of data an ill-equipped descriptor of the localized region of $\mathbf{x}$ [39, 67]. For example, for an evaluated point $\mathbf{x}^i$, the distance to the center of data $\varphi(\mathbf{x}^i, \bar{\mathbf{x}})$ could be quite large, creating an unnecessarily large difference between $f(\mathbf{x}^i)$ and $\hat{f}(\mathbf{x}^i) - L\varphi(\mathbf{x}^i, \bar{\mathbf{x}})$. By contrast, using the reference point $\bar{\mathbf{x}}^i$ to be the nearest evaluated point, and if $\hat{f}(\mathbf{x}) = f(\mathbf{x})$ at all evaluated points, as in a non-interpolating $\hat{f}$ model, then at an evaluated point, the bound in this research becomes the function value itself since $\varphi(\mathbf{x}, \bar{\mathbf{x}}^i) = 0$.

In this paper, we conduct a comparative theoretical and empirical analysis of using alternative distance metrics for Lipschitz constant estimation. Our theoretical analysis focuses on the use of different distance metrics, including superlinear, linear, and sublinear distance metrics, for estimating Lipschitz constants and computing lower bounds on the objective function. We prove a key result, which establishes the relationship between the lower bounds obtained by these distance metrics on a set of unevaluated data points.

The remainder of this paper is organized as follows. In Section 2, we conduct a thorough literature review, summarizing the most recent advancements in the field. Section 2.1 outlines our research contribution and identifies the research gap it aims to address. The methodology for estimating the Lipschitz constant based on distance metric functions and theoretical development is presented and discussed in Section 3. Section 4 discusses improving lower bound estimates. In Section 5, we present comparative results of our method against approaches utilizing a linear distance metric. These results are obtained through experiments conducted on various benchmark problems. Finally, we conclude the paper by summarizing our contributions and providing insights into potential avenues for future research.

2 LITERATURE REVIEW

A significant body of research has been dedicated to addressing global optimization problems and utilizing the Lipschitz constant for lower-bound construction. In this section, we review relevant studies that have contributed to the understanding and application of Lipschitz constant-based techniques in optimization algorithms.

In the literature, lower bounds are crucial in optimization algorithms as they guide the search for optimal solutions, accelerate convergence, and help avoid local optima [12, 23, 29, 42, 65]. Moreover, lower bounds can be used to determine optimality gaps and serve as stopping criteria to terminate the optimization process when a satisfactory solution is obtained [4, 34, 41]. Lower bounds also play a crucial role in determining the feasible region of an optimization problem [8, 17, 66]. For instance, in branch-and-bound algorithms commonly used for combinatorial optimization problems, lower bounds aid in pruning unpromising branches [5, 46, 60, 61, 66].

In the field of global optimization, several methods have been developed to estimate the lower bound on the objective function within the search space [1, 2]. One popular approach is Convex Relaxation, which involves approximating a non-convex optimization problem with a convex one. By relaxing the constraints or objective function of the problem to a convex form, it becomes possible to obtain a lower bound on the global optimum. Convex relaxation techniques, such as semidefinite programming, linear programming, or convex hull approximation, are commonly employed to construct convex lower bounds [35, 50]. Interval analysis is another widely used technique, which utilizes interval arithmetic to compute bounds on functions. By propagating intervals of values for the decision variables through the objective function and constraints, bounds on the function's values can be obtained. Interval analysis provides guaranteed lower bounds by considering the range of possible values for each variable within given intervals [7, 21, 37].

Statistical methods, such as Working-Hotelling inference, offer an additional approach for estimating lower bounds based on sampled data. These techniques enable the inference and approximation of the lower bound of the objective function using available sample data [6, 56]. Furthermore, Lipschitz continuity-based methods establish lower bounds on the objective function by bounding the rate of change of the function. These methods utilize the Lipschitz constant, which characterizes the local smoothness of the function and quantifies its maximum rate of change between any two points in its domain [18, 25, 26].

Estimating the Lipschitz constant has been the subject of various techniques proposed in the literature. These techniques can be broadly classified into two main approaches: analytical and data-driven. The analytical approach derives an upper bound on the Lipschitz constant based on some known properties or assumptions about the function, such as its derivatives, convexity, or smoothness [22, 31, 32]. For instance, if the function is differentiable, and its gradient is bounded by a constant M, then the Lipschitz constant is at most M. However, this approach may not be feasible or accurate when the function is complex, nonlinear, or unknown. The data-driven approach estimates the Lipschitz constant from data by computing the maximum ratio of the function values and the distances between any two points in the domain [e.g., 48]. This approach does not require any prior knowledge or assumptions about the function, but it may face high computational complexity or poor scalability when the dimension or the number of data points is large.

Several techniques have been proposed to improve the efficiency and accuracy of data-driven Lipschitz estimation methods. One technique is to use sampling strategies to reduce the number of data points or subspaces that need to be evaluated. For example, Fazlyab et al. [11] proposes a branch and bound algorithm that uses adaptive sampling to select promising subspaces based on their lower bounds and prune unpromising ones based on their upper bounds. Another technique is to use convex optimization techniques to formulate and solve the Lipschitz estimation problem as a semidefinite program (SDP) or a second-order cone program (SOCP). For example, Jin, Khajenejad, and Yong [27] present a convex optimization framework that interprets activation functions as gradients of convex potential functions and uses quadratic constraints to describe their properties. This allows them to pose the Lipschitz estimation problem as an SDP that can be adapted to increase either the estimation accuracy or scalability.

These techniques cannot be applied directly to global optimization problems that involve expensive black-box functions. Recent research has focused on Lipschitz constant estimation in such settings. One way to exploit the Lipschitz constant in the global optimization of black-box functions is to use response surfaces such as the Radial Basis Function [e.g., 20] and Gaussian Processes [e.g., 3], which are approximate functions of the objective function [e.g., 38]. By using response surfaces, researchers can estimate the Lipschitz constant and use it to guide the optimization process more efficiently and accurately. In particular, it can be used to identify promising regions in the problem domain as proposed in the diagonal partitioning method [e.g., 54] or space-filling curves [e.g., 55]. More recently, the Lipschitz constant has been estimated using a linear similarity measurement between sampled points. This approximate function is then employed to establish a lower bound for the objective function, which directs the search towards promising regions of the problem space [3, 38]. This approach combines the advantages of surrogate modeling and Lipschitz constant estimation to enhance the efficiency and effectiveness of global optimization techniques.

2.1 Contribution

While previous studies have dedicated efforts to construct lower bounds based on Lipschitz constant estimations, there is a significant research gap in comprehensively evaluating the effectiveness and accuracy of these lower bounds across different problem domains. This study aims to bridge this gap by conducting a systematic and theoretical analysis of the quality of Lipschitz constant-based lower bounds.

In particular, the existing studies often concentrate on linear or piecewise linear approximations for Lipschitz constant estimations [3, 38], neglecting the potential benefits that nonlinear techniques, such as a nonlinear distance metric, may offer. The investigation of nonlinear distance metrics for estimating the Lipschitz constant and their impact on the quality of lower bounds remains unexplored in the optimization context. Exploring the use of nonlinear distance metrics in Lipschitz constant estimation could provide valuable insights into enhancing the accuracy and efficiency of lower bounds, particularly for functions with nonlinear characteristics or complex problem domains. By incorporating a nonlinear distance metric into Lipschitz constant estimation, we can better capture the nonlinear behavior of the underlying function. Nonlinear distance metrics allow for more flexible and expressive representations of the function, enabling the estimation algorithm to capture intricate relationships and variations in the function's smoothness. This leads to more accurate Lipschitz constant estimates, which in turn yields improvements in the quality of the lower bounds. Consequently, optimization algorithms guided by these nonlinear lower bounds can explore the search space more efficiently, avoiding unnecessary evaluations in unpromising regions and focusing on potentially superior solution subspaces.

Therefore, this study aims to address these gaps by promoting Hölder-style nonlinear distance metrics for Lipschitz constant-based lower bounds, providing theoretical properties of them, and conducting an empirical and theoretical evaluation of them across diverse optimization problems. By investigating the quality of these lower bounds and their performance in guiding optimization algorithms, this research will contribute to a deeper understanding of Lipschitz-based techniques and their practical implications.

In this paper, we present a theoretical analysis, focusing on the use of superlinear, linear, and sublinear distance metrics for estimating Lipschitz constants and computing lower bounds on the objective function. Our key theoretical contribution lies in establishing a fundamental result that reveals the relationship between the lower bounds obtained by these distance metrics on a set of unevaluated data points. We demonstrate, under certain assumptions, that the lower bounds obtained using the superlinear, linear, and sublinear distance metric exhibit a specific order: $\hat{f}_{lb}^{sup}(\mathbf{x}) < \hat{f}_{lb}^{lin}(\mathbf{x}) < \hat{f}_{lb}^{sub}(\mathbf{x})$. This order of lower bounds provides insights into the tightness of the Lipschitz constant estimation. We

propose an evaluation metric to assess the quality of lower bounds generated using Lipschitz estimates. Our evaluation metric considers both point-wise evaluations, comparing the reliability of lower bounds at individual data points, and comparative analysis of lower bounds generated by different estimators. By comparing the lower bounds obtained from various approaches, we can identify the most precise and reliable Lipschitz estimator for a specific application.

3 METHODOLOGY

3.1 Problem Formulation

Let $f(\mathbf{x}) : \mathbb{R}^d \to \mathbb{R}$ represent the objective function to be minimized (or maximized) in an optimization problem, and let $\mathbf{x}^* \in \mathbb{R}^d$ be the global optimizer. The global optimizer $\mathbf{x}^*$ satisfies the property that $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for all feasible $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X}$ represents the solution space, which is a set of possible values for the decision variables. The feasible region, denoted as $\Omega \subseteq \mathcal{X}$, encompasses the set of all feasible $\mathbf{x} \in \mathcal{X}$ that satisfy the constraints of the optimization problem. A lower bound, denoted as $f_{lb}(\mathbf{x})$, is a value or function that satisfies the following condition for all feasible $\mathbf{x} \in \Omega \subseteq \mathcal{X}$:

$$f_{lb}(\mathbf{x}) \leq f(\mathbf{x}) \tag{5}$$

In this study, we aim to explore the use of Hölder-style nonlinear distance metrics for estimating the Lipschitz constant and subsequently constructing lower bounds for an objective function. Lipschitz continuity refers to a property of functions where there exists a Lipschitz constant that bounds the function rate of change.

Formally, Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function defined over the d-dimensional Euclidean space. Function f is said to be Lipschitz continuous if there exists a constant $l_f > 0$ such that for any two points $\mathbf{x}^i$ and $\mathbf{x}^j$ in the domain $\mathbb{R}^d$, the following inequality holds:

$$|f(\mathbf{x}^i) - f(\mathbf{x}^j)| \leq l_f \|\mathbf{x}^i - \mathbf{x}^j\| \tag{6}$$

where $\|\cdot\|$ denotes the Euclidean norm (L^2-norm). The Lipschitz constant l_f of the function $f(\mathbf{x})$ defined below represents an upper bound on the slope or rate of change of the function:

$$\hat{l}_f = \sup_{\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}} \frac{|f(\mathbf{x}^i) - f(\mathbf{x}^j)|}{\|\mathbf{x}^i - \mathbf{x}^j\|}, \tag{7}$$

In our approach, we define a lower bound as an approximation of the minimum value of the objective function at a given point. The lower bound at point $\mathbf{x}^i$ is obtained by subtracting the estimated Lipschitz constant $\hat{l}_f$ multiplied by the Euclidean distance between $\mathbf{x}^i$ and a reference point $\mathbf{x}^j$ from the objective function value $f(\mathbf{x}^i)$. Given the unknown nature of the Lipschitz constant, our approach involves the utilization of techniques to estimate its value, $\hat{l}_f$. In addition, estimating l_f using Equation 7 can be challenging, especially in cases where the function $f(\mathbf{x}^i)$ is unknown or can only be evaluated at specific points.

We extend the standard Lipschitz constant estimation framework by considering a more general form of Equation 7,

$$\hat{l}_f = \sup_{\mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}} \frac{f(\mathbf{x}^i) - f(\mathbf{x}^j)}{\varphi(\mathbf{x}^i, \mathbf{x}^j)} \tag{8}$$

where the numerator of the fraction represents the difference in function values between two points $\mathbf{x}^i$ and $\mathbf{x}^j$, while the denominator is determined by the non-negative distance metric $\varphi(\cdot)$. For $\varphi(\cdot) = \|\cdot\|^p$ and $p > 0$, $\hat{l}_f$ in Equation 7 is referred to as a Hölder constant [63]. The distance metric can take on a linear or nonlinear form, providing flexibility in capturing the relationship between input points. Further details on the construction and selection of suitable distance metrics will be discussed in section 3.2.

In practical scenarios, the true objective function $f(\mathbf{x}^i)$ may be unavailable or computationally expensive to evaluate for every pair of points in the input space. Hence, we rely on an estimate of the objective function, denoted as $\hat{f}(\mathbf{x}^i)$, to approximate the function. The use of $\hat{f}(\mathbf{x}^i)$ allows us to overcome the limitations of direct access to $f(\mathbf{x}^i)$ and enables the estimation of the Lipschitz constant.

Our approach leverages a surrogate model that is trained on a limited set of evaluated points $\mathcal{D} = \{X\}$, where the true output values are known. By fitting the surrogate model to the evaluated dataset, we can estimate the lower bound of the underlying function for an unevaluated set of observations $\mathcal{U}$, which is unknown to the trained model.

To estimate the Lipschitz constant, we solve an LP using the set of evaluated points $\mathcal{D}$. Specifically, let $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$ be a partition of the set $\mathcal{D}$. For each $\mathbf{x}^i \in \mathcal{D}$, let $\mathcal{D}^i$ be the subset within the partition that includes the evaluated point $\mathbf{x}^i$, let $\bar{\mathcal{D}}^i$ be the complementary set of evaluated points such that $\bar{\mathcal{D}}^i = \mathcal{D} \setminus \mathcal{D}^i$, and let $\hat{f}_{\bar{\mathcal{D}}^i}$ be a surrogate model trained similarly to that of $\hat{f}$ but with only the evaluated points in $\bar{\mathcal{D}}^i$. Note that the evaluated point $\mathbf{x}^i$ is not in the data that trains $\hat{f}_{\bar{\mathcal{D}}^i}$. Moreover, let $\bar{\mathbf{x}}^i \in \bar{\mathcal{D}}^i$ be a nearest point to $\mathbf{x}^i$ based upon the distance metric φ; that is,

$$\bar{\mathbf{x}}^i \in \arg\min_{\mathbf{x} \in \bar{\mathcal{D}}^i} \varphi(\mathbf{x}^i, \mathbf{x}). \tag{9}$$

Let η_i be a nonnegative slack variable representing the difference between the lower bound at $\mathbf{x}^i$ and its true evaluated function $f(\mathbf{x}^i)$. The LP formulation, as given in Equation 10, seeks to find the Lipshitz constant $\hat{l}_f$ that minimizes the maximum η_i for all evaluated points.

$$\min_{i=1,\ldots,|\mathcal{D}|} \max \eta_i \tag{10a}$$

s.t.

$$\hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i) - \hat{l}_f \varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i) + \eta_i = f(\mathbf{x}^i), \qquad\qquad \forall \mathbf{x}^i \in \mathcal{D}, \tag{10b}$$

$$\eta_i \geq 0, \qquad\qquad \forall i = 1, \ldots, |\mathcal{D}|. \tag{10c}$$

By solving the LP, we obtain estimates of the Lipschitz constant $\hat{l}_f$, which can approximate the lower bound of the function with Equation 11.

$$\hat{f}_{lb}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \hat{l}_f \varphi(\mathbf{x}, \bar{\mathbf{x}}) \approx f_{lb}(\mathbf{x}), \tag{11}$$

where $\bar{\mathbf{x}}$ is a nearest evaluated point to $\mathbf{x}$ in $\mathcal{D}$ based upon the distance metric φ.

This approach allows us to handle cases where the true function evaluations are limited or unavailable, enabling us to estimate the Lipschitz constant using the available surrogate model and evaluated data. Moreover, Equation 11 provides a conservative estimate of the lower bound, which can guide the optimization algorithm in exploring regions with the potential for improving the objective function value while avoiding local optima. It can also help estimate an optimality gap with Equation 12.

$$\text{GAP} = \min_{\mathbf{x} \in \mathcal{D}} f(\mathbf{x}) - \min_{\mathbf{x} \in \Omega} \hat{f}_{lb}(\mathbf{x}). \tag{12}$$

It is worth noting that the quality of the lower bound approximation depends on the accuracy of $\hat{f}$ in representing the true objective function within a certain region of interest. Therefore, the selection and construction of an appropriate surrogate $\hat{f}$ play a crucial role in obtaining reliable and meaningful lower bounds.

In general, we want $\hat{f}$ to have a flexible architecture so that it can accurately represent the true function f. However, highly flexible surrogate models can lead to overfitting the evaluated points and consequently be inaccurate in estimating

unevaluated points. To mitigate overfitting and estimate the variance of surrogate models, cross-validation (CV) is commonly employed in statistics and machine learning, especially when the dataset is limited in size, which is common in surrogate optimization.

By incorporating an approach similar to CV into the Lipschitz constant estimation process, we can account for the uncertainty in the surrogate model's predictions and improve the robustness of the lower bound construction. Specifically, CV in its traditional machine learning setting trains multiple models on different subsets of the data and validates them on the complementary datasets. Similarly, in the LP, multiple models are trained, but the Lipschitz constant is estimated using evaluated points in complementary datasets $\bar{\mathcal{D}}$ as shown in Equation 10b.

Algorithm 1 outlines the steps involved in estimating the Lipschitz constant using distance metrics and surrogate models.

Algorithm 1 Lipschitz Constant Estimation Using Distance Metrics

---

1: Partition $\mathcal{D}$ into $\{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$
2: **for** $k = 1, \ldots, K$ **do**
3:     Train a surrogate model $\hat{f}_{\bar{\mathcal{D}}_k}$ on $\bar{\mathcal{D}}_k$.
4:     **for** $\mathbf{x}^i \in \mathcal{D}_k$ **do**
5:         Calculate $\hat{f}_{\bar{\mathcal{D}}_k}(\mathbf{x}^i)$.
6:         Find $\bar{\mathbf{x}}^i$ from Equation 9.
7:         Calculate $\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)$
8:     **end for**
9: **end for**
10: Solve the linear program (LP) using Equation 10.
11: Return the estimated Lipschitz constant $\hat{l}_f$ [2]

The algorithm iterates over the K subsets within the partition $\mathcal{D}_1, ..., \mathcal{D}_K$. This partitioning allows us to create a surrogate model $\hat{f}_{\bar{\mathcal{D}}_k}$ by fitting it on $\bar{\mathcal{D}}_k$; which is the complementary subset of $\mathcal{D}_k$ such that $\bar{\mathcal{D}}_k = \mathcal{D} \setminus \mathcal{D}_k$. Then for each point $\mathbf{x}^i \in \mathcal{D}_k$, the surrogate model is used to predict the function value at $\mathbf{x}^i$, resulting in $\hat{f}_{\bar{\mathcal{D}}_k}(\mathbf{x}^i)$. Next, the algorithm identifies the closest point to $\mathbf{x}^i$ in $\bar{\mathcal{D}}_k$. This step determines the reference point $\bar{\mathbf{x}}^i$ that is used in the distance metric calculations. The distance metric value $\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)$ is computed, representing the relationship between the current point $\mathbf{x}^i$ and the reference point $\bar{\mathbf{x}}^i$. The process continues for each point in each subset of the partition, resulting in a set of distance metric values for each point. Finally, the algorithm solves the linear program (LP) given in Equation 10. Upon solving the LP, the algorithm returns the estimated Lipschitz constant $\hat{l}_f$.

3.2 Distance Metrics and Similarity Measurements

The Lipschitz constant is a crucial parameter that characterizes the smoothness and regularity of the objective function in optimization problems. It represents the maximum rate of change of the function across its entire domain. The Lipschitz constant, denoted as l, satisfies the Lipschitz condition, which imposes a bound on the difference between the function values at any two points $\mathbf{x}^i$ and $\mathbf{x}^j$ in the search space. Specifically, the Lipschitz condition is defined as below [62]:

$$|f(\mathbf{x}^i) - f(\mathbf{x}^j)| \leq l \|\mathbf{x}^i - \mathbf{x}^j\|, \tag{13}$$

---

[2]Throughout the remainder of the paper, l refers to l_f.

where $\|\mathbf{x}^i - \mathbf{x}^j\|$ represents the Euclidean distance between points $\mathbf{x}^i$ and $\mathbf{x}^j$.

Our study encompasses three distinct types of distance metrics exhibiting diverse linearity properties. Nevertheless, before delving into the definition of each distance metric, it is crucial to examine the concept of linearity properties.

Definition 1. Let $\mathcal{X}$ be a vector space over the set of real numbers $\mathbb{R}$ or complex numbers $\mathbb{C}$. A real-valued function $g : \mathcal{X} \to \mathbb{R}$ is called *superlinear*, *linear*, or *sublinear* if it is non-negative homogeneous and satisfies the respective additivity property [47, 52] as defined in Table 1.

Table 1. Definitions of Linearity Properties

Property	Mathematical Definition	
Non-negative homogeneity	$g(\mathbf{u}.\mathbf{x}^i) = \mathbf{u}.g(\mathbf{x}^i)$ $\quad\quad \forall \mathbf{u} \geq 0, \mathbf{u} \in \mathbb{R},$	$\forall \mathbf{x}^i \in \mathcal{X}$
Subadditivity	$g(\mathbf{x}^i + \mathbf{x}^j) \leq g(\mathbf{x}^i) + g(\mathbf{x}^j)$	$\forall \mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$
Additivity	$g(\mathbf{x}^i + \mathbf{x}^j) = g(\mathbf{x}^i) + g(\mathbf{x}^j)$	$\forall \mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$
Superadditivity	$g(\mathbf{x}^i + \mathbf{x}^j) \geq g(\mathbf{x}^i) + g(\mathbf{x}^j)$	$\forall \mathbf{x}^i, \mathbf{x}^j \in \mathcal{X}$

We can now characterize the general forms of superlinear, linear, and sublinear distance metrics by considering their specific linearity traits, all while relaxing the traditional requirement of triangle inequality from distance metrics. For a given point-wise distance-based metric $\varphi(\mathbf{x}^i, \mathbf{x}^j)$ expressed in the following form:

$$\varphi(\mathbf{x}^i, \mathbf{x}^j) = \alpha\|\mathbf{x}^i - \mathbf{x}^j\|^p \tag{14}$$

Definition 2. A *superlinear distance metric* $\varphi(\mathbf{x}^i, \mathbf{x}^j)$ is a type of distance metric that satisfies Equation (14) where $\mathbf{x}^i$ and $\mathbf{x}^j$ are input vectors, α is a scaling factor, and p is a positive exponent greater than 1. The superlinear distance metric captures complex relationships between the input vectors, and its growth rate is faster than linear, as $p > 1$.

Similarly, we can define *sublinear distance metric* and *linear distance metric* based on the magnitude of the p constant in the Equation 14, in which $p < 1$ would give us the *sublinear distance metric*, and $p = 1$ gives the *linear distance metric*.

By studying these specific distance metrics, we can effectively measure the similarity or distance between input vectors and use them in the context of estimating the Lipschitz constant. Each distance metric has its own characteristics, enabling us to approximate the Lipschitz constant efficiently based on the linearity properties captured by the respective distance metric. The choice of distance metric depends on the specific requirements and properties of the problem at hand.

For instance, the superlinear distance metric captures nonlinear relationships and allows for a faster growth rate than the linear distance metric. This distance metric is suitable for modeling complex relationships between input vectors. On the other hand, the sublinear distance metric gives less weight to larger distances, making it appropriate for sparse and high-dimensional data. The linear distance metric, based on the Euclidean distance, measures similarity by computing the dot product of feature vectors. It is a simple and computationally efficient distance metric that serves as a baseline for comparison with more complex distance metrics. However, its linearity limits its ability to capture non-linearity.

Through the application of these distance metrics and their inherent linear properties, we can proficiently approximate the Lipschitz constant. This constant serves as a key indicator of the smoothness and regularity inherent in the objective

function. Its significance lies in optimization problems, where it enables us to restrict the disparity between the values of the objective function at distinct points within the search space.

The parameters of the distance metric definitions play a crucial role and can lead to different outcomes. While the right-hand side (RHS) of the LP in Equation 10b remains the same for all distance metrics, the left-hand side (LHS) differs in terms of similarity measurement. The φ values, which are derived by applying the corresponding distance metric to the distance between each pair of $(\mathbf{x}^i, \bar{\mathbf{x}}^i)$, contribute to this difference. Consequently, we obtain three distinct Lipschitz constant values, namely l^{sup}, l^{lin}, and l^{sub}. These constants represent the estimated Lipschitz constants derived from the LP. Theorem 3.1 provides a closed-form solution for the Lipschitz constant l based on Equation 10, which we will explain in detail below.

In this study, we will explore three different distance metrics, each with its own set of parameters and properties. It is important to note that practitioners may not necessarily use or define all of these distance metrics, but our comparison aims to provide some guidance for their selection.

Theorem 3.1. *For a given function $f : \mathbb{R}^d \to \mathbb{R}$ with a set of evaluated data points $\mathcal{D}$ and a constructed surrogate model $\hat{f}$, there exists a data point $\mathbf{x}^i \in \mathcal{D}$, where $\eta_i = 0$, that maximizes the fraction:*

$$\hat{l} = \max_{\mathbf{x}^i \in \mathcal{D}} \frac{\hat{f}_{\mathcal{D}^i}(\mathbf{x}^i) - f(\mathbf{x}^i)}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)} \tag{15}$$

where $\bar{\mathbf{x}}^i$ represents the closest point to $\mathbf{x}^i$ from the data set $\mathcal{D}$ given by Equation 9. This maximization provides an optimal solution for the Lipschitz constant estimate $\hat{l}$.

PROOF OF THEOREM 3.1. The proof is given in Appendix A. □

Corollary 3.2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a function for which a set of data is given and a surrogate model is constructed. Let $\hat{l}^{sup}$, $\hat{l}^{lin}$, and $\hat{l}^{sub}$ denote the Lipschitz constant approximations obtained using superlinear, linear, and sublinear distance metrics, respectively. Then, the Lipschitz constant approximations follow the inequality:*

$$\hat{l}^{sup} < \hat{l}^{lin} < \hat{l}^{sub}$$

PROOF OF COROLLARY 3.2. The proof is given in Appendix A. □

3.3 Bound Construction

Once the estimated Lipschitz constant $\hat{l}$ has been obtained using the evaluated data set $\mathcal{D}$, we can use Equation 11 to estimate the lower bound of f for a set of unevaluated data points $\mathcal{U} \subseteq \mathcal{X} \setminus \mathcal{D}$. Observe that from Equation 14, $\varphi(\mathbf{x}, \bar{\mathbf{x}}) = 0$ when $\bar{\mathbf{x}} \in \mathcal{D}$.

Let $\rho(\mathbf{x})$ be defined by Equation 16.

$$\rho(\mathbf{x}) = \hat{l}\varphi(\mathbf{x}, \bar{\mathbf{x}}), \forall \mathbf{x} \in \mathcal{X} \tag{16}$$

Combining Equations 11 and 16, we have Equation 17.

$$\hat{f}_{lb}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \rho(\mathbf{x}), \forall \mathbf{x} \in \mathcal{X} \tag{17}$$

We obtain three distinct $\rho(\mathbf{x})$ functions for each of their respective distance metrics: $\rho^{sup}(\mathbf{x})$, $\rho^{lin}(\mathbf{x})$, and $\rho^{sub}(\mathbf{x})$.

Based on the ordering of the $\rho(\mathbf{x})$, we can assert the following sequence for the lower bounds produced by the three distinct distance metrics, using Corollary 3.2:

Theorem 3.3. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be the objective function, $\mathcal{D}$ be a given set of evaluated data points, $\mathcal{U}$ be a set of unevaluated data points, and $\hat{f}$ be a surrogate model trained on $\mathcal{D}$. Let $\tilde{\mathbf{x}}$ be a unique point where $\tilde{\eta} = 0$ as discussed in proof of Theorem 3.1, and let $\bar{\tilde{\mathbf{x}}}$ be given by Equation 9.*

$$\forall \mathbf{x} \in \mathcal{U} : \ \varphi(\mathbf{x}, \bar{\mathbf{x}}) > \varphi(\tilde{\mathbf{x}}, \bar{\tilde{\mathbf{x}}}) \tag{18}$$

$$\hat{f}_{lb}^{sup}(\mathbf{x}) < \hat{f}_{lb}^{lin}(\mathbf{x}) < \hat{f}_{lb}^{sub}(\mathbf{x}), \tag{19}$$

where f_{lb}^{sup}, f_{lb}^{lin}, and f_{lb}^{sub} are the lower bounds computed using the superlinear, linear, and sublinear distance metrics, respectively.

PROOF OF THEOREM 3.3. The proof is given in Appendix A. □

While this proof is based on the definitions and properties of the superlinear, linear, and sublinear distance metrics, it is important to consider the characteristics and properties of the bounds used for Lipschitz constant estimation. Some of these properties are:

- **Conservative vs. Aggressive:** A *conservative bound* is guaranteed to be a valid bound but may not be tight, while an *aggressive bound* is more likely to be tight but not necessarily guaranteed to be valid. The choice of a bound should balance between conservatism and aggressiveness.
- **Tightness:** *Tight bounds* provide more accurate estimates of the Lipschitz constant. In practice, it is desirable to have bounds that are as tight as possible.
- **Computational Complexity:** Some bounds may be more computationally expensive to compute than others. This can be a consideration when working with large datasets or limited computational resources.
- **Generality:** Some bounds may be more general and applicable to a wider range of functions or data distributions than others. This can be important when working with diverse datasets or when generalization is a priority.

Regarding the characteristics and properties of specific distance metrics used in Lipschitz estimation, it is worth noting that linear distance metrics are widely used in the literature due to their simplicity and desirable properties. However, other distance metrics such as sublinear and superlinear distance metrics can also be useful in certain contexts. For example, the sublinear distance metric can handle sparse data effectively, while the superlinear distance metric can capture nonlinear relationships between data points. It is important to consider that some distance metrics, such as the superlinear distance metric, may be computationally expensive, which can be a limitation in certain applications.

4 LOWER BOUND QUALITY

The quality of the lower bounds constructed using Lipschitz estimates is an essential aspect of assessing their accuracy and usefulness in practical applications. One way to evaluate the quality of the lower bounds is through *point-wise evaluation*. This approach focuses on assessing the accuracy of the lower bound at individual data points. Given a set of data points $\mathcal{D}$, we can compare the true function values $f(\mathbf{x}^i)$ with the corresponding lower bounds $\hat{f}_{lb}(\mathbf{x}^i)$ obtained using Lipschitz estimates. For each point $\mathbf{x}^i \in \mathcal{D}$, we compute the absolute difference between the true function value and the lower bound: $|f(\mathbf{x}^i) - \hat{f}_{lb}(\mathbf{x}^i)|$.

By comparing the lower bounds generated by different estimators or parameter configurations, we can identify which approach produces more aggressive and conservative lower bounds. This comparative analysis allows us to

select the most suitable Lipschitz estimator or parameter setting for a specific application, optimizing the quality of the lower bounds.

To obtain tighter lower bounds, certain adjustments can be considered. One of the possible improvements is enhancing the accuracy of the surrogate model. The surrogate model is used to predict the output for a set of unevaluated observations. Depending on the specific context of the study, various types of surrogate models can be considered. When evaluating the quality of the lower bounds, the objective is to minimize the deviation from the true function values. Equation 17 highlights two key aspects to consider to achieve this objective:

- The accuracy and proximity of the surrogate model predictions to the true function values: The closer the predictions are to the true function values, the smaller the difference between the prediction and the true function value will be.
- The choice of $\rho(\mathbf{x})$ for each data point: Smaller $\rho(\mathbf{x})$ leads to smaller differences between the surrogate model predictions and the estimated lower bound.

It is important to note that improving the accuracy of the surrogate model is a separate discussion. However, based on how the $\rho(\mathbf{x})$ function is determined, some general guidelines can be considered to influence the $\rho(\mathbf{x})$ function. To minimize $\rho(\mathbf{x})$, it is necessary to minimize each component of the expression in Equation 16. Therefore, it is crucial to investigate the behavior of $\rho(\mathbf{x})$ on evaluated and unevaluated points individually.

In the following sections, we analyze the influence of different distance metric choices and different data distributions on the quality of the lower bounds obtained for a given data set. The specific distance metric definitions in Table 2 were chosen for the example to illustrate the necessary linearity characteristics of the distance metric. Although these definitions do not represent the general form given by Equation 14, they capture the essential behavior of the distance metric we are considering.

The superlinear distance metric accentuates small distances by incorporating an exponential term, allowing it to emphasize close similarities between data points. The linear distance metric, on the other hand, directly scales the distance between points. Lastly, the sublinear distance metric applies a logarithmic transformation to the distance, which effectively downplays larger distances.

By using these specific definitions, we can demonstrate the distinctive behaviors of the different distance metric types and their impact on the evaluation of $\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)$. This facilitates a more intuitive understanding of the characteristics of each distance metric and its suitability for specific problems.

Table 2. Distance Metric Definitions

Superlinear distance metric	$\varphi(\mathbf{x}^i, \mathbf{x}^j) = \exp(m\|\mathbf{x}^i - \mathbf{x}^j\|) - 1$
Linear distance metric	$\varphi(\mathbf{x}^i, \mathbf{x}^j) = m\|\mathbf{x}^i - \mathbf{x}^j\|$
Sublinear distance metric	$\varphi(\mathbf{x}^i, \mathbf{x}^j) = \ln(m\|\mathbf{x}^i - \mathbf{x}^j\| + 1)$

4.1 Data Distribution

The data distribution plays a crucial role in this study, particularly in determining the lower bounds. The distribution can be divided into two distinct sections: evaluated and unevaluated data points. The distribution characteristics of these sections directly affect the computation of the Lipschitz constant, as they influence the distance and $\varphi(\mathbf{x}, \bar{\mathbf{x}})$ involved in the process.

Evaluated Data Distribution.

The distribution of the evaluated points $\mathcal{D}$ and their pairwise distances impact the calculation of the Lipschitz constant using Equation 10, which is a key factor in determining the quality of the lower bounds. The evaluated data points serve as a reference for determining the Lipschitz constant, which in turn influences the estimation of the lower bounds for the unevaluated points.

Let us consider a set of evaluated points, denoted as $\mathcal{D}$. Specifically, we focus on a point $\tilde{\mathbf{x}} \in \mathcal{D}$, where $\tilde{\eta} = 0$ as defined in Equation 15. Considering the solution to Equation 15 for estimating the Lipschitz constant (Corollary 3.2), and assuming all other terms in the numerator of this fraction are equal, we observe that the larger the distance (distance metric values), the smaller the Lipschitz constant becomes. Consequently, data sets with a greater spread and exploration of points compared to datasets with more concentrated points will result in a smaller Lipschitz constant. In other words, a well-spread data set, where the evaluated points are distributed across a wider range, leads to a smaller Lipschitz constant.

Unevaluated Data Distribution.

The distribution of unevaluated data points also plays a significant role in assessing the effectiveness of the distance metric on unseen observations. The pairwise distances between the unevaluated points $\mathcal{U}$ and their corresponding closest point in the evaluated set $\mathcal{D}$ are used to calculate $\rho(\mathbf{x})$ as shown in Equation 16.

Using the estimated Lipschitz constant $\hat{l}$ obtained from the evaluated dataset $\mathcal{D}$, a tighter lower bound would be created by subtracting a smaller $\rho(\mathbf{x})$ in Equation 17. Let us consider two unevaluated points $\mathbf{x}^i, \mathbf{x}^j \in \mathcal{U}$. Referring to the general definition of a distance metric in Equation 14, we observe that $\varphi(\mathbf{x}, \bar{\mathbf{x}})$ is monotonically increasing with respect to the Euclidean distance between the two points. Consequently, we can make the following inference:

$$\|\mathbf{x}^i - \bar{\mathbf{x}}^i\| > \|\mathbf{x}^j - \bar{\mathbf{x}}^j\| \implies \varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i) > \varphi(\mathbf{x}^j, \bar{\mathbf{x}}^j) \implies \rho(\mathbf{x}^i) > \rho(\mathbf{x}^j) \tag{20}$$

Consequently, unevaluated data points that have a smaller Euclidean distance from their closest evaluated point will result in tighter lower bounds.

4.2 Evaluation Metric

In this study, we propose an evaluation metric to select the *winning distance metric* based on the lower bound they generate. Specifically, let $\mathbf{x} \in \mathcal{X}$ be a sampled point, and consider two estimated lower bounds $\hat{f}_{lb}^{\varphi_1}$ and $\hat{f}_{lb}^{\varphi_2}$ derived from distance metrics φ_1 and φ_2. Assuming $\hat{f}_{lb}^{\varphi_1}(\mathbf{x}) \neq \hat{f}_{lb}^{\varphi_2}(\mathbf{x})$, we declare φ_1 the *winning distance metric* at $\mathbf{x}$ if any of the conditions in Equation 21 are true.

$$\hat{f}_{lb}^{\varphi_2}(\mathbf{x}) < \hat{f}_{lb}^{\varphi_1}(\mathbf{x}) \leq f(\mathbf{x}), \tag{21a}$$

$$\hat{f}_{lb}^{\varphi_1}(\mathbf{x}) \leq f(\mathbf{x}) < \hat{f}_{lb}^{\varphi_2}(\mathbf{x}), \tag{21b}$$

$$f(\mathbf{x}) < \hat{f}_{lb}^{\varphi_1}(\mathbf{x}) < \hat{f}_{lb}^{\varphi_2}(\mathbf{x}). \tag{21c}$$

The metric considers three different scenarios. First, it determines whether the distance metric produces *valid lower bounds*. If both distance metrics meet this criterion, it selects the distance metric with the smallest gap as the *winning distance metric* as in Equation 21a. Additionally, in the case where neither distance metric estimates a lower bound, the metric still determines the distance metric with the smallest gap as in Equation 21c to be the winner.

4.3 Distance Metric Effects on Lower Bound Construction

In this section, we analyze the influence of different distance metrics on the quality of the lower bounds obtained for a given data set. Variations in $\varphi(\mathbf{x}, \bar{\mathbf{x}})$ values for a given distribution of data points lead to a tighter or wider lower bound on the function value. This variation in the magnitude of the obtained $\varphi(\mathbf{x}, \bar{\mathbf{x}})$ value can happen either due to a different choice of distance metric or a different distribution of data points (As discussed earlier in Section 4.1). The quality of distance metric lower bounds can be closely related to the characteristics of the underlying true function. Therefore, the selection of the choice of the distance metric type should be based on a careful analysis of the specific characteristics of the problem at hand. Mathematical properties such as linearity, superlinearity, or sublinearity of the function can determine the appropriateness of certain distance metrics as lower bounds.

Three different empirical tests and sensitivity analyses have been conducted to assess the performance and generality of different distance metric types in the context of the optimization problem. In particular, we expect a direct relationship between the function attributes and the choice of a suitable distance metric. Meaning that the distance metric bound quality improves as the type of the distance metric and the type of the function match. Moreover, we expect poor performance of the traditional lower bound construction methods against the distance metric bounds in general.

The following are the traditional methodologies that are taken into consideration for this study:

Working-Hotelling (WH) [45]: This method uses the Working-Hotelling formula to construct simultaneous confidence intervals for the mean responses at different levels of the independent variable. The formula is given by:

$$\hat{f}(\mathbf{x}) \pm W \hat{s}_{\hat{f}(\mathbf{x})} \tag{22}$$

where $\hat{f}(\mathbf{x})$ is the estimated mean response at point $\mathbf{x}$, $\hat{s}_{\hat{f}(\mathbf{x})}$ is the estimated standard error of $\hat{f}(\mathbf{x})$, and W is a constant that depends on the number of levels and the desired confidence level.

Standard Lipschitz Estimator [28] (Lipschitz): Consider the problem of finding the global minimum of a function $f(\mathbf{x})$ defined on $\mathcal{X}$. Standard Lipschitzian algorithms assume that there exists a finite bound on the rate of change of the function; that is, for any $\mathbf{x} \in \mathcal{U}$ and its closest point $\bar{\mathbf{x}} \in \mathcal{D}$ there exists a positive constant L, called the Lipschitz constant, such that:

$$|f(\mathbf{x}) - f(\bar{\mathbf{x}})| \leq L \|\mathbf{x} - \bar{\mathbf{x}}\|, \tag{23}$$

where $f(\mathbf{x})$ and $f(\bar{\mathbf{x}})$ are the objective function values for $\mathbf{x} \in \mathcal{U}$ and $\bar{\mathbf{x}} \in \mathcal{D}$ respectively. In this research, we estimate L using Equation 10 but with $\hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i) = f(\bar{\mathbf{x}}^i)$.

To approximate the response surface (as in Equation 10) of the objective function, we employ a Radial Basis Function (RBF) surrogate model, as commonly used in the literature [e.g., 15, 58]. By examining these three examples, we can assess the performance of different distance metrics in terms of their ability to capture the linearity traits of the underlying functions and provide suitable lower bounds. The empirical results will be presented and discussed in the subsequent sections.

Linear Function.

Consider the function $f(\mathbf{x})$ as a piecewise linear function given by Equation 24.

$$f(\mathbf{x}) = \min\{|\mathbf{x} - 4|, |\mathbf{x} + 4|\} \tag{24}$$

Figure 1a represents the graph of the piecewise test function as defined in Equation 24 over the domain interval $[-6, 6]$. In this example, three unevaluated data points around -4 are marked in orange, while seven evaluated data points are marked in blue. Figure 1b demonstrates the lower bounds constructed on the three unevaluated points.

The "True-y" label indicates the true values of the test function at the unevaluated data points. Additionally, the "f_hat" labels indicate the predictions made by the RBF surrogate model for the unevaluated data points. As shown in Figure 1b, the surrogate model accurately predicts the function values of unevaluated data points that are close to the true function values, resulting in overlapping "True-y" and "f_hat" values. The same labeling and presentation approach has been followed for the sublinear and superlinear test functions.
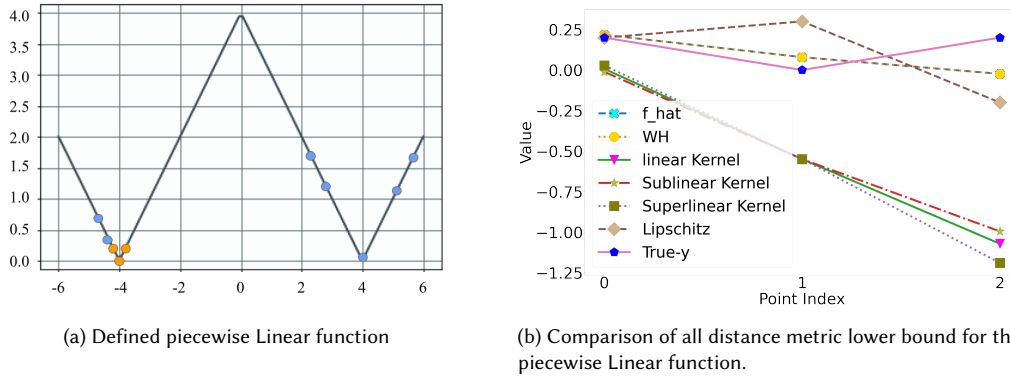

(a) Defined piecewise Linear function

(b) Comparison of all distance metric lower bound for the piecewise Linear function.

Fig. 1. Empirical test case for a Linear function.

Table 3. Distance metric lower bound comparison for Linear test function

	Linear	Sublinear	Superlinear	WH	Lipschitz	Total
Linear	*	2	2	2	2	8
Sublinear	1	*	2	2	2	7
Superlinear	1	1	*	2	2	6
Working - Hotelling	1	1	1	*	2	5
Lipschitz	1	1	1	1	*	4
Total	4	5	6	7	8	*

As depicted in Figure 1b, we can observe that the linear distance metric yields the closest lower bound for the piecewise linear function, while the standard Lipschitz method fails to determine a true lower bound in two of the three unevaluated points. In addition, Table 3 indicates the cumulative number of times each distance metric outperforms the other distance metrics based upon the metric described in section 4.2. Table 3 also indicates that the linear distance

metric produces a better estimated lower bound in eight out of twelve instances compared to the other distance metrics for the unevaluated points in the domain.

Sublinear Function. Consider the piecewise sublinear function defined as $f(\mathbf{x})$ in Equation 25.

$$f(\mathbf{x}) = \min\{\ln{(|\mathbf{x} - 4| + 1)}, \ln{(|\mathbf{x} + 4| + 1)}\} \tag{25}$$

In Figure 2b and Table 4, we observe that the sublinear distance metric produces the most accurate lower bound for the sublinear function as it is designed to capture sublinear behavior.
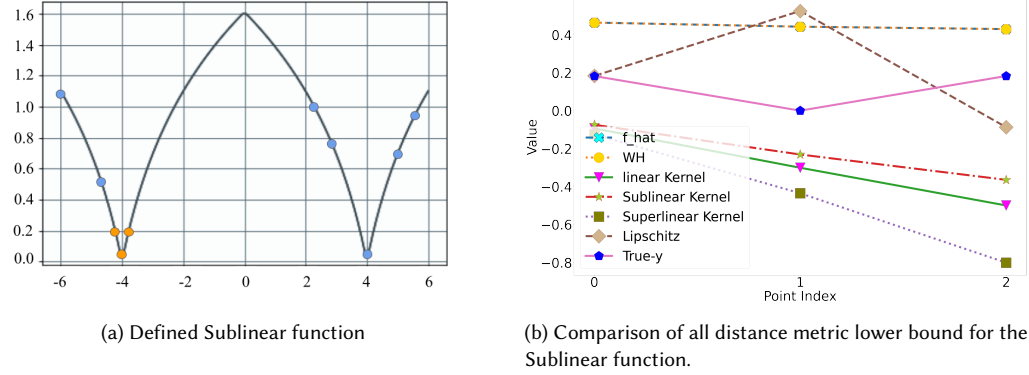

(a) Defined Sublinear function

(b) Comparison of all distance metric lower bound for the Sublinear function.

Fig. 2. Empirical test case for a Sublinear function.

Table 4. Distance metric lower bound comparison for Sublinear test function

	Linear	Sublinear	Superlinear	WH	Lipschitz	Total
Linear	*	0	3	3	2	8
Sublinear	3	*	3	3	2	11
Superlinear	0	0	*	3	2	5
Working - Hotelling	0	0	0	*	1	1
Lipschitz	1	1	1	2	*	5
Total	4	1	7	11	7	*

Superlinear Function.

Let us consider a piecewise-defined superlinear function given by Equation 26.

$$f(\mathbf{x}) = \min\{\exp{(|\mathbf{x} - 4|)} - 1, \exp{(|\mathbf{x} + 4|)} - 1\} \tag{26}$$

As anticipated, the superlinear distance metric is expected to yield a superior lower bound compared to other distance metrics due to its ability to capture superlinear relationships between data points. Superlinear distance metrics, such as the one defined here, are known for their capability to capture complex nonlinear patterns more effectively than linear and sublinear distance metrics. This is particularly evident when examining specific points, as illustrated in Figure 3.

In conclusion, Table 5 demonstrates the significant advantage of the superlinear distance metric over other distance metrics and approaches in generating high-quality lower bounds. The results are based on a dataset comprising
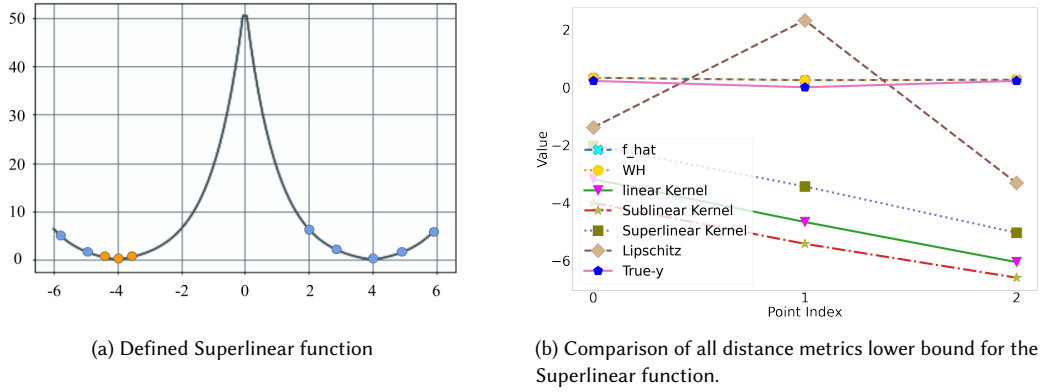
16

(a) Defined Superlinear function

(b) Comparison of all distance metrics lower bound for the Superlinear function.

Fig. 3. Empirical test case for a Superlinear function.

Table 5. Distance metric lower bound comparison for Superlinear test function

	Linear	Sublinear	Superlinear	WH	Lipschitz	Total
Linear	*	3	0	3	1	7
Sublinear	0	*	0	3	1	4
Superlinear	3	3	*	3	1	10
Working - Hotelling	0	0	0	*	1	1
Lipschitz	2	2	2	2	*	8
Total	5	8	2	11	4	*

ten points, with three of them selected as unevaluated points. It is noteworthy that the superlinear distance metric consistently produces lower bounds with a narrower disparity in ten out of twelve instances compared to the other distance metrics, indicating its superior performance in capturing the characteristics of the underlying function.

Ultimately, we can assert that these observations validate our assumption that there is a direct relationship between the characteristics of the underlying function, such as linearity, and the suitability of a certain distance metric for producing accurate lower bounds. The linear distance metric is well-suited for approximating the linear behavior of the function, as it provides a more accurate estimation of the lower bound compared to other distance metrics that may not effectively capture the linearity of the function.

While the sublinear distance metric may generally perform well in capturing sublinear behavior, there can still be cases where it fails to accurately estimate the lower bound for certain instances of a sublinear function. This can be the result of certain traits of the function or distance metric restrictions. For instance, a function may exhibit irregularities, such as sharp jumps or spikes, or it may have local maxima or minima in certain locations, making it difficult for a sublinear distance metric to calculate the lower bound with reliability, resulting in invalid estimates.

As discussed in Section 4.3, it can be observed that the $\varphi(\mathbf{x}, \bar{\mathbf{x}})$ values generated by the superlinear distance metric exhibit greater consistency within their range. Consequently, the superlinear distance metric demonstrates the ability to generate a tighter lower bound for superlinear functions compared to other distance metrics and methods. It is

important to note that the standard Lipschitz method and the Working-Hotelling method fail to find high-quality lower bounds due to their limited flexibility in capturing the nonlinearity aspects of the underlying functions.

5 EXPERIMENTAL ANALYSIS

5.1 Experiment Setup

In this section, we present the experimental analysis of our proposed method for estimating the Lipschitz constant and constructing lower bounds using different distance metrics. Similar to Section 4.3 to approximate the response surface of the objective function, we employ an RBF surrogate model.

To evaluate the performance of our method for estimating the Lipschitz constant and constructing lower bounds using different distance metrics, we use ten benchmark functions from the SFU optimization test problems library [59]. Table 6 provides the details of each test function, including the dimensionality indicated under the "Dimension" column along with their defined domain under the "Domain" column. The known global minimum of each test function is presented under the "Global Minimum" column.

Table 6. Test Functions Definition

Test Function	Dimension	Domain	Global Minimum
Branin	2	$[-5, 10]^2$	0.397887
Six Hump Camel	2	$[-3, 3]^2$	-1.0316
Goldstein-Price	2	$[-2, 2]^2$	3
Shubert	2	$[-5.12, 5.12]^2$	-186.7309
Cross-in-Tray	2	$[-10, 10]^2$	-2.06261
Holder Table	2	$[-10, 10]^2$	-19.2085
Rosenbrock	4	$[-5, 10]^4$	0
Rastrigin	4	$[-5.12, 5.12]^4$	0
Sphere	4	$[-5.12, 5.12]^4$	0
Ackley	4	$[-32.768, 32.678]^4$	0

We conduct 30 random seeded runs for each test function in Table 6, and we determine the best distance metric for each run based on the metric defined by Equation 21. We also report the mean and standard deviation (STD) of the total number of times the winning distance metric has won over 30 different runs at each test problem. In each of the 30 runs, we generate 150 points uniformly in the corresponding domain, as shown in Table 6. We use 100 points for the evaluated set $\mathcal{D}$ to estimate the Lipschitz constant $\hat{l}_f$. The partition of $\mathcal{D}$ is just the set of 100 individual points; that is, $\mathcal{D}^i = \{\mathbf{x}^i\}, \forall i = 1, \ldots, |\mathcal{D}|$. We use the remaining 50 points, to evaluate the winning distance metric and the gaps.

We use a computer with a 2 GHz quad-core Intel Core i5 CPU, Python 3.9.7, and Gurobi Optimizer version 9.5.2 to solve the LP in Equation 10 to run the experiments.

5.2 Results

Tables 7 and 8 show the final results for each of the 10 test functions after 30 runs. Some distance metrics may tie for the highest score after 30 runs, reflecting the same gap tightness. We assign a win to each of these distance metrics, which may result in more than 30 wins across methods. We indicate the scores of the tied distance metric with an asterisk ($\star$).

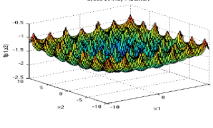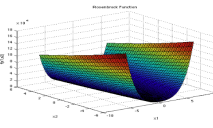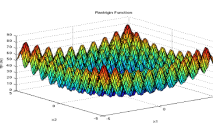Table 7. Test functions with the Linear and Sublinear distance metrics as the winning distance metric
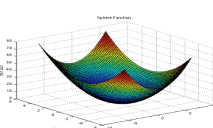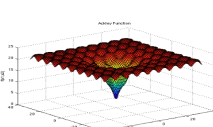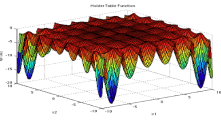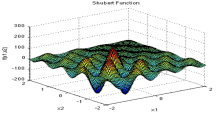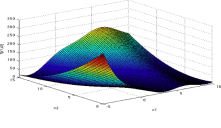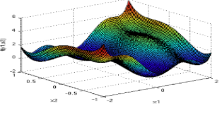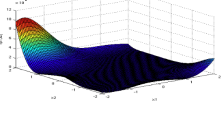
Test Function	Distance Metric	Wins	STD	Mean	Illustration
Cross in Tray	Linear	8	22.80	116.30	
	Sublinear	12	27.99	121.83	
	Superlinear	4	29.07	84.900	
	WH	2	15.79	99.770	
	Lipschitz	4	38.84	77.200	
Rosenbrock	Linear	23*	12.99	144.07	
	Sublinear	6	20.21	124.00	
	Superlinear	1*	22.12	93.670	
	WH	1	19.25	102.50	
	Lipschitz	0	8.860	35.770	
Rastrigin	Linear	15	14.47	134.47	
	Sublinear	7*	22.55	113.30	
	Superlinear	6	23.50	119.13	
	WH	3*	16.94	105.37	
	Lipschitz	0	6.220	27.730	
Sphere	Linear	15	14.79	135.43	
	Sublinear	6	22.39	109.53	
	Superlinear	6	22.74	119.70	
	WH	3	15.84	108.63	
	Lipschitz	0	5.360	26.700	
Ackley	Linear	16	16.82	140.37	
	Sublinear	13	21.28	141.33	
	Superlinear	0	12.83	39.570	
	WH	0	14.39	93.900	
	Lipschitz	1	16.02	84.830	

Table 8. Test functions with the Superlinear distance metric as the winning distance metric

Test Function	Distance Metric	Wins	STD	Mean	Illustration
Holder Table	Linear	7★	8.540	126.47	
	Sublinear	10★	26.96	105.40	
	Superlinear	17★	25.01	134.27	
	WH	1	12.79	108.40	
	Lipschitz	0	3.740	25.467	
Shubert	Linear	5★	7.930	126.93	
	Sublinear	11★	26.16	109.63	
	Superlinear	16★	28.25	134.50	
	WH	0	14.08	102.43	
	Lipschitz	0	4.210	26.500	
Branin	Linear	6	13.06	126.47	
	Sublinear	9	29.71	105.40	
	Superlinear	15	29.18	134.27	
	WH	0	19.17	108.40	
	Lipschitz	0	5.710	25.467	
Camel	Linear	4	8.050	128.70	
	Sublinear	11	32.42	115.63	
	Superlinear	15	30.39	130.53	
	WH	0	14.33	97.530	
	Lipschitz	0	4.850	27.600	
Goldstein	Linear	1	6.970	124.73	
	Sublinear	11	34.31	107.87	
	Superlinear	18	31.22	139.17	
	WH	0	12.96	95.300	
	Lipschitz	0	9.900	32.930	

Based on the results, the superlinear distance metric outperforms the linear and sublinear distance metrics individually in generating lower bounds for half of the benchmark functions as presented in Table 8. This matches the number of times where the linear and sublinear distance metrics together won as shown in Table 7.

The comparative results consistently demonstrate that distance metricized Lipschitz constant estimation outperforms other methods for estimating the Lipschitz constant and constructing lower bounds, including WH and the standard Lipschitz technique. The utilization of the distance metricized Lipschitz constant estimator leads to the construction of tighter and more accurate lower bounds.

In test functions where the global optima are located in flat regions such as Shubert, Branin, Six-Hump Camel, and Goldstein-Price, the superlinear distance metric tends to generate more conservative lower bounds. Consequently, the lower bound produced by the superlinear distance metric for such functions is more general compared to the lower bounds generated by other distance metrics.

For test functions where the global optima occurs on the boundaries of the function, such as the Holder Table, the superlinear distance metric demonstrates greater flexibility in capturing those minimum values.

Table 9. Benchmark test function results (STD and Mean of the number of times the winning distance metric has won)

Test Function	Best Lower Bound	Score out of 30	STD	Mean
ROSENBROCK	Linear	23*	12.99	144.07
RASTRIGIN	Linear	15	14.47	134.47
SPHERE	Linear	15	14.79	135.43
ACKLEY	Linear	16	16.82	140.37
C.T.F.	Sublinear	12	27.99	121.83
HOLDER TABLE	Superlinear	17*	25.01	134.27
SHUBERT	Superlinear	16*	28.25	134.50
BRANIN	Superlinear	15	29.18	135.20
CAMEL	Superlinear	15	30.39	130.53
GOLDSTEIN-PRICE	Superlinear	18	31.22	139.17

In functions that exhibit a consistent pattern of change in the function value (either increasing or decreasing), the lower bounds obtained by the linear and sublinear distance metrics are more general. However, when it comes to establishing a lower bound, the sublinear distance metric is more aggressive than the linear distance metric. This means that for functions with characteristics similar to sublinear functions, like the Rosenbrock function, the linear distance metric outperforms the sublinear distance metric in terms of the number of valid lower bounds produced (satisfying Equation 21a). Although the sublinear distance metric generates lower bounds for fewer points, those lower bounds are tighter compared to the linear distance metric's lower bounds for the same points. Consequently, the linear distance metric's lower bound can be considered more general and conservative, while the sublinear distance metric is more aggressive and provides tighter lower bounds.

Furthermore, analyzing Table 9 and observing the standard deviation (STD) of the total number of times the winning distance metric has won for each test function reveals that the linear distance metric exhibits less volatility in generating lower bounds across the 30 executions with different random data distributions. This characteristic can be attributed to the complexity associated with the properties of the test functions and their sensitivity to variations in data distribution.
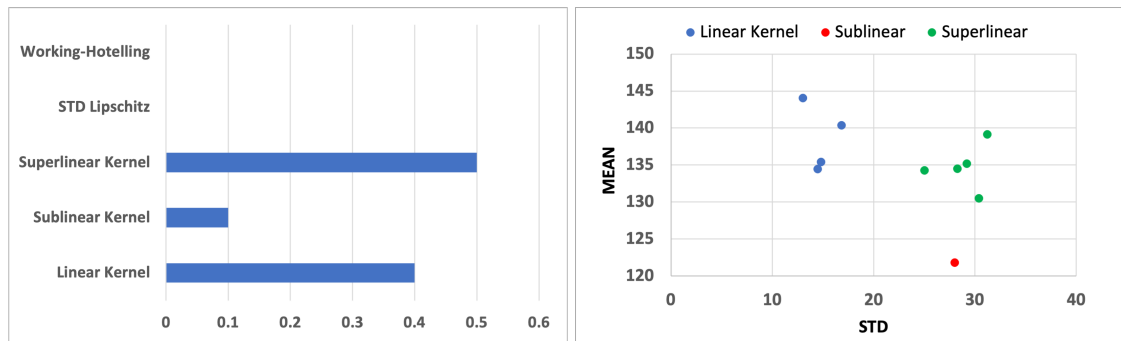
(a) Fraction of problems each distance metric won

(b) STD and Mean of the number of times the winning distance metric has won

Fig. 4. Comparison of STD, Mean, and Score of Winning Distance Metric

Figure 4b illustrates the relationship between the standard deviation and mean of the total number of times the winning distance metric has won over 30 different runs at each test problem. It can be observed that the lower bound generated by the superlinear distance metric exhibits a higher standard deviation. This increased volatility causes the distance metric to occasionally fail to produce a lower bound in cases where other more aggressive distance metrics, like the linear distance metric, can create a tighter gap between their lower bound and the true function value. This behavior can be interpreted as the superlinear distance metric being conservative, as it allows for more fluctuation around the true function value and provides room for correction.

Conversely, the linear distance metric tends to be more aggressive, generating lower bounds with higher means and lower standard deviations. By staying closer to the true function value, the linear distance metric leaves little room for the lower bound to adjust across the entire domain. Consequently, we can conclude that the linear distance metric is more aggressive compared to other distance metrics, such as the superlinear distance metric, which exhibits a more conservative and general behavior across a wider range of functions.

Figure 4a illustrates the fraction of problems in which each distance metric-derived lower bound outperformed the others across all test functions. The results indicate the superiority of our utilized distance metric. Specifically, the Superlinear distance metric achieved a winning fraction of 50%, followed by the Linear distance metric with 40%, and the Sublinear distance metric with 10%. These values significantly outperformed both the standard Lipschitz algorithm and the lower bounds generated by the Working-Hotelling method, highlighting the effectiveness of our approach.

6 CONCLUSION AND FUTURE DIRECTION

In this paper, we have presented a novel approach for estimating the Lipschitz constant and constructing lower bounds using various Hölder-style distance metrics. Our method utilizes a surrogate model to approximate the response surface of the objective function and compares the performance of three types of distance metrics: superlinear, sublinear, and linear to estimate the Lipschitz constant. Through experiments on various benchmark test functions, we have evaluated the effectiveness of each method by measuring the gap between the lower bound and the true function value for a set of unevaluated observations. Additionally, we have identified the best distance metric for each run based on this metric.

Our findings demonstrate the superiority of our proposed method in estimating the Lipschitz constant and constructing lower bounds when compared to alternative approaches such as the Working-Hotelling statistical method and the standard Lipschitz method. We have observed that the superlinear distance metric generally provides more conservative and versatile results across a broader range of functions, while the linear distance metric tends to be more aggressive, yielding tighter lower bounds for certain functions. Consequently, we have concluded that the choice of distance metric should depend on the characteristics of the objective function and the desired trade-off between tightness and generality.

Looking ahead, there are two key directions for future research. Firstly, we intend to explore the application of alternative distance metrics, such as polynomial and Gaussian distance metrics, to further investigate their impact on Lipschitz estimation and lower bound construction. This exploration will provide a more comprehensive understanding of the relationship between distance metrics and their performance in our approach. Additionally, we plan to apply our method to real-world optimization problems, specifically focusing on black-box optimization, in various domains including engineering, machine learning, and other relevant fields. By applying our method to these practical scenarios, we will assess its utility and effectiveness in solving complex optimization challenges.

In conclusion, our proposed method offers a novel and effective approach for estimating the Lipschitz constant and constructing lower bounds. By leveraging surrogate models and comparing different Hölder-style distance metrics,

we have demonstrated superior performance over existing methods. With future enhancements and applications, we anticipate that our method will contribute to advancements in optimization techniques and find practical utility in a wide range of domains.

7 ACKNOWLEDGEMENTS

8 STATEMENTS AND DECLARATIONS

The authors have no relevant financial or non-financial interests to disclose.

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

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

The authors have no financial or proprietary interests in any material discussed in this article.

Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

REFERENCES

[1] Claire S Adjiman, Ioannis P Androulakis, Costas D Maranas, and Christodoulos A Floudas. 1996. A global optimization method, αBB, for process design. *Computers & Chemical Engineering* 20 (1996), S419–S424.

[2] Alekh Agarwal, Martin J Wainwright, Peter Bartlett, and Pradeep Ravikumar. 2009. Information-theoretic lower bounds on the oracle complexity of convex optimization. *Advances in Neural Information Processing Systems* 22 (2009).

[3] Mohamed Osama Ahmed, Sharan Vaswani, and Mark Schmidt. 2020. Combining Bayesian optimization and Lipschitz optimization. *Machine Learning* 109 (2020), 79–102.

[4] Yossi Arjevani, Yair Carmon, John C Duchi, Dylan J Foster, Nathan Srebro, and Blake Woodworth. 2023. Lower bounds for non-convex stochastic optimization. *Mathematical Programming* 199, 1-2 (2023), 165–214.

[5] Fani Boukouvala, Ruth Misener, and Christodoulos A Floudas. 2016. Global optimization advances in mixed-integer nonlinear programming, MINLP, and constrained derivative-free optimization, CDFO. *European Journal of Operational Research* 252, 3 (2016), 701–727.

[6] Peter Bubenik et al. 2015. Statistical topological data analysis using persistence landscapes. *J. Mach. Learn. Res.* 16, 1 (2015), 77–102.

[7] Leocadio G Casado, José A MartÍnez, Inmaculada GarcÍa, and Ya D Sergeyev. 2003. New interval analysis support functions using gradient information in a global minimization algorithm. *Journal of Global Optimization* 25 (2003), 345–362.

[8] Hoon Sung Chwa and Jinkyu Lee. 2023. Tight necessary feasibility analysis for recurring real-time tasks on a multiprocessor. *Journal of Systems Architecture* 135 (2023), 102808.

[9] Aurore Delaigle, Peter Hall, and Farshid Jamshidi. 2015. Confidence bands in non-parametric errors-in-variables regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 77, 1 (2015), 149–169.

[10] William R Esposito and Christodoulos A Floudas. 2000. Deterministic global optimization in nonlinear optimal control problems. *Journal of global optimization* 17 (2000), 97–126.

[11] Mahyar Fazlyab, Alexander Robey, Hamed Hassani, Manfred Morari, and George Pappas. 2019. Efficient and accurate estimation of lipschitz constants for deep neural networks. *Advances in Neural Information Processing Systems* 32 (2019).

[12] Christodoulos A Floudas. 2013. Deterministic global optimization: theory, methods and applications.

[13] Christodoulos A Floudas and Panos M Pardalos. 2013. State of the art in global optimization: computational methods and applications.

[14] Alexander IJ Forrester and Andy J Keane. 2009. Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences* 45, 1-3 (2009), 50–79.

[15] Alexander IJ Forrester, András Sóbester, and Andy J Keane. 2007. Multi-fidelity optimization via surrogate modelling. *Proceedings of the Royal Society Mathematical, Physical and Engineering Sciences* 463, 2088 (2007), 3251–3269.

[16] Mario Francisco-Fernandez and Alejandro Quintela-del Rio. 2016. Comparing simultaneous and pointwise confidence intervals for hydrological processes. *PLoS One* 11, 2 (2016), e0147505.

[17] Hormozd Gahvari and William Gropp. 2010. An introductory exascale feasibility study for FFTs and multigrid. In *2010 IEEE International Symposium on Parallel & Distributed Processing (IPDPS)*. IEEE, https://doi.org/10.1109/IPDPS.2010.5470417, 1–9.

[18] Henry Gouk, Eibe Frank, Bernhard Pfahringer, and Michael J Cree. 2021. Regularisation of neural networks by enforcing lipschitz continuity. *Machine Learning* 110 (2021), 393–416.

[19] Jean-Bastien Grill, Michal Valko, and Rémi Munos. 2015. Black-box optimization of noisy functions with unknown smoothness. *Advances in Neural Information Processing Systems* 28 (2015), 1–9.

[20] H-M Gutmann. 2001. A radial basis function method for global optimization. *Journal of Global Optimization* 19, 3 (2001), 201–227.

[21] Eldon Hansen and G William Walster. 2003. *Global optimization using interval analysis: revised and expanded*. CRC Press, https://doi.org/10.1201/9780203026922. 285–418 pages.

[22] Pierre Hansen and Brigitte Jaumard. 1995. *Lipschitz optimization*. Springer, https://doi.org/10.1007/978-1-4615-2025-2_9.

[23] Reiner Horst and Panos M Pardalos. 2013. *Handbook of global optimization*. Springer Science & Business Media, https://doi.org/10.1007/978-1-4615-2025-2.

[24] Hao Huang, Pariyakorn Maneekul, Danielle F Morey, Zelda B Zabinsky, and Giulia Pedrielli. 2022. A Computational Study of Probabilistic Branch and Bound with Multilevel Importance Sampling. In *2022 Winter Simulation Conference (WSC)*. IEEE, https://doi.org/10.1109/WSC57314.2022.10015267, 3251–3262.

[25] Martin Hutzenthaler, Arnulf Jentzen, and Peter E. Kloeden. 2012. Strong convergence of an explicit numerical method for SDEs with nonglobally Lipschitz continuous coefficients. *The Annals of Applied Probability* 22, 4 (2012), 1611 – 1641. https://doi.org/10.1214/11-AAP803

[26] Jin-ichi Itoh and Minoru Tanaka. 2001. The Lipschitz continuity of the distance function to the cut locus. *Trans. Amer. Math. Soc.* 353, 1 (2001), 21–40.

[27] Zeyuan Jin, Mohammad Khajenejad, and Sze Zheng Yong. 2020. Data-driven model invalidation for unknown Lipschitz continuous systems via abstraction. In *2020 American Control Conference (ACC)*. IEEE, https://doi.org/10.23919/acc45564.2020.9147725, 2975–2980.

[28] Donald R Jones, Cary D Perttunen, and Bruce E Stuckman. 1993. Lipschitzian optimization without the Lipschitz constant. *Journal of Optimization Theory and Applications* 79 (1993), 157–181.

[29] Ramkumar Karuppiah and Ignacio E Grossmann. 2006. Global optimization for the synthesis of integrated water systems in chemical processes. *Computers & Chemical Engineering* 30, 4 (2006), 650–673.

[30] Jakub Kudela and Radomil Matoušek. 2023. Combining Lipschitz and RBF surrogate models for high-dimensional computationally expensive problems. *Information Sciences* 619 (2023), 457–477.

[31] Dmitri Evgenievich Kvasov and Yaroslav Dmitrievich Sergeyev. 2003. A multidimensional global optimization algorithm based on adaptive diagonal curves. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki* 43, 1 (2003), 42–59.

[32] Dmitri E Kvasov and Yaroslav D Sergeyev. 2009. A univariate global search working with a set of Lipschitz constants for the first derivative. *Optimization Letters* 3, 2 (2009), 303–318.

[33] Dmitri E Kvasov and Ya D Sergeyev. 2013. Lipschitz global optimization methods in control problems. *Automation and Remote Control* 74 (2013), 1435–1448.

[34] Guanghui Lan. 2012. An optimal method for stochastic composite optimization. *Mathematical Programming* 133, 1-2 (2012), 365–397.

[35] Leo Liberti. 2004. Reformulation and convex relaxation techniques for global optimization. *Quarterly Journal of the Belgian, French and Italian Operations Research Societies* 2, 3 (2004), 255–258.

[36] Leo Liberti. 2008. Introduction to global optimization. *Ecole Polytechnique* 1 (2008), 1–43.

[37] Youdong Lin and Mark A Stadtherr. 2007. Deterministic global optimization of nonlinear dynamic systems. *AIChE Journal* 53, 4 (2007), 866–875.

[38] Haitao Liu, Shengli Xu, Ying Ma, and Xiaofang Wang. 2015. Global optimization of expensive black box functions using potential Lipschitz constants and response surfaces. *Journal of Global Optimization* 63, 2 (2015), 229–251.

[39] Yiping Liu, Dunwei Gong, Xiaoyan Sun, and Yong Zhang. 2017. Many-objective evolutionary optimization based on reference points. *Applied Soft Computing* 50 (2017), 344–355.

[40] Marco Locatelli and Fabio Schoen. 2013. *Global optimization: theory, algorithms, and applications*. SIAM, https://doi.org/10.1137/1.9781611972672.

[41] AV Lyamin and SW Sloan. 2002. Lower bound limit analysis using non-linear programming. *Internat. J. Numer. Methods Engrg.* 55, 5 (2002), 573–611.

[42] Kaiwen Ma, Luis Miguel Rios, Atharv Bhosekar, Nikolaos V Sahinidis, and Sreekanth Rajagopalan. 2023. Branch-and-Model: a derivative-free global optimization algorithm. *Computational Optimization and Applications* 85, 2 (2023), 337–367.

[43] Cédric Malherbe and Nicolas Vayatis. 2017. Global optimization of Lipschitz functions. In *International Conference on Machine Learning*. PMLR, https://proceedings.mlr.press/v70/malherbe17a.html, 2314–2323.

[44] Nadia Martinez, Hadis Anahideh, Jay M Rosenberger, Diana Martinez, Victoria CP Chen, and Bo Ping Wang. 2017. Global optimization of non-convex piecewise linear regression splines. *Journal of Global Optimization* 68 (2017), 563–586.

[45] Rupert G. Miller. 2012. *Simultaneous statistical inference*. Springer Science & Business Media, https://doi.org/10.1007/978-1-4613-8122-8.

[46] Ruth Misener and Christodoulos A Floudas. 2014. ANTIGONE: algorithms for continuous/integer global optimization of nonlinear equations. *Journal of Global Optimization* 59, 2-3 (2014), 503–526.

[47] Lawrence Narici and Edward Beckenstein. 2010. *Topological vector space*. CRC Press, https://doi.org/10.1201/9781584888673.

[48] Sebastian A Nugroho, Ahmad F Taha, and Vu Hoang. 2021. Nonlinear dynamic systems parameterization using interval-based global optimization: Computing lipschitz constants and beyond. *IEEE Trans. Automat. Control* 67, 8 (2021), 3836–3850.

[49] János D Pintér. 2002. Global optimization: software, test problems, and applications. *Handbook of Global Optimization* 2, 2 (2002), 515–569.

[50] David M Rosen, Charles DuHadway, and John J Leonard. 2015. A convex relaxation for approximate global optimization in simultaneous localization and mapping. In *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, https://doi.org/10.1109/icra.2015.7140014, 5822–5829.

[51] Todd Rowland and Eric W Weisstein. 2016. Lipschitz Function. https://mathworld.wolfram.com/LipschitzFunction.html.

[52] Eric Schechter. 1996. *Handbook of Analysis and its Foundations*. Academic Press, https://doi.org/10.1016/b978-012622760-4/50011-x.

[53] Yaroslav D Sergeyev, Antonio Candelieri, Dmitri E Kvasov, and Riccardo Perego. 2020. Safe global optimization of expensive noisy black-box functions in the δ-Lipschitz framework. *Soft Computing* 24, 23 (2020), 17715–17735.

[54] Yaroslav D Sergeyev and Dmitri E Kvasov. 2006. Global search based on efficient diagonal partitions and a set of Lipschitz constants. *SIAM Journal on Optimization* 16, 3 (2006), 910–937.

[55] Yaroslav D Sergeyev, Paolo Pugliese, and Domenico Famularo. 2003. Index information algorithm with local tuning for solving multidimensional global optimization problems with multiextremal constraints. *Mathematical Programming* 96, 3 (2003), 489–512.

[56] Alexander Shapiro. 2003. Statistical inference of multistage stochastic programming problems. *Math. Methods of Oper. Res* 58 (2003), 57–68.

[57] Nicola Soave, Hugo Tavares, Susanna Terracini, and Alessandro Zilio. 2016. Hölder bounds and regularity of emerging free boundaries for strongly competing Schrödinger equations with nontrivial grouping. *Nonlinear Analysis* 138 (2016), 388–427.

[58] Xueguan Song, Liye Lv, Wei Sun, and Jie Zhang. 2019. A radial basis function-based multi-fidelity surrogate model: exploring correlation between high-fidelity and low-fidelity models. *Structural and Multidisciplinary Optimization* 60 (2019), 965–981.

[59] Sonja Surjanovic and Derek Bingham. 2013. Virtual Library of Simulation Experiments, Test Functions and Datasets. https://www.sfu.ca/~ssurjano/optimization.html. [Last accessed on 18-April-2023].

[60] Mohit Tawarmalani and Nikolaos V Sahinidis. 2004. Global optimization of mixed-integer nonlinear programs: A theoretical and computational study. *Mathematical Programming* 99, 3 (2004), 563–591.

[61] Mohit Tawarmalani and Nikolaos V Sahinidis. 2005. A polyhedral branch-and-cut approach to global optimization. *Mathematical Programming* 103, 2 (2005), 225–249.

[62] GR Wood and BP Zhang. 1996. Estimation of the Lipschitz constant of a function. *Journal of Global Optimization* 8 (1996), 91–103.

[63] Junru Wu. 2020. On a linearity between fractal dimension and order of fractional calculus in Hölder space. *Appl. Math. Comput.* 385 (2020), 125433.

[64] Henry P Wynn and P Bloomfield. 1971. Simultaneous confidence bands in regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological)* 33, 2 (1971), 202–217.

[65] Zelda B Zabinsky, Robert L Smith, and Birna P Kristinsdottir. 2003. Optimal estimation of univariate black-box Lipschitz functions with upper and lower error bounds. *Computers & Operations Research* 30, 10 (2003), 1539–1553.

[66] Jianyuan Zhai and Fani Boukouvala. 2022. Data-driven spatial branch-and-bound algorithms for box-constrained simulation-based optimization. *Journal of Global Optimization* 82, 2 (2022), 21–50.

[67] Jialing Zhou, Yuezu Lv, Changyun Wen, and Guanghui Wen. 2022. Solving specified-time distributed optimization problem via sampled-data-based algorithm. *IEEE Transactions on Network Science and Engineering* 9, 4 (2022), 2747–2758.

A PROOFS OF THEOREMS AND COROLLARIES

PROOF OF THEOREM 3.1. We aim to prove that there exists a data point $\mathbf{x}^i \in \mathcal{D}$ such that $\eta_i = 0$ in the LP formula, and it maximizes the fraction in Equation 15.

Let $\tilde{\eta}$ be the maximum of all η_i in Equation 10. We then rewrite the LP formula as follows:

$$\min \tilde{\eta} : \tag{27a}$$

$$\text{s.t.}$$

$$- \hat{l}_f \varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i) + \eta_i = f(\mathbf{x}^i) - \hat{f}_{\tilde{\mathcal{D}}^i}(\mathbf{x}^i) \qquad \forall i = 1, \ldots, |\mathcal{D}| \tag{27b}$$

$$\tilde{\eta} \geq \eta_i \qquad \forall i = 1, \ldots, |\mathcal{D}| \tag{27c}$$

$$\eta_i \geq 0 \qquad \forall i = 1, \ldots, |\mathcal{D}| \tag{27d}$$

Let y_i be the dual of Equation 27b, and let π_i be the dual of the slack between $\tilde{\eta}$ and all η_i in Equation 27c. To establish the existence of a data point $\mathbf{x}^i$ such that $\eta_i = 0$, we introduce the complementary slackness as follows:

$$(y_i - \pi_i)\eta_i = 0 \qquad \forall i = 1, \ldots, |\mathcal{D}| \tag{28a}$$

$$(\tilde{\eta} - \eta_i)\pi_i = 0 \qquad \forall i = 1, \ldots, |\mathcal{D}| \tag{28b}$$

We now write the dual problem:

$$\max \qquad \sum_{i=1}^{|\mathcal{D}|} (f(\mathbf{x}^i) - \hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i))y_i \tag{29a}$$

$$\text{s.t.}$$

$$y_i - \pi_i \leq 0 \qquad \forall i = 1, \ldots, |\mathcal{D}| \qquad (\text{constraint for } \eta_i) \tag{29b}$$

$$\sum_{i=1}^{|\mathcal{D}|} -\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)y_i = 0 \qquad (\text{constraint for } \hat{l}_f) \tag{29c}$$

$$\sum_{i=1}^{|\mathcal{D}|} \pi_i = 1 \qquad \forall i = 1, \ldots, |\mathcal{D}| \qquad (\text{constraint for } \tilde{\eta}) \tag{29d}$$

$$\pi_i \geq 0 \qquad \forall i = 1, \ldots, |\mathcal{D}|. \tag{29e}$$

Now, assume to the contrary that $\eta_i > 0$ for all $i = 1, \ldots, |\mathcal{D}|$. According to Equation 28a, we have $y_i = \pi_i$ for all $i = 1, \ldots, |\mathcal{D}|$.

Therefore, by constraints 29d and 29e,

$$y_i \geq 0 \quad \text{and} \quad \sum_{i}^{|\mathcal{D}|} y_i = 1 \qquad \forall i = 1, \ldots, |\mathcal{D}|. \tag{30a}$$

Considering the values of $\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)$ and the fact that $\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)$ is always nonnegative, we have:

$$\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i) > 0 \rightarrow -\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i) < 0 \qquad \forall i = 1, \ldots, |\mathcal{D}| \tag{31a}$$

$$\xrightarrow[\exists y_i > 0]{\text{Equation 30a}} -\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)y_i < 0 \Rightarrow \sum_{i}^{|\mathcal{D}|} -\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)y_i < 0 \tag{31b}$$

However, this contradicts Equation 29c.

Now that we have established the existence of a data point $\mathbf{x}^i \in \mathcal{D}$ where $\eta_i = 0$ in the LP formula, our goal is to demonstrate that the fraction in Equation 15 achieves its maximum value at $\mathbf{x}^i$. Assume $\tilde{\mathbf{x}}$ represents the point where $\tilde{\eta} = 0$, and $\mathbf{x}^i \in \mathcal{D}$ represents another point within the set of evaluated data points.

Recalling Equation 10b we have:

$$\hat{l}_f = \frac{\hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i) - f(\mathbf{x}^i) - \eta_i}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)} \rightarrow \frac{\hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i) - f(\mathbf{x}^i)}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)} - \frac{\eta_i}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)}, \forall i = 1, \ldots, |\mathcal{D}| \tag{32}$$

Since we know that $\eta_i \geq 0$ and $\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i) > 0$, it follows that:

$$\frac{\eta_i}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)} \geq 0 \implies \frac{\hat{f}_{\bar{\mathcal{D}}^i}(\tilde{\mathbf{x}}) - f(\tilde{\mathbf{x}})}{\varphi(\tilde{\mathbf{x}}, \bar{\tilde{\mathbf{x}}})} \geq \frac{\hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i) - f(\mathbf{x}^i)}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)} - \frac{\eta_i}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)} \tag{33}$$

We can observe that for $\tilde{\mathbf{x}} \in \mathcal{D}$ where $\tilde{\eta} = 0$, Equation 15 is always greater than the right-hand side (RHS) for other points such as $\mathbf{x}^i$ where $\eta_i > 0$, as depicted in Equation 33. Thus, the claim holds true. □

PROOF OF COROLLARY 3.2. We assume that the surrogate model is trained on the given data $\mathcal{D}$ and has an accuracy sufficient to provide reasonable Lipschitz constant approximations. Using Theorem 3.1, we know that the Lipschitz constant $\hat{l}$ for each distance metric is given by:

$$\hat{l} = \max_{\mathbf{x}^i \in \mathcal{D}} \frac{\hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i) - f(\mathbf{x}^i)}{\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)} \tag{34}$$

Without loss of generality, we can consider the numerator $(\hat{f}_{\bar{\mathcal{D}}^i}(\mathbf{x}^i) - f(\mathbf{x}^i))$ to be the same for all the distance metrics. Moreover, since each distance metric is increasing by Equation 14, the maximum points are the same for all the distance metrics; that is,

$$\tilde{\mathbf{x}} \in \arg\max_{\mathbf{x}^i \in \mathcal{D}} \varphi^{\sup}(\mathbf{x}^i, \bar{\mathbf{x}}^i) = \arg\max_{\mathbf{x}^i \in \mathcal{D}} \varphi^{\lin}(\mathbf{x}^i, \bar{\mathbf{x}}^i) = \arg\max_{\mathbf{x}^i \in \mathcal{D}} \varphi^{\sub}(\mathbf{x}^i, \bar{\mathbf{x}}^i) \tag{35}$$

Because the numerator of Equation 34 is independent of the distance metric, we have the following equality.

$$\hat{l}^{\sup}\varphi^{\sup}(\tilde{\mathbf{x}}, \bar{\tilde{\mathbf{x}}}) = \hat{l}^{\lin}\varphi^{\lin}(\tilde{\mathbf{x}}, \bar{\tilde{\mathbf{x}}}) = \hat{l}^{\sub}\varphi^{\sub}(\tilde{\mathbf{x}}, \bar{\tilde{\mathbf{x}}}) \tag{36}$$

Thus, the ordering of the $\hat{l}$ values will be determined by the denominator $\varphi(\mathbf{x}^i, \bar{\mathbf{x}}^i)$.

Based on the definition of the distance metric, we have:

$$\varphi^{\sup}(\mathbf{x}^i, \bar{\mathbf{x}}^i) > \varphi^{\lin}(\mathbf{x}^i, \bar{\mathbf{x}}^i) > \varphi^{\sub}(\mathbf{x}^i, \bar{\mathbf{x}}^i) \tag{37}$$

Therefore, the ordering of the Lipschitz constant approximations $\hat{l}^{sup}$, $\hat{l}^{lin}$, and $\hat{l}^{sub}$ will be opposite to the ordering of the φ values:

$$\hat{l}^{\sup} < \hat{l}^{\lin} < \hat{l}^{\sub} \tag{38}$$

□

PROOF OF THEOREM 3.3. Recalling from Equations 16 and 17 we construct lower bounds as follows:

$$\rho^{\sup}(\mathbf{x}) = \hat{l}^{\sup} \cdot \varphi^{\sup}(\mathbf{x}, \bar{\mathbf{x}}) \longrightarrow \hat{f}_{lb}^{\sup}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \rho^{\sup}(\mathbf{x})$$
$$\rho^{\lin}(\mathbf{x}) = \hat{l}^{lin} \cdot \varphi^{\lin}(\mathbf{x}, \bar{\mathbf{x}}) \longrightarrow \hat{f}_{lb}^{\lin}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \rho^{\lin}(\mathbf{x})$$
$$\rho^{\sub}(\mathbf{x}) = \hat{l}^{\sub} \cdot \varphi^{sub}(\mathbf{x}, \bar{\mathbf{x}}) \longrightarrow \hat{f}_{lb}^{\sub}(\mathbf{x}) = \hat{f}(\mathbf{x}) - \rho^{\sub}(\mathbf{x})$$

Based upon Equation 36 from Corollary 3.2, we the following equality.

$$\rho^{\sup}(\tilde{\mathbf{x}}) = \rho^{\lin}(\tilde{\mathbf{x}}) = \rho^{\sub}(\tilde{\mathbf{x}}) \tag{39}$$

Now, for any point $\mathbf{x} \in \mathcal{U}$ that satisfies the assumption in Equation 18, we can conclude:

$$\rho^{\sup}(\mathbf{x}) > \rho^{\lin}(\mathbf{x}) > \rho^{\sub}(\mathbf{x}) \tag{40}$$

Finally, subtracting the $\rho(\mathbf{x})$ from the surrogate model predictions, we obtain the lower bounds for all $\mathbf{x}$. Since $\rho^{\sup}(\mathbf{x}) > \rho^{\lin}(\mathbf{x}) > \rho^{\sub}(\mathbf{x})$ holds for all $\mathbf{x} \in \mathcal{U}$ based on the order of Lipschitz constant approximations and the properties of the distance metrics, we can conclude that:

$$\hat{f}_{lb}^{\sup}(\mathbf{x}) < \hat{f}_{lb}^{\lin}(\mathbf{x}) < \hat{f}_{lb}^{\sub}(\mathbf{x}). \tag{41}$$

$\square$

28