The Sample Complexities of Global Lipschitz Optimization

François Bachoc, Tommaso R Cesari, Sébastien Gerchinovitz

The Sample Complexities of Global Lipschitz Optimization

François Bachoc[1], Tommaso R. Cesari[2], and Sébastien Gerchinovitz[3]

[1]Institut de Mathématiques de Toulouse & University Paul Sabatier
[2]Toulouse School of Economics
[3]IRT Saint Exupéry & Institut de Mathématiques de Toulouse

March 9, 2021

Abstract

We study the problem of black-box optimization of a Lipschitz function f defined on a compact subset $\mathcal{X}$ of $\mathbb{R}^d$, both via algorithms that certify the accuracy of their recommendations and those that do not. We investigate their sample complexities, i.e., the number of samples needed to either reach or certify a given accuracy ε. We start by proving a tighter bound for the well-known DOO algorithm [Perevozchikov, 1990, Munos, 2011] that matches the best existing upper bounds for (more computationally challenging) non-certified algorithms. We then introduce and analyze a new certified version of DOO and prove a matching f-dependent lower bound (up to logarithmic terms). Afterwards, we show that this optimal quantity is proportional to $\int_{\mathcal{X}} \mathrm{d}\boldsymbol{x}/(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon)^d$, solving as a corollary a three-decade-old conjecture by Hansen et al. [1991]. Finally, we show how to control the sample complexity of state-of-the-art non-certified algorithms with an integral reminiscent of the Dudley-entropy integral.

Keywords: global optimization, bandit optimization.

1 Introduction

In this paper, $f\colon \mathcal{X} \to \mathbb{R}$ denotes a function defined on a compact non-empty subset $\mathcal{X}$ of $\mathbb{R}^d$. We consider the following global optimization problem: with only black-box access to f, find an ε-optimal point $\boldsymbol{x}_n^\star \in \mathcal{X}$ of f with as little evaluations of f as possible.

We make the following weak Lipschitz assumption, where the norm $\|\cdot\|$ and the constant L are known to the learner. For some of our results, we will instead assume f to be globally L-Lipschitz.

Assumption 1 (Lipschitzness around a maximizer)**.** *The function f attains its maximum at some $\boldsymbol{x}^\star \in \mathcal{X}$, and there exist a constant $L > 0$ and a norm $\|\cdot\|$ such that, for all $\boldsymbol{x} \in \mathcal{X}$,*

$$f(\boldsymbol{x}) \geq f(\boldsymbol{x}^\star) - L\,\|\boldsymbol{x}^\star - \boldsymbol{x}\| \ .$$

1.1 Setting: Black-Box Optimization, With or Without Certificates

We study the case in which f is black-box, i.e., except for some *a priori* knowledge on its smoothness, we can only access f by sequentially querying its values at a sequence $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \in [0,1]^d$ of points of our choice (see Online Protocol 1 below). At every round $i \geq 1$, the query point $\boldsymbol{x}_i$ can be chosen as a deterministic function of the values $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_{i-1})$ observed so far. At the end of round i, the learner outputs a recommendation $\boldsymbol{x}_i^\star \in \mathcal{X}$ with some other optional information. The goal is to minimize the *optimization error* (or *simple regret*): $\max_{\boldsymbol{x}\in\mathcal{X}} f(\boldsymbol{x}) - f(\boldsymbol{x}_i^\star) = f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_i^\star)$.

In all the sequel, we consider two different types of algorithms:

Online Protocol 1: Certified/non-certified algorithm

input: accuracy $\varepsilon > 0$ (*certified algorithms only*)
for $i = 1, 2, \ldots$ **do**
  pick the next query point $\boldsymbol{x}_i \in \mathcal{X}$
  observe the value $f(\boldsymbol{x}_i) \in \mathbb{R}$
  output a recommendation $\boldsymbol{x}_i^\star \in \mathcal{X}$
  output a certificate $\gamma_i \in \{0, 1\}$ (*certified algorithms only*)

- *certified algorithms* output an *accuracy certificate* $\gamma_i \in \{0, 1\}$ along with the recommendation $\boldsymbol{x}_i^\star$ at the end of each step i. By definition, certified algorithms take as input an accuracy $\varepsilon > 0$ and are such that $f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_i^\star) \leq \varepsilon$ whenever $\gamma_i = 1$, for any function f satisfying Assumption 1. In other words, the accuracy certificate γ_i can only equal 1 if the algorithm can guarantee that the recommendation $\boldsymbol{x}_i^\star$ is ε-optimal;
- *non-certified algorithms* only output the recommendation $\boldsymbol{x}_i^\star \in \mathcal{X}$ at the end of step i.

The goal of finding an ε-optimal point $\boldsymbol{x}_n^\star \in \mathcal{X}$ of f with as little evaluations of f as possible corresponds to minimizing the sample complexity. More precisely, we associate a natural notion of sample complexity to each of the two types of algorithms above. We will discuss existing bounds and algorithms in Section 1.2, and summarize our main contributions in Section 1.3.

Non-certified sample complexity. For non-certified algorithms, a natural and classical performance criterion is the minimum number of queries made before outputting ε-optimal recommendations only. More precisely, we define the *non-certified sample complexity* $\zeta(A, f, \varepsilon)$ of a non-certified algorithm A for a function f with accuracy $\varepsilon > 0$ as

$$\zeta(A, f, \varepsilon) := \inf\{n \geq 1 : \forall i \geq n,\ \boldsymbol{x}_i^\star \in \mathcal{X}_\varepsilon\} \in \{1, 2, \ldots\} \cup \{+\infty\}\,. \tag{1}$$

Certified sample complexity. For certified algorithms, it is more natural to look at a notion of sample complexity that does not depend on a oracle knowing f but instead that can be computed on the basis of the outputs of the algorithm only. We define the *certified sample complexity* $\sigma(A, f, \varepsilon)$ of a certified algorithm A for a target function f with accuracy $\varepsilon > 0$ as

$$\sigma(A, f, \varepsilon) = \inf\{i \geq 1 : \gamma_i = 1\} \in \{1, 2, \ldots\} \cup \{+\infty\}\,. \tag{2}$$

This corresponds to the first time when we can stop the algorithm while being sure to have an ε-optimal recommendation $\boldsymbol{x}_i^\star$.

1.2 Existing Bounds and Algorithms

The most naive approach to output an ε-optimal point knowing that f satisfies Assumption 1 is to perform a uniform grid search in $\mathcal{X}$ with step-size $\approx \varepsilon/L$, requiring roughly $(L/\varepsilon)^d$ evaluations of f. Though this approach is optimal in the worst case (see, e.g., Theorem 1.1.2 by Nesterov 2003 where a small Lipschitz bump function f is adversarially constructed for any given algorithm), it is clearly suboptimal for functions that take "very suboptimal" values for a "large fraction" of the domain $\mathcal{X}$. For such more benign functions, sequential algorithms can reasonably quickly rule out very suboptimal regions and then mostly explore closer-to-optimality regions. Indeed many papers have considered sequential algorithms with improved bounds for benign functions, as discussed below.

Many of these bounds rely on the notions of sets of near-optimal points and of packing numbers, which we recall first. For all $\varepsilon > 0$, the set of *ε-optimal points* of f is defined by $\mathcal{X}_\varepsilon := \big\{\boldsymbol{x} \in \mathcal{X} : f(\boldsymbol{x}) \geq f(\boldsymbol{x}^\star) - \varepsilon\big\}$. We also denote its complement (i.e., the set of *ε-suboptimal points*) by $\mathcal{X}_\varepsilon^c$ and, for all $0 \leq a < b$, the $(a, b]$-*layer* by $\mathcal{X}_{(a,b]} := \mathcal{X}_a^c \cap \mathcal{X}_b = \big\{\boldsymbol{x} \in \mathcal{X} : a < f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) \leq b\big\}$ (i.e., the set of points that are b-optimal but a-suboptimal). Since f is L-Lipschitz around $\boldsymbol{x}^\star$, every point in $\mathcal{X}$ is ε_0-optimal, with ε_0 defined by $\varepsilon_0 := L \sup_{\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}} \|\boldsymbol{x} - \boldsymbol{y}\|$. In other words, $\mathcal{X}_{\varepsilon_0} = \mathcal{X}$. For this reason, without loss of generality we will only consider values of accuracy ε smaller than or equal to ε_0.

Recall also that for any bounded set $A \subset \mathbb{R}^d$ and any real number $r > 0$, the *r-packing number* of A is the largest cardinality of an r-packing of A, that is, $\mathcal{N}(A, r) := \sup\{k \in \mathbb{N}^* : \exists \boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in A, \min_{i \neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| > r\}$ if A is nonempty, zero otherwise. Well-known and useful properties of packing (and covering) numbers are recalled in Appendix D.

Bounds on the non-certified sample complexity. Several (non worst-case) f-dependent upper bounds involving sets of ε-optimal points $\mathcal{X}_\varepsilon$ or layers $\mathcal{X}_{(a,b]}$ of f have been derived over the past decades for the non-certified sample complexity. For instance, for the DOO algorithm, Munos [2011, Theorem 1] proved a bound roughly of the form $\sum_{i=1}^{\delta^{-1}(\varepsilon)} \mathcal{N}(\mathcal{X}_{\delta(i)}, c\delta(i))$ for some constant $c > 0$ independent of f, and where the sequence $i \mapsto \delta(i)$ is decreasing. In the special case where $\delta(i) \approx \gamma^i$ with $\gamma \in (0, 1)$, and for (weakly) Lipschitz functions f such that $\mathcal{N}(\mathcal{X}_\varepsilon, c\varepsilon) \lesssim (1/\varepsilon)^{d^\star}$ with $d^\star \in [0, d]$ (we say that f has near-optimality dimension smaller than $d^\star$), this non-certified sample complexity bound implies a bound roughly of $\log(1/\varepsilon)$ if $d^\star = 0$ or $(1/\varepsilon)^{d^\star}$ if $d^\star > 0$. A similar bound was proved by Perevozchikov [1990] for a branch-and-bound algorithm close to the DOO algorithm, with an assumption on the volume of $\mathcal{X}_\varepsilon$ (instead of a packing number), or by Malherbe and Vayatis [2017] for a stochastic version of the Piyavskii-Shubert algorithm under a stronger assumption on the shape of f around its maximizer.

The matching worst-case lower bounds of Nesterov [2003, Theorem 1.1.2] when $d^\star = d$, of Horn [2006] when $d^\star = d/2$, and the lower bounds of Bubeck et al. [2011] for any $d^\star$ but in the stochastic case suggest that this bound is worst-case optimal over the class of functions with given near-optimality dimension $d^\star$. However, the bounds mentioned above have several limitations. First, as noted earlier by Kleinberg et al. [2019, remark after Theorem 4.4], bounds involving a single notion of dimension such as the near-optimality or zooming dimension might be too crude. Indeed, functions that feature different shapes at different scales (such as, e.g., different values of $d^\star$ for different ranges of ε) are better analyzed in a non-asymptotic framework through sums of packing numbers at different scales (see discussion at the end of Section 5). Second, considering packing numbers of the near-optimal sets $\mathcal{X}_{\delta(i)}$ as in the first bound mentioned above can also be quite suboptimal. For instance, for constant functions f, the sets $\mathcal{X}_{\delta(i)}$ are equal to the whole domain $\mathcal{X}$ (maximal packing number) while f is optimized by any algorithm after only a single evaluation of f.

The second limitation was overcome in general metric spaces by Kleinberg et al. [2019] (see also Kleinberg et al. 2008) by considering packing numbers of layers $\mathcal{X}_{(a,b]}$ instead of near-optimal sets $\mathcal{X}_\varepsilon$. In the special deterministic setting with compact domain $\mathcal{X} \subset \mathbb{R}^d$ considered here, Bouttier et al. [2020, Theorem 1] proved that the non-certified sample complexity of the Piyavskii-Shubert algorithm [Piyavskii, 1972, Shubert, 1972] is at most of $1 + S_{\mathrm{NC}}(f, \varepsilon)$, where[1]

$$S_{\mathrm{NC}}(f, \varepsilon) := \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k, \varepsilon_{k-1}]}, \frac{\varepsilon_k}{L}\right) , \tag{3}$$

where $m_\varepsilon := \left\lceil \log_2(\varepsilon_0/\varepsilon) \right\rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$ and, for each $k \in \{0, 1, \dots, m_\varepsilon - 1\}$, $\varepsilon_k := \varepsilon_0 2^{-k}$.

The bound (3) on the non-certified sample complexity is proved by noting that highly suboptimal regions $\mathcal{X}_{(\varepsilon_k, \varepsilon_{k-1}]}$ (with $\varepsilon_k \gg \varepsilon$) need not be explored too much thanks to Assumption 1. Any reasonable algorithm can therefore "quickly" output recommendations $\boldsymbol{x}_i^\star$ closer to ε-optimality.

Notably, the two algorithms achieving (3) or similar bounds are computationally challenging. Indeed, the zooming algorithm of Kleinberg et al. [2008, 2019] designed for general metric spaces is based on a "covering oracle" that takes as input a collection of balls and either declares it to be a covering of the input space, or outputs a point which is not covered. As for the Piyavskii-Shubert algorithm [Piyavskii, 1972, Shubert, 1972], it requires at every step n to solve an inner global Lipschitz optimization problem whose computational complexity might resemble that of the computation of a Voronoi diagram (see discussion in Bouttier et al. 2020, Section 1.1).

In Section 2.1 we show that the more computationally tractable DOO algorithm matches the same bound (3). For better interpretability, we also show in Section 5 that this bound is proportional to an integral reminiscent of the Dudley-entropy integral.

[1] The bound in Bouttier et al. [2020, Theorem 1] with $\alpha = 0$ differs a little in the smallest level considered, which can be as small as $\varepsilon/2$, instead of the more natural level ε. Their bound can however be straightforwardly improved to get ε instead.

Bounds on the certified sample complexity. This notion of sample complexity seems to have been less studied in the past. In dimension $d = 1$, several authors derived upper bounds for a certified version of the Piyavskii-Shubert algorithm. In particular, Hansen et al. [1991] proved that its certified sample complexity for globally L-Lipschitz functions $f\colon [0,1] \to \mathbb{R}$ is at most proportional to the integral $\int_0^1 \big(f(x^\star) - f(x) + \varepsilon\big)^{-1} \,\mathrm{d}x$, and left the question of extending the results to arbitrary dimensions open. Recently, Bouttier et al. [2020, Theorem 2] proved a bound valid for any dimension $d \geq 1$ roughly of this form:

$$S_{\mathrm{C}}(f,\varepsilon) := \mathcal{N}\left(\mathcal{X}_\varepsilon, \frac{\varepsilon}{L}\right) + \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\varepsilon_{k-1}]}, \frac{\varepsilon_k}{L}\right) , \tag{4}$$

where, again, $m_\varepsilon := \lceil \log_2(\varepsilon_0/\varepsilon) \rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$ and, for each $k \in \{0, 1, \dots, m_\varepsilon - 1\}$, $\varepsilon_k := \varepsilon_0 2^{-k}$.

We show in Section 2.2 how to obtain (4) with the more computationally tractable DOO algorithm, and prove a matching f-dependent lower bound (up to log factors) in Section 3. In Section 4 we also show that the bound (4) is actually proportional to $\int_{\mathcal{X}} \big(f(x^\star) - f(x) + \varepsilon\big)^{-d} \,\mathrm{d}x$ for globally L-Lipschitz functions $f : \mathcal{X} \to \mathbb{R}$ under a mild assumption on $\mathcal{X}$, solving the question left open by Hansen et al. [1991] three decades ago.

Though the two bounds (3) and (4) look very similar, they differ in the first term being either 1 or $\mathcal{N}\left(\mathcal{X}_\varepsilon, \frac{\varepsilon}{L}\right)$. This difference is not negligible in general and can be explained with a simple example. Indeed, consider a constant function $c\colon [0,1]^d \to \mathbb{R}$. Any non-certified algorithm has sample complexity 1 for c at any scale ε, because all points in the domain are maxima. However, the only way to *certify* that the output is ε-optimal is essentially to perform a grid-search of $[0,1]^d$ with step-size roughly ε/L, so as to be sure there is no hidden bump of height more than ε. This is reflected in the term $\mathcal{N}\left(\mathcal{X}_\varepsilon, \frac{\varepsilon}{L}\right)$, which is of order $(L/\varepsilon)^d$ for constant functions. At a high level, the more "constant" a function is, the easier it is to recommend an ε-optimal point, but the harder it is to certify that such recommendation is actually a good recommendation.

1.3 Outline and Main Contributions

In Section 2, we refine the analysis of the DOO algorithm, showing that its certified (resp., non-certified) sample complexity can be upper bounded (up to constants) by $S_{\mathrm{C}}(f,\varepsilon)$ (resp., $S_{\mathrm{NC}}(f,\varepsilon)$). In Section 3, we prove an f-dependent lower bound on the sample complexity of all certified algorithms, which matches $S_{\mathrm{C}}(f,\varepsilon)$ up to logarithmic terms. In Section 4, we show for globally Lipschitz functions $f\colon \mathcal{X} \to \mathbb{R}$ (with a mild condition on $\mathcal{X}$) that $S_{\mathrm{C}}(f,\varepsilon)$ is proportional to the integral $\int_{\mathcal{X}} \mathrm{d}\boldsymbol{x}/(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon)^d$, which thus characterizes the optimal certified sample complexity of global Lipschitz optimization. In Section 5, we also show that $S_{\mathrm{NC}}(f,\varepsilon)$ is proportional to the Dudley-type integral over accuracies $\int_\varepsilon^\infty \mathcal{N}\left(\mathcal{X}_{(\max(\varepsilon,\, u/4),\, u]},\, {}^{u}\!/_{L}\right)/u \,\mathrm{d}u$. Some proofs and useful lemmas are deferred to the appendix, together with a comparison of packing-based and volume-based bounds (Appendix C).

What we do not cover There are some interesting directions that would be worth investigating in the future but we did not cover in this paper, such as noisy observations (see, e.g, Bubeck et al. 2011, Kleinberg et al. 2019) or adaptivity to smoothness (e.g., Munos 2011, Bartlett et al. 2019; we consider L-Lipschitz functions f with L known, although our lower bound suggests that no adaptivity could be possible for certified algorithms). Finally, even if the results of Section 2 could be easily extended to general pseudo-metric spaces as in Munos [2014] and related works, our other results are finite-dimensional and exploit the normed space structure.

1.4 Additional Notation

We denote the set of positive integers $\{1, 2, \dots\}$ by $\mathbb{N}^*$ and let $\mathbb{N} := \mathbb{N}^* \cup \{0\}$. For all $n \in \mathbb{N}^*$, we denote by $[n]$ the set of the first n integers $\{1, \dots, n\}$. We denote by $A + B$ the Minkowski sum of two sets A, B and for any set A and all $\lambda \in \mathbb{R}$, we let $\lambda A := \{\lambda \boldsymbol{a} : \boldsymbol{a} \in A\}$.

We denote the Lebesgue measure of a Lebesgue-measurable set A by $\mathrm{vol}(A)$ and we simply refer to it as *volume*. For all $\rho > 0$ and $\boldsymbol{x} \in \mathbb{R}^d$, we denote by $B_\rho(\boldsymbol{x})$ the closed ball with radius ρ centered at $\boldsymbol{x}$. We also write B_ρ for the ball with radius ρ centered at the origin and denote by v_ρ its volume.

2 A Tighter Analysis of the DOO Algorithm: Two Sample Complexity Bounds

In this section, we show that the DOO algorithm of Perevozchikov [1990], Munos [2011] and a new certified version of it match the two f-dependent bounds (3) and (4) previously derived for computationally more challenging algorithms. In particular, our analysis of the non-certified sample complexity of DOO improves over that of Munos [2011]. A nearly matching lower bound on the certified sample complexity of all certified algorithms will be proved in Section 3.

The definition of the DOO algorithm and the associated assumptions are recalled (and slightly adjusted) in Appendix E.1, together with its new certified version.

2.1 Upper Bound on the Non-Certified Sample Complexity

The next theorem shows that the quantity $S_{\mathrm{NC}}(f,\varepsilon)$ defined in (3) is a tight upper bound on the non-certified sample complexity of the DOO algorithm. This improves over the bound of Munos [2011, Theorem 1] since our bound involves a sum of disjoint layers $\mathcal{X}_{(\varepsilon_k,\varepsilon_{k-1}]}$ whose values are all above ε, while the bound of Munos [2011] involves a sum over overlapping layers of the form $\mathcal{X}_{[0,u)}$ for all values $0 < u \leq \varepsilon_0$. See Section 1.2 and Remark 17 in Appendix E.2 for further discussion. We also note that the result could be extended to more general layers than those based on $(\varepsilon_k)_{k\in\{0,\dots,m_\varepsilon\}}$, as detailed in Remark 18.

Recall that $\zeta(A,f,\varepsilon)$ denotes the non-certified sample complexity of an algorithm A (see (1)), and that $m_\varepsilon := \lceil \log_2(\varepsilon_0/\varepsilon) \rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$ and, for each $k \in \{0,1,\dots,m_\varepsilon - 1\}$, $\varepsilon_k := \varepsilon_0 2^{-k}$.

Theorem 1. *Assume that f satisfies Assumption 1, and that Assumptions 4 and 5 in Appendix E.1 hold. Then, the non-certified DOO algorithm (Algorithm 3 in Appendix E.1) has a non-certified sample complexity bounded as follows: for any $\varepsilon \in (0,\varepsilon_0]$,*

$$\zeta(\textit{non-certified DOO}, f, \varepsilon) \leq 1 + C \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\varepsilon_{k-1}]}, \frac{\varepsilon_k}{L}\right) ,$$

where $C = K\big(\mathbf{1}_{\nu/R\geq 1} + \mathbf{1}_{\nu/R<1}(4R/\nu)^d\big)$.

Note that the above bound equals $1 + C \cdot S_{\mathrm{NC}}(f,\varepsilon)$, where $S_{\mathrm{NC}}(f,\varepsilon)$ was defined in (3) in Section 1.2. It therefore matches the bound of Bouttier et al. [2020, Theorem 1] on the Piyavskii-Shubert algorithm up to constant factors, but for the more computionally tractable DOO algorithm.

As discussed in Appendix E.1, when $\mathcal{X} = [0,1]^d$ and $\|\cdot\|$ is the sup norm, we have $K = 2^d$, $R = 1, \delta = 1/2$ and $\nu = 1/2$. Hence, in this case, the multiplicative constant C above equals 16^d.

The proof is postponed to Appendix E.2 and shares common arguments with the proofs of Perevozchikov [1990], Munos [2011]. As noted in Remark 17, some steps are however tighter, leading to a tighter bound. The key change is to partition the values of f instead of partitioning the domain $\mathcal{X}$ at any depth h in the tree (see Munos 2011) when counting the representatives selected at all levels. The idea of using layers $\mathcal{X}_{(\varepsilon_i,\ \varepsilon_{i-1}]}$ was already present in Kleinberg et al. [2008, 2019] and Bouttier et al. [2020] but for more computionally challenging algorithms (see discussion in Section 1.2).

2.2 Upper Bound on the Certified Sample Complexity

We now analyze the certified version of the DOO algorithm that we design in Appendix E.1, and show that its certified sample complexity is at most proportional to $S_{\mathrm{C}}(f,\varepsilon)$ defined in (4).

Recall that $\sigma(A,f,\varepsilon)$ denotes the certified sample complexity of an algorithm A (see (2)), and that $m_\varepsilon := \lceil \log_2(\varepsilon_0/\varepsilon) \rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$ and, for each $k \in \{0,1,\dots,m_\varepsilon - 1\}$, $\varepsilon_k := \varepsilon_0 2^{-k}$.

Theorem 2. *Assume that Assumptions 4 and 5 in Appendix E.1 hold. Then the certified DOO algorithm (Algorithm 3 in Appendix E.1) is indeed a certified algorithm. Furthermore, for any f satisfying Assumption 1 and for any $\varepsilon \in (0,\varepsilon_0]$, its certified sample complexity is bounded as*

$$\sigma(\textit{certified DOO}, f, \varepsilon) \leq C\left(\mathcal{N}\left(\mathcal{X}_\varepsilon, \frac{\varepsilon}{L}\right) + \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\varepsilon_{k-1}]}, \frac{\varepsilon_k}{L}\right)\right) ,$$

where $C = 1 + K\big(\mathbf{1}_{\nu/R\geq 1} + \mathbf{1}_{\nu/R<1}(4R/\nu)^d\big)$.

Note that the above bound is proportional to the quantity $S_{\mathrm{C}}(f,\varepsilon)$ defined in (4) in Section 1.2. It therefore matches the bound of Bouttier et al. [2020, Theorem 2] on the certified version of the Piyavskii-Shubert algorithm up to constant factors, but for the more computionally tractable DOO algorithm.

Similarly to Theorem 1, the above result could be easily extended to more general layers than those based on $(\varepsilon_k)_{0\leq k\leq m_\varepsilon}$ (see Remark 18 for further details).

The proof is postponed to Appendix E.3. It follows approximately the same lines as that of Theorem 1, except that considering the stopping time $\sigma(\text{certified DOO}, f, \varepsilon)$ introduces the additional term $\mathcal{N}\left(\mathcal{X}_\varepsilon, \frac{\varepsilon}{L}\right)$, which cannot be avoided as discussed in Section 1.2 and proved formally in Section 3. Similar arguments were used by Bouttier et al. [2020], which we adapt here to handle the discretization inherent to the DOO algorithm (this discretization makes the analysis slightly less natural but the algorithm more computationally tractable).

3 Lower Bounds for the Certified Sample Complexity

In this section, we will focus on f-dependent lower bounds on the certified sample complexity of certified algorithms applied to globally Lipschitz functions. Formally, we assume the following.

Assumption 2 (Global Lipschitzness)**.** *We say that* $L_0 := \sup_{\boldsymbol{u},\boldsymbol{v}\in\mathcal{X},\boldsymbol{u}\neq\boldsymbol{v}} \big|f(\boldsymbol{u}) - f(\boldsymbol{v})\big| / \|\boldsymbol{u}-\boldsymbol{v}\|$*, where* $\|\cdot\|$ *is a norm, is the* Lipschitz constant *of* f *and that* f *is* globally L-Lipschitz *if* $L \geq L_0$.

It is immediate to see that Assumption 1 is a relaxation of Assumption 2. Similarly to Section 1.1, we say that A is a *certified algorithm for L-Lipschitz functions* if for any globally L-Lipschitz function f, accuracy $\varepsilon \in (0, \varepsilon_0]$, and time step $i \in \mathbb{N}^*$, the accuracy certificate γ_i of A at time i when applied to f with accuracy ε is equal to 1 only if its recommendation $\boldsymbol{x}_i^\star$ satisfies $\boldsymbol{x}_i^\star \in \mathcal{X}_\varepsilon$.

The main result of this section (Theorem 3) shows that when a certified algorithm for L-Lipschitz functions is applied to a function f that is globally L'-Lipschitz, with L' bounded away below L, then its certified sample complexity is (up to a log factor) at least $S_{\mathrm{C}}(f,\varepsilon)$. This is the quantity that we presented in Sections 1 and 2, which upper bounds the certified sample complexity of algorithms such as DOO and Piyavskii-Shubert. Putting these upper and lower bounds together proves that the *optimal* f-dependent certified sample complexity (of certified algorithms for L-Lipschitz functions) is of order $S_{\mathrm{C}}(f,\varepsilon)$, up to a log factor, at least for globally L'-Lipschitz functions with L' bounded away below L. The boundary case where $L' = L$ is discussed later.

Theorem 3. *Let* $0 < L' < L$*,* $K := 16L/(L-L')$*, and* $c := (1/8K)^d/4$*. Then, the certified sample complexity of any certified algorithm A for L-Lipschitz functions satisfies, for any globally L'-Lipschitz function f and all* $\varepsilon \in (0, \varepsilon_0]$*,*

$$\sigma(A,f,\varepsilon) > \frac{c}{1+m_\varepsilon} S_{\mathrm{C}}(f,\varepsilon)\ . \tag{5}$$

Before proving Theorem 3, we introduce the insightful notions of *worst-case error* and *worst-case sample complexity*. For an algorithm A and a function f, we denote the point queried by A when applied to f at time n by $\boldsymbol{x}_n(A,f)$, the recommendation by $\boldsymbol{x}_n^\star(A,f)$, and the accuracy certificate by $\gamma_n(A,f)$. For any $n \in \mathbb{N}^\star$ we define the *worst-case error* of A when applied to f at time n as

$$E_L(A,f,n) := \sup\Big\{\max(g) - g\big(\boldsymbol{x}_n^\star(A,f)\big) : g \text{ is globally } L\text{-Lipschitz and } g(\boldsymbol{x}) = f(\boldsymbol{x}) \text{ for all } \boldsymbol{x} \in \big\{\boldsymbol{x}_1(A,f),\dots,\boldsymbol{x}_n(A,f)\big\}\Big\}. \tag{6}$$

Note that if f is globally L-Lipschitz, the optimization error of A at time n is no larger than (but possibly equal to) $E_L(A,f,n)$. We then define the *worst-case minimax error* for f at step n as $\inf_A E_L(A,f,n)$, where the infimum is over all algorithms (not only certified ones). A classic performance measure in global optimization is the minimax error $\inf_A \sup_g\big(\max(g) - g\big(\boldsymbol{x}_n^\star(A,g)\big)\big)$, where the supremum is over all gobally L-Lipschitz functions and the infimum is over all algorithms (e.g., see Nesterov 2003). Compared to ours, this is a more pessimistic notion, that does not depend on f. Based on our worst-case minimax error, we define the the *worst-case sample complexity* of a globally L-Lipschitz function f with accuracy $\varepsilon \in (0, \varepsilon_0]$, as

$$\tau(f,\varepsilon) := \min\big\{n \in \mathbb{N}^* : \inf_A E_L(A,f,n) \leq \varepsilon\big\}\ ,$$

where the infimum is over all algorithms. It is immediate to prove that $\tau(f,\varepsilon)$ is finite, by considering an algorithm that queries a dense sequence of points (independently of the observed function values) and outputs as a recommendation the maximizer of the observed values.

Crucially, the worst-case sample complexity lower bounds the certified sample complexity.

Lemma 4. *For any certified algorithm A for L-Lipschitz functions, any globally L-Lipschitz function f, and all $\varepsilon \in (0, \varepsilon_0]$, we have $\sigma(A,f,\varepsilon) \geq \tau(f,\varepsilon)$.*

Proof. Let $N = \sigma(A,f,\varepsilon)$. Then $\gamma_N(A,f) = 1$. Assume that $N < \tau(f,\varepsilon)$. Then we have $E_L(A,f,N) \geq \inf_{A'} E_L(A',f,N) > \varepsilon$ by definition of $\tau(f,\varepsilon)$. This means that there exists a globally L-Lipschitz function g, coinciding with f on $\boldsymbol{x}_1(A,f),\dots,\boldsymbol{x}_N(A,f)$ and such that $\max(g) - \boldsymbol{x}_N^\star(A,g) > \varepsilon$. Then by definition of the certificate, $\gamma_N(A,g) = 0$. But $\gamma_N(A,f) = \gamma_N(A,g)$, which yields a contradiction and concludes the proof. □

We can now present a full proof of Theorem 3, the main result of this section. The high-level idea is the following. For a given certified algorithm A (for L-Lipschitz functions) and a (globally L'-Lipschitz) f, we create an adversarial function by adding a perturbation ($\pm g_{\tilde\varepsilon}$ in the proof) to f. To show that the resulting function is globally L-Lipschitz, we sum the Lipschitz constants of f and $\pm g_{\tilde\varepsilon}$. This is why we require the Lipschitz constant L' of f to be bounded away below L.

Proof of Theorem 3. Because of Lemma 4, it is sufficient to show that $\tau(f,\varepsilon) > cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$. If $S_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon) < 3(8K)^d$, then $\tau(f,\varepsilon) \geq 1 > 3/4 > cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$. Consider then from now on that $S_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon) \geq 3(8K)^d$.

We will now upper bound $S_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$ with (an upper bound of) the largest summand in (4). Let $\tilde\varepsilon$ be the scale achieving the maximum in (4), that is

$$\tilde\varepsilon = \begin{cases} \varepsilon\,, & \text{if } \mathcal{N}\left(\mathcal{X}_\varepsilon,\ \frac{\varepsilon}{L}\right) \geq \max_{i\in\{1,\dots,m_\varepsilon\}} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_i,\ \varepsilon_{i-1}]},\ \frac{\varepsilon_i}{L}\right)\,, \\ \varepsilon_{i^\star-1}\,, & \text{otherwise, where } i^\star \in \operatorname{argmax}_{i\in\{1,\dots,m_\varepsilon\}} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_i,\ \varepsilon_{i-1}]},\ \frac{\varepsilon_i}{L}\right)\,. \end{cases}$$

Since $\mathcal{N}\left(\mathcal{X}_\varepsilon,\ \varepsilon/L\right) \leq \mathcal{N}\left(\mathcal{X}_\varepsilon,\ \varepsilon/2L\right)$ and $\mathcal{N}\left(\mathcal{X}_{(\varepsilon_i,\ \varepsilon_{i-1}]},\ \varepsilon_i/L\right) \leq \mathcal{N}\left(\mathcal{X}_{\varepsilon_{i-1}},\ \varepsilon_{i-1}/2L\right)$, we then have $S_{\mathrm{C}}(f,\varepsilon) \leq (m_\varepsilon+1)\mathcal{N}\left(\mathcal{X}_{\tilde\varepsilon,\ \tilde\varepsilon/2L}\right)$. Let now $n \leq cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$. We then have $n \leq c\mathcal{N}\left(\mathcal{X}_{\tilde\varepsilon,\ \tilde\varepsilon/2L}\right)$. From Lemma 14,

$$\mathcal{N}\left(\mathcal{X}_{\tilde\varepsilon},\ \frac{K\tilde\varepsilon}{L}\right) \geq \left(\frac{1}{8K}\right)^d \mathcal{N}\left(\mathcal{X}_{\tilde\varepsilon},\ \frac{\tilde\varepsilon}{2L}\right) \geq \left(\frac{1}{8K}\right)^d \frac{S_{\mathrm{C}}(f,\varepsilon)}{m_\varepsilon+1} \geq 3\,,$$

as considered earlier. Then we have $n \leq c(8K)^d\mathcal{N}\left(\mathcal{X}_{\tilde\varepsilon,\ K\tilde\varepsilon/L}\right)$. Since $c(8K)^d = 1/4$, we thus obtain $n \leq \mathcal{N}\left(\mathcal{X}_{\tilde\varepsilon,\ K\tilde\varepsilon/L}\right) - 2$.

Consider a certified algorithm A for L-Lipschitz functions. Fix a $K\tilde\varepsilon/L$ packing $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ of $\mathcal{X}_{\tilde\varepsilon}$ with cardinality $N = \mathcal{N}\left(\mathcal{X}_{\tilde\varepsilon},\ K\tilde\varepsilon/L\right)$. Then the open balls of centers $\boldsymbol{x}_1,\dots,\boldsymbol{x}_N$ and radius $K\tilde\varepsilon/2L$ are disjoint and two of them, with centers, say, $\tilde{\boldsymbol{x}}_1$ and $\tilde{\boldsymbol{x}}_2$, do not contain any of the $\boldsymbol{x}_1(A,f),\dots,\boldsymbol{x}_n(A,f)$. Let, for $\boldsymbol{x} \in \mathcal{X}$,

$$g_{\tilde\varepsilon}(\boldsymbol{x}) := \left(8\tilde\varepsilon - \tfrac{16L}{K}\left\|\boldsymbol{x} - \tilde{\boldsymbol{x}}_1\right\|\right)\mathbb{I}\left(\boldsymbol{x} \in \mathcal{X} \cap B_{K\tilde\varepsilon/2L}(\tilde{\boldsymbol{x}}_1)\right).$$

Then $g_{\tilde\varepsilon}(x)$ is globally $16L/K = L - L'$ Lipschitz. Hence $f + g_{\tilde\varepsilon}$ and $f - g_{\tilde\varepsilon}$ are L-Lipschitz. Observe that f, $f + g_{\tilde\varepsilon}$ and $f - g_{\tilde\varepsilon}$ coincide on $\boldsymbol{x}_1(A,f),\dots,\boldsymbol{x}_n(A,f)$. As a consequence, A has the same recommendation for them: $\boldsymbol{x}_n^\star(A,f) = \boldsymbol{x}_n^\star(A,f+g_{\tilde\varepsilon}) = \boldsymbol{x}_n^\star(A,f-g_{\tilde\varepsilon})$.

Consider first the case $\boldsymbol{x}_n^\star(A,f) \in B(\tilde{\boldsymbol{x}}_1, K\tilde\varepsilon/4L)$. Then, we have, by definition of $g_{\tilde\varepsilon}$ and the fact that $\tilde{\boldsymbol{x}}_2 \in \mathcal{X}_{\tilde\varepsilon}$, $f(\tilde{\boldsymbol{x}}_2) - g_{\tilde\varepsilon}(\tilde{\boldsymbol{x}}_2) - f(\boldsymbol{x}_n^\star(A,f)) + g_{\tilde\varepsilon}(\boldsymbol{x}_n^\star(A,f)) \geq -\tilde\varepsilon + 8\tilde\varepsilon - \frac{16L}{K}\frac{K\tilde\varepsilon}{4L} = 3\tilde\varepsilon$.

Now consider the case $\boldsymbol{x}_n^\star(A,f) \notin B(\tilde{\boldsymbol{x}}_1, K\tilde\varepsilon/4L)$. Then, we have, by definition of $g_{\tilde\varepsilon}$ and the fact that $\tilde{\boldsymbol{x}}_1 \in \mathcal{X}_{\tilde\varepsilon}$, $f(\tilde{\boldsymbol{x}}_1) + g_{\tilde\varepsilon}(\tilde{\boldsymbol{x}}_1) - f(\boldsymbol{x}_n^\star(A,f)) - g_{\tilde\varepsilon}(\boldsymbol{x}_n^\star(A,f)) \geq -\tilde\varepsilon + 8\tilde\varepsilon - 8\tilde\varepsilon + \frac{16L}{K}\frac{K\tilde\varepsilon}{4L} = 3\tilde\varepsilon$.

Finally, in both cases $E_L(A,f,n) \geq 3\tilde\varepsilon > \varepsilon$. Hence $\inf_A E_L(A,f,n) > \varepsilon$. Since this has been shown for any $n \leq cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$ we thus have $\tau(f,\varepsilon) > cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$. □

We conclude the section by discussing the limit case $L' \to L$, in which the constant in Theorem 3 diverges. The next result shows that in dimension $d = 1$, the limit case $L = L'$ does not hinder the validity of the result. The proof is deferred to Appendix A.1.

Proposition 5. *If $d = 1$, let $K = 8$ and $c = 1/96K$. Then, the certified sample complexity of any certified algorithm A for L-Lipschitz functions satisfies, for any globally L-Lipschitz function f and all $\varepsilon \in (0, \varepsilon_0]$, $\sigma(A, f, \varepsilon) > (c/1+m_\varepsilon) S_{\mathrm{C}}(f, \varepsilon)$.*

The final result of this section shows that, in higher dimension $d \geq 2$, the improvement of Proposition 5 when L' is close to L is not possible in general. The proof is deferred to Appendix A.2.

Proposition 6. *Let $d \geq 2$, $\mathcal{X} := B_1$, and $\|\cdot\|$ be a norm. The certified Piyavskii-Shubert algorithm (for L-Lipschitz functions)[2] with initial guess $\boldsymbol{x}_1 := \mathbf{0}$ is a certified algorithm (for L-Lipschitz functions) satisfying, for the globally L-Lipschitz function $f := L\|\cdot\|$ and any $\varepsilon \in (0, \varepsilon_0)$, σ(Piyavskii-Shubert, $f, \varepsilon) = 2 \ll c/\varepsilon^{d-1} \leq S_{\mathrm{C}}(f, \varepsilon)$, where $c > 0$ is a constant independent of ε.*

4 An Integral Characterization of the Optimal Certified Sample Complexity S_{C}

In dimension $d = 1$, an elegant bound on the certified sample complexity was derived by Hansen et al. [1991] for a certified version of the Piyavskii-Shubert algorithm.[2] They proved that if f is globally Lipschitz, for any accuracy $\varepsilon \in (0, \varepsilon_0]$, the smallest number of time steps $\sigma := \sigma$(certified Piyavskii-Shubert, f, ε) before outputting $\gamma_\sigma = 1$ is at most proportional to $\int_0^1 \big(f(x^\star) - f(x) + \varepsilon\big)^{-1} \,\mathrm{d}x$. As pointed out in Bouttier et al. [2020], given that σ for this algorithm can be upper bounded with S_{C}, this suggests a relationship between the two quantities. However, Hansen et al. [1991] rely heavily on the one-dimensional setting and the specific form of the Piyavskii-Shubert algorithm, claiming it that even the simpler task of "*Extending the results of this paper to the multivariate case appears to be difficult*". In this section, we show an equivalence between S_{C} and this type of integral bound in any dimension d. As a corollary, this solves the long-standing problem raised by Hansen et al. [1991] three decades ago.

To tame the wild spectrum of shapes that compact subsets may have, we will assume that $\mathcal{X}$ satisfies the following additional assumption. At a high level, it says that a constant fraction of each (sufficiently small) ball centered at a point in $\mathcal{X}$ is included in $\mathcal{X}$. This removes sets containing isolated points or "peaked" corners, but includes most domains that are typically used, such as non-degenerate polytopes, ellipsoids, finite unions of them, etc. This natural assumption has already appeared in the past (e.g., Hu et al. 2020) and is weaker than another classical assumption in the statistics literature (the rolling ball assumption, Cuevas et al. 2012, Walther 1997).

Assumption 3. *There exist two constants $r_0 > 0, \gamma \in (0, 1]$ such that, for any $\boldsymbol{x} \in \mathcal{X}$ and all $r \in (0, r_0)$, $\mathrm{vol}\big(B_r(\boldsymbol{x}) \cap \mathcal{X}\big) \geq \gamma v_r$.*

We can now state the main result of this section. Its proof relies on some additional technical results that are deferred to the appendix.

Theorem 7. *If f is globally L-Lipschitz[3] and $\mathcal{X}$ satisfies Assumption 3 with $r_0 > \varepsilon_0/2L, \gamma \in (0, 1]$, then there exist $c, C > 0$ (e.g., $c := 1/v_{1/L}$ and $C := 1/(\gamma v_{1/128L})$) such that, for all $\varepsilon \in (0, \varepsilon_0]$,*

$$c \int_{\mathcal{X}} \frac{\mathrm{d}\boldsymbol{x}}{\big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon\big)^d} \leq S_{\mathrm{C}}(f, \varepsilon) \leq C \int_{\mathcal{X}} \frac{\mathrm{d}\boldsymbol{x}}{\big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon\big)^d} \,.$$

In light of Theorem 3, this also shows that the integral bound characterizes (up to a log factor) the optimal certified sample complexity for globally Lipschitz functions in any dimension d, outside of boundary cases.

Corollary 8. *Assume that Assumption 3 holds with $r_0 > \varepsilon_0/2L$. Let $0 < L' < L$. Then, there exist two constants $c, C > 0$ such that the optimal certified sample complexity of any certified algorithm A for L-Lipschitz functions satisfies, for any globally L'-Lipschitz function f and all $\varepsilon \in (0, \varepsilon_0]$,*

$$\frac{c}{1 + m_\varepsilon} \int_{\mathcal{X}} \frac{\mathrm{d}\boldsymbol{x}}{\big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon\big)^d} \leq \inf_A \sigma(A, f, \varepsilon) \leq C \int_{\mathcal{X}} \frac{\mathrm{d}\boldsymbol{x}}{\big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon\big)^d} \,.$$

[2] See Appendix A for more details on the certified Piyavskii-Shubert algorithm.

[3] We actually prove a stronger result. The first inequality holds more generally for any f that is L-Lipschitz around a maximizer (Assumption 1) and Lebesgue-measurable, and does not require $\mathcal{X}$ to satisfy Assumption 3.

The previous result follows directly from Theorems 2, 3 and 7. We now prove Theorem 7.

Proof of Theorem 7. Fix any $\varepsilon \in (0, \varepsilon_0]$ and recall that $m_\varepsilon := \lceil \log_2(\varepsilon_0/\varepsilon) \rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$, and for all $k \leq m_\varepsilon - 1$, $\varepsilon_k := \varepsilon_0 2^{-k}$. Partition the domain of integration $\mathcal{X}$ into the following $m_\varepsilon + 1$ sets: the set of ε-optimal points $\mathcal{X}_\varepsilon$ and the m_ε layers $\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]}$, for $k \in [m_\varepsilon]$. We begin by proving the first inequality:

$$\begin{aligned}
\int_{\mathcal{X}} \frac{\mathrm{d}\boldsymbol{x}}{\big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon\big)^d} &\leq \frac{\mathrm{vol}(\mathcal{X}_\varepsilon)}{\varepsilon^d} + \sum_{k=1}^{m_\varepsilon} \frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]}\big)}{(\varepsilon_k + \varepsilon)^d} \\
&\leq \frac{\mathcal{M}\big(\mathcal{X}_\varepsilon,\, \frac{\varepsilon}{L}\big) \cdot v_1\big(\frac{\varepsilon}{L}\big)^d}{\varepsilon^d} + \sum_{k=1}^{m_\varepsilon} \frac{\mathcal{M}\big(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]},\, \frac{\varepsilon_k}{L}\big) \cdot v_1\big(\frac{\varepsilon_k}{L}\big)^d}{\varepsilon_k^d} \\
&\leq \frac{v_1}{L^d} \left(\mathcal{N}\left(\mathcal{X}_\varepsilon,\, \frac{\varepsilon}{L}\right) + \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]},\, \frac{\varepsilon_k}{L}\right) \right) ,
\end{aligned}$$

where the first inequality follows by lower bounding $f(\boldsymbol{x}^\star) - f$ with its infimum on the partition, the second one by dropping $\varepsilon > 0$ from the second denominator and upper bounding the volume of a set with the volume of the balls of a smallest ε_k/L-cover, and the last one by the fact that covering numbers are always smaller than packing numbers (14). This proves the first part of the theorem.

For the second one, note that

$$\begin{aligned}
\int_{\mathcal{X}} \frac{\mathrm{d}\boldsymbol{x}}{\big(f(\boldsymbol{x}^\star) - f(\boldsymbol{x}) + \varepsilon\big)^d} &\geq \frac{\mathrm{vol}(\mathcal{X}_\varepsilon)}{(\varepsilon + \varepsilon)^d} + \sum_{k=1}^{m_\varepsilon} \frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]}\big)}{(\varepsilon_{k-1} + \varepsilon)^d} \\
&\geq \frac{1}{2^d} \frac{\mathrm{vol}(\mathcal{X}_\varepsilon)}{\varepsilon^d} + \frac{1}{4^d} \sum_{k=1}^{m_\varepsilon} \frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]}\big)}{\varepsilon_k^d} \geq \frac{1}{32^d} \left(\frac{\mathrm{vol}(\mathcal{X}_{2\varepsilon})}{\varepsilon^d} + \sum_{k=1}^{m_\varepsilon} \frac{\mathrm{vol}\left(\mathcal{X}_{\left(\frac{1}{2}\varepsilon_k,\, 2\varepsilon_{k-1}\right]}\right)}{\varepsilon_{k-1}^d} \right) \\
&\geq \frac{\mathcal{N}\left(\mathcal{X}_\varepsilon,\, \frac{\varepsilon}{L}\right) \mathrm{vol}\left(\frac{\varepsilon}{2L} B_1\right)}{(32\,\varepsilon)^d/\gamma} + \sum_{k=1}^{m_\varepsilon} \frac{\mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]},\, \frac{\varepsilon_k}{L}\right) \mathrm{vol}\left(\frac{\varepsilon_k}{2L} B_1\right)}{(32\,\varepsilon_{k-1})^d/\gamma} \\
&\geq \gamma v_{1/64L} \mathcal{N}\left(\mathcal{X}_\varepsilon,\, \frac{\varepsilon}{L}\right) + \gamma v_{1/128L} \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]},\, \frac{\varepsilon_k}{L}\right) ,
\end{aligned}$$

where the first inequality follows by upper bounding $f(\boldsymbol{x}^\star) - f$ with its supremum on the partition, the second one by $\varepsilon \leq \varepsilon_{k-1}$ (for all $k \in [m_\varepsilon + 1]$) and $\varepsilon_{k-1} \leq 2\varepsilon_k$ (for all $k \in [m_\varepsilon]$), the third one by Lemma 15,[4] and the fourth one by the elementary inclusions $\mathcal{X}_{2\varepsilon} \supseteq \mathcal{X}_{\frac{3}{2}\varepsilon}$ and $\mathcal{X}_{\left(\frac{1}{2}\varepsilon_k,\, 2\varepsilon_{k-1}\right]} \supseteq \mathcal{X}_{\left(\frac{1}{2}\varepsilon_k,\, \frac{3}{2}\varepsilon_{k-1}\right]}$ (for all $k \in [m_\varepsilon]$) followed by Proposition 11. □

5 A Dudley-Integral Characterization of the Non-Certified Sample Complexity Bound S_{NC}

As we mentioned earlier, the non-certified sample complexity of some algorithms can be upper bounded with a sum of packing numbers (3) (e.g., for the DOO and Piyavskii-Shubert algorithms). Notably, these types of bounds appear in fields beyond optimization (e.g., see Bachoc et al. 2020, Theorem 2). In this section, we show how this quantity can be controlled by a more elegant Dudley-type entropy integral bound. The idea of the reduction is leveraging the usual sum-integral approximation carried out, e.g., when deriving the Dudley entropy integral bound (see, e.g., Dudley 1967, Boucheron et al. 2013), with the additional care that here the integrand is not necessarily monotone.

[4] In addition to Lemma 15, in Appendix D, we prove a more general result on controlling the sum of the volumes of a family of overlapping sets covering $\mathcal{X}$ with the sum of volumes over a partition of $\mathcal{X}$ (Proposition 16). As we do here, this can be translated to packing numbers.

Theorem 9. *If f is L-Lipschitz around a maximizer and $(x, y, z) \mapsto \mathcal{N}\big(\mathcal{X}_{(x,y]}, z\big)$ is Lebesgue-measurable, then there exist $c, C > 0$ (e.g., $c = 1/5$ and $C := 2 \cdot 16^d$) such that, for all $\varepsilon \in (0, \varepsilon_0]$,*

$$c \int_\varepsilon^\infty \frac{\mathcal{N}\left(\mathcal{X}_{(\max(\varepsilon,\, u/4),\, u]},\ {}^{u}/_{L}\right)}{u} \,\mathrm{d}u \leq S_{\mathrm{NC}}(f, \varepsilon) \leq C \int_\varepsilon^\infty \frac{\mathcal{N}\left(\mathcal{X}_{(\max(\varepsilon,\, u/4),\, u]},\ {}^{u}/_{L}\right)}{u} \,\mathrm{d}u \ .$$

Proof. Fix any $\varepsilon \in (0, \varepsilon_0]$ and recall that $m_\varepsilon := \left\lceil \log_2({}^{\varepsilon_0}/_{\varepsilon}) \right\rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$, and for all $k \leq m_\varepsilon - 1$, $\varepsilon_k := \varepsilon_0 2^{-k}$. The proof of the first inequality is deferred to Appendix B. For the second one, for all $k \in [m_\varepsilon]$, define $\nu_k := \inf_{u \in [\varepsilon_{k-1},\, \varepsilon_{k-2}]} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u]},\ {}^{u}/_{L}\right)$, fix any $\eta_k > 0$, and take $u_k \in [\varepsilon_{k-1}, \varepsilon_{k-2}]$ such that $\mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u_k]},\ {}^{u_k}/_{L}\right) \leq \nu_k + \eta_k$. Then, for all $k \in [m_\varepsilon]$,[5]

$$\mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]},\ \frac{\varepsilon_k}{L}\right) \leq \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u_k]},\ \frac{\varepsilon_k}{L}\right) \leq \left(4\frac{u_k}{\varepsilon_k}\right)^d \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u_k]},\ \frac{u_k}{L}\right) \leq 16^d(\nu_k + \eta_k) \ ,$$

where the second inequality follows from Lemma 14. Thus

$$\begin{aligned}
S_{\mathrm{NC}}(f, \varepsilon) &= \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, \varepsilon_{k-1}]},\ \frac{\varepsilon_k}{L}\right) \leq \sum_{k=1}^{m_\varepsilon} \frac{16^d\,(\nu_k + \eta_k)}{\varepsilon_{k-1}} (\varepsilon_{k-2} - \varepsilon_{k-1}) \\
&= 16^d \sum_{k=1}^{m_\varepsilon} \frac{\inf_{u \in [\varepsilon_{k-1},\, \varepsilon_{k-2}]} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u]},\ {}^{u}/_{L}\right)}{\varepsilon_{k-1}} (\varepsilon_{k-2} - \varepsilon_{k-1}) + 16^d \sum_{k=1}^{m_\varepsilon} \eta_k \\
&\leq 2 \cdot 16^d \sum_{k=1}^{m_\varepsilon} \int_{\varepsilon_{k-1}}^{\varepsilon_{k-2}} \frac{\mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u]},\ {}^{u}/_{L}\right)}{u} \,\mathrm{d}u + 16^d \sum_{k=1}^{m_\varepsilon} \eta_k \ .
\end{aligned}$$

The result follows by noting that $\varepsilon_k \geq \max(u/4,\ \varepsilon)$ for any $k \in [m_\varepsilon]$ and all $u \in [\varepsilon_{k-1}, \varepsilon_{k-2}]$, summing, recalling that $\mathcal{X}_{(a,b]} = \varnothing$ for all $\varepsilon_0 \leq a < b$, and taking the limits for $\eta_k \to 0$. □

The previous result yields two direct corollaries, obtained by upper bounding $\mathcal{X}_{(\max(u/4,\, \varepsilon),\, u]}$ with either $\mathcal{X}_{(u/4,\, u]}$ or $\mathcal{X}_{(\varepsilon,\, u]}$. The latter has immediate consequences in terms of *near-optimality dimension* [Bubeck et al., 2011]. Indeed, if there exist two constants $C^\star > 0$ and $d^\star \in (0, d]$ such that $\mathcal{N}\left(\mathcal{X}_u, {}^{u}/_{L}\right) \leq C^\star / u^{d^\star}$ for all $u \in (0, \varepsilon_0]$, then immediately there exists $C > 0$ such that $\int_\varepsilon^\infty \mathcal{N}\left(\mathcal{X}_{(\varepsilon,\, u]}, {}^{u}/_{L}\right) / u \,\mathrm{d}u \leq C / \varepsilon^{d^\star}$ for all $\varepsilon \in (0, \varepsilon_0]$. Similarly, if there exists $C^\star > 0$ such that $\mathcal{N}\left(\mathcal{X}_u, {}^{u}/_{L}\right) \leq C^\star$ for all $u \in (0, \varepsilon_0]$, then there exists $C > 0$ such that $\int_\varepsilon^\infty \mathcal{N}\left(\mathcal{X}_{(\varepsilon,\, u]}, {}^{u}/_{L}\right) / u \,\mathrm{d}u \leq C \log(\varepsilon_0/\varepsilon)$ for all $\varepsilon \in (0, \varepsilon_0]$. Furthermore, such an integral bound makes it easy to handle multiple different regimes at the same time. This is the case if, for example, there exist $C_1^\star, C_2^\star > 0$ and $u_0 \in (0, \varepsilon_0]$ such that $\mathcal{N}\left(\mathcal{X}_u, {}^{u}/_{L}\right) \leq C_1^\star$ for all $u \in [u_0, \varepsilon_0]$ but $\mathcal{N}\left(\mathcal{X}_u, {}^{u}/_{L}\right) \leq C_2^\star / \varepsilon^{d^\star}$ for all $u \in (0, u_0)$. In this case, our integral bound reflects that there exist $C_1, C_2 > 0$ such that $\int_\varepsilon^\infty \mathcal{N}\left(\mathcal{X}_{(\varepsilon,\, u]}, {}^{u}/_{L}\right) / u \,\mathrm{d}u \leq C_1 \log(\varepsilon_0/\varepsilon)$ for all $\varepsilon \in [u_0, \varepsilon_0]$ but $\int_\varepsilon^\infty \mathcal{N}\left(\mathcal{X}_{(\varepsilon,\, u]}, {}^{u}/_{L}\right) / u \,\mathrm{d}u \leq C_2\big(\log(\varepsilon_0/u_0) + 1/\varepsilon^{d^\star}\big)$ if $\varepsilon \in (0, u_0)$.

Note that if for some $\varepsilon \in (0, \varepsilon_0]$, most points in $[0, 1]^d$ are ε-optimal, i.e., if $\mathcal{X}_\varepsilon$ is large, then $\mathcal{N}\left(\mathcal{X}_\varepsilon, {}^{\varepsilon}/_{L}\right) \approx 1/\varepsilon^d$ is also large and does not reflect the fact that the non-certified problem is easy. This is a folklore criticism that has been made to these types of bounds. Such a mismatching behavior is solved by expressing bounds as we did, in terms of layers $\mathcal{X}_{(\varepsilon,\, u]}$ instead.

ACKNOWLEDGEMENTS

The work of Tommaso Cesari and Sébastien Gerchinovitz has benefitted from the AI Interdisciplinary Institute ANITI, which is funded by the French "Investing for the Future – PIA3" program under the Grant agreement ANR-19-P3IA-0004. Sébastien Gerchinovitz gratefully acknowledges the support of the DEEL project.[6]. This work benefited from the support of the project BOLD from the French national research agency (ANR).

[5] The trick of using $\nu_k := \inf_{u \in [\varepsilon_{k-1}, \varepsilon_{k-2}]} \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u]},\ {}^{u}/_{L}\right)$ is key to handle the cases when $u \mapsto \mathcal{N}\left(\mathcal{X}_{(\varepsilon_k,\, u]},\ {}^{u}/_{L}\right)$ is not non-increasing.

[6] https://www.deel.ai/

References

François Bachoc, Tommaso Cesari, and Sébastien Gerchinovitz. The sample complexity of level set approximation. *arXiv preprint arXiv:2010.13405*, 2020.

Peter L. Bartlett, Victor Gabillon, and Michal Valko. A simple parameter-free and adaptive approach to optimization under a minimal local smoothness assumption. In Aurélien Garivier and Satyen Kale, editors, *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 184–206, Chicago, Illinois, 22–24 Mar 2019. PMLR.

Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: a nonasymptotic theory of independence*. Oxford University Press, 2013.

Clément Bouttier, Tommaso Cesari, and Sébastien Gerchinovitz. Regret analysis of the Piyavskii-Shubert algorithm for global lipschitz optimization. *arXiv preprint arXiv:2002.02390*, 2020.

Sébastien Bubeck, Rémi Munos, Gilles Stoltz, and Csaba Szepesvári. X-armed bandits. *Journal of Machine Learning Research*, 12(May):1655–1695, 2011.

Antonio Cuevas, Ricardo Fraiman, and Beatriz Pateiro-López. On statistical properties of sets fulfilling rolling-type conditions. *Advances in Applied Probability*, 44(2):311–329, 2012.

Richard M. Dudley. The sizes of compact subsets of hilbert space and continuity of Gaussian processes. *Journal of Functional Analysis*, 1(3):290 – 330, 1967.

Pierre Hansen, Brigitte Jaumard, and S-H Lu. On the number of iterations of Piyavskii's global optimization algorithm. *Mathematics of Operations Research*, 16(2):334–350, 1991.

Matthias Horn. Optimal algorithms for global optimization in case of unknown Lipschitz constant. *Journal of Complexity*, 22(1):50–70, 2006.

Yichun Hu, Nathan Kallus, and Xiaojie Mao. Smooth contextual bandits: Bridging the parametric and non-differentiable regret regimes. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 2007–2010. PMLR, 09–12 Jul 2020.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Multi-armed bandits in metric spaces. In *Proceedings of the fortieth annual ACM symposium on Theory of computing*, pages 681–690, 2008.

Robert Kleinberg, Aleksandrs Slivkins, and Eli Upfal. Bandits and experts in metric spaces. *Journal of the ACM*, 66 (4), 2019.

Cédric Malherbe and Nicolas Vayatis. Global optimization of Lipschitz functions. In *Proceedings of the 34th International Conference on Machine Learning*, volume PMLR 70, pages 2314–2323, 2017.

Rémi Munos. Optimistic optimization of a deterministic function without the knowledge of its smoothness. In *Advances in Neural Information Processing Systems 24 (NIPS 2011)*, pages 783–791, 2011.

Rémi Munos. From bandits to Monte-Carlo tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends in Machine Learning*, 7(1):1–130, 2014.

Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2003.

A.G. Perevozchikov. The complexity of the computation of the global extremum in a class of multi-extremum problems. *USSR Computational Mathematics and Mathematical Physics*, 30(2):28–33, 1990.

S.A. Piyavskii. An algorithm for finding the absolute extremum of a function. *USSR Computational Mathematics and Mathematical Physics*, 12(4):57–67, 1972.

Bruno O Shubert. A sequential method seeking the global maximum of a function. *SIAM Journal on Numerical Analysis*, 9(3):379–388, 1972.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2019.

Guenther Walther. Granulometric smoothing. *The Annals of Statistics*, 25(6):2273–2299, 1997.

A Missing proofs of Section 3

In this section we provide all missing details and proofs from Section 3.

A.1 Lower Bound on the Certified Sample Complexity in Dimension $d = 1$

We begin by proving our lower bound on the certified sample complexity in dimension $d = 1$. It shows that in the special one-dimensional case $d = 1$, $S_{\mathrm{C}}(f, \varepsilon)$ provides a tight lower bound on the certified sample complexity of all certified algorithms, up to the factor $1 + m_\varepsilon$, even in the boundary case in which f is globally L-Lipschitz.

Proposition (Proposition 5)**.** *If $d = 1$, let $K = 8$ and $c = 1/96K$. Then, the certified sample complexity of any certified algorithm A for L-Lipschitz functions satisfies, for any globally L-Lipschitz function f and all $\varepsilon \in (0, \varepsilon_0]$, $\sigma(A, f, \varepsilon) > (c/1+m_\varepsilon) S_{\mathrm{C}}(f, \varepsilon)$.*

Proof. As for the proof of Theorem 3, it is sufficient to show that $\tau(f, \varepsilon) > cS_{\mathrm{C}}(f, \varepsilon)/(1 + m_\varepsilon)$. If $cS_{\mathrm{C}}(f, \varepsilon)/(1 + m_\varepsilon) < 1$, then the result follows by $\tau(f, \varepsilon) \geq 1$. Consider then from now on that $cS_{\mathrm{C}}(f, \varepsilon)/(1 + m_\varepsilon) \geq 1$.

Defining $\tilde{\varepsilon}$ as in the proof of Theorem 3, one can prove similarly that $cS_{\mathrm{C}}(f, \varepsilon)/(1+m_\varepsilon) \leq c\mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, \tilde{\varepsilon}/2L}\right)$. From Lemma 14,

$$\mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, \frac{K\tilde{\varepsilon}}{L}}\right) \geq \frac{1}{8K}\mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, \frac{\tilde{\varepsilon}}{2L}}\right) \geq \frac{1}{8K}\frac{S_{\mathrm{C}}(f, \varepsilon)}{m_\varepsilon + 1} \geq 12,$$

because $c = 1/96K$ and $cS_{\mathrm{C}}(f, \varepsilon)/(1 + m_\varepsilon) \geq 1$. Let now $n \leq cS_{\mathrm{C}}(f, \varepsilon)/(1 + m_\varepsilon)$. Then we have $n \leq c(8K)\mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, K\tilde{\varepsilon}/L}\right)$. Thus, by $c(8K) = 1/12$, $n \leq \mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, K\tilde{\varepsilon}/L}\right)/12$, and $\mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, K\tilde{\varepsilon}/L}\right) \geq 12$, we have

$$n \leq \left\lfloor \frac{\mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, \frac{K\tilde{\varepsilon}}{L}}\right)}{2} \right\rfloor - 4. \tag{7}$$

Consider a certified algorithm A for L-Lipschitz functions. Let us consider a $K\tilde{\varepsilon}/L$ packing $0 \leq x_1 < \cdots < x_N \leq 1$ of $\mathcal{X}_{\tilde{\varepsilon}}$ with $N = \mathcal{N}\left(\mathcal{X}_{\tilde{\varepsilon},\, K\tilde{\varepsilon}/L}\right)$. Consider the $\lfloor N/2 \rfloor - 1$ disjoint open segments $(x_1, x_3), (x_3, x_5), \ldots, (x_{2\lfloor N/2 \rfloor - 3}, x_{2\lfloor N/2 \rfloor - 1})$. Then from (7) there exists $i \in \left\{1, 3, \ldots, 2\lfloor N/2 \rfloor - 3\right\}$ such that the segment (x_i, x_{i+2}) does not contain any of the $x_1(A, f), \ldots, x_n(A, f)$. Assume that $x_{i+1} - x_i \leq x_{i+2} - x_{i+1}$ (the case $x_{i+1} - x_i > x_{i+2} - x_{i+1}$ can be treated analogously; we omit these straightforward details for the sake of conciseness). Consider the function $h_{+,\tilde{\varepsilon}} : \mathcal{X} \to \mathbb{R}$ defined by

$$h_{+,\tilde{\varepsilon}}(x) = \begin{cases} f(x) & \text{if } x \in \mathcal{X} \backslash [x_i, x_{i+2}] \\ f(x_i) + L(x - x_i) & \text{if } x \in \mathcal{X} \cap [x_i, x_{i+1}] \\ f(x_i) + L(x_{i+1} - x_i) + (x - x_{i+1})\frac{f(x_{i+2}) - f(x_i) - L(x_{i+1} - x_i)}{x_{i+2} - x_{i+1}} & \text{if } x \in \mathcal{X} \cap (x_{i+1}, x_{i+2}]. \end{cases}$$

We see that $h_{+,\tilde{\varepsilon}}$ is L-Lipschitz (since $x_{i+1} - x_i \leq x_{i+2} - x_{i+1}$). Furthermore, $h_{+,\tilde{\varepsilon}}$ coincides with f on $x_1(A, f), \ldots, x_n(A, f)$. Similarly, consider the function $h_{-,\tilde{\varepsilon}} : \mathcal{X} \to \mathbb{R}$ defined by

$$h_{-,\tilde{\varepsilon}}(x) = \begin{cases} f(x) & \text{if } x \in \mathcal{X} \backslash [x_i, x_{i+2}] \\ f(x_i) - L(x - x_i) & \text{if } x \in \mathcal{X} \cap [x_i, x_{i+1}] \\ f(x_i) - L(x_{i+1} - x_i) + (x - x_{i+1})\frac{f(x_{i+2}) - f(x_i) + L(x_{i+1} - x_i)}{x_{i+2} - x_{i+1}} & \text{if } x \in \mathcal{X} \cap (x_{i+1}, x_{i+2}]. \end{cases}$$

As before, $h_{-,\tilde{\varepsilon}}$ is L-Lipschitz and coincides with f on $x_1(A,f),\dots,x_n(A,f)$.

Consider the case (1) where $x_n^\star(A,f) \in \mathcal{X}\backslash[x_i, x_{i+2}]$. Then, we have, since $x_i \in \mathcal{X}_{\tilde{\varepsilon}}$ and $x_{i+1} - x_i \geq K\tilde{\varepsilon}/L$,

$$h_{+,\tilde{\varepsilon}}(x_{i+1}) - h_{+,\tilde{\varepsilon}}(x_n^\star(A,f)) = f(x_i) + L(x_{i+1} - x_i) - f(x_n^\star(A,f)) \geq -\tilde{\varepsilon} + L\frac{K\tilde{\varepsilon}}{L} = 7\tilde{\varepsilon}.$$

Consider the case (2) where $x_n^\star(A,f) \in \mathcal{X} \cap [x_i, (x_i + x_{i+1})/2]$. Then, we have, since $x_{i+1} - x_i \geq K\tilde{\varepsilon}/L$,

$$\begin{aligned} h_{+,\tilde{\varepsilon}}(x_{i+1}) - h_{+,\tilde{\varepsilon}}(x_n^\star(A,f)) &= f(x_i) + L(x_{i+1} - x_i) - f(x_i) - L(x_n^\star(A,f) - x_i) \\ &\geq L\frac{(x_{i+1} - x_i)}{2} \geq \tilde{\varepsilon}\frac{K}{2} = 4\tilde{\varepsilon}. \end{aligned}$$

Consider the case (3) where $x_n^\star(A,f) \in \mathcal{X} \cap [(x_i + x_{i+1})/2, x_{i+1}]$. Then, we have, since $x_{i+1} - x_i \geq K\tilde{\varepsilon}/L$,

$$h_{-,\tilde{\varepsilon}}(x_i) - h_{-,\tilde{\varepsilon}}(x_n^\star(A,f)) = f(x_i) - f(x_i) + L(x_n^\star(A,f) - x_i) \geq L\frac{(x_{i+1} - x_i)}{2} \geq \tilde{\varepsilon}\frac{K}{2} = 4\tilde{\varepsilon}.$$

Consider the case (4) where $x_n^\star(A,f) \in \mathcal{X} \cap [x_{i+1}, (x_{i+1} + x_{i+2})/2]$. Then, we have, since $x_{i+1} - x_i \geq K\tilde{\varepsilon}/L$, since $x_i, x_{i+2} \in \mathcal{X}_{\tilde{\varepsilon}}$ and since $h_{-,\tilde{\varepsilon}}$ is linear increasing on $[x_{i+1}, x_{i+2}]$ with left value $f(x_i) - L(x_{i+1} - x_i)$ and right value $f(x_{i+2})$,

$$\begin{aligned} h_{-,\tilde{\varepsilon}}(x_i) - h_{-,\tilde{\varepsilon}}(x_n^\star(A,f)) &\geq f(x_i) - \frac{f(x_i) - L(x_{i+1} - x_i) + f(x_{i+2})}{2} \\ &= \frac{f(x_i) - f(x_{i+2})}{2} + L\frac{(x_{i+1} - x_i)}{2} \geq -\frac{\tilde{\varepsilon}}{2} + \frac{K}{2}\tilde{\varepsilon} \geq 3\tilde{\varepsilon}. \end{aligned}$$

Consider the case (5) where $x_n^\star(A,f) \in \mathcal{X} \cap [(x_{i+1} + x_{i+2})/2, x_{i+2}]$. Then, we have, since $x_{i+1} - x_i \geq K\tilde{\varepsilon}/L$, since $x_i, x_{i+2} \in \mathcal{X}_{\tilde{\varepsilon}}$ and since $h_{+,\tilde{\varepsilon}}$ is linear decreasing on $[x_{i+1}, x_{i+2}]$ with left value $f(x_i) + L(x_{i+1} - x_i)$ and right value $f(x_{i+2})$,

$$\begin{aligned} h_{+,\tilde{\varepsilon}}(x_{i+1}) - h_{+,\tilde{\varepsilon}}(x_n^\star(A,f)) &\geq f(x_i) + L(x_{i+1} - x_i) - \frac{f(x_i) + L(x_{i+1} - x_i) + f(x_{i+2})}{2} \\ &= \frac{f(x_i) - f(x_{i+2})}{2} + L\frac{(x_{i+1} - x_i)}{2} \geq -\frac{\tilde{\varepsilon}}{2} + \frac{K}{2}\tilde{\varepsilon} \geq 3\tilde{\varepsilon}. \end{aligned}$$

Hence, in all cases $E_L(A,f,n) \geq 3\tilde{\varepsilon} > \varepsilon$. Hence $\inf_A E_L(A,f,n) > \varepsilon$. Since this has been shown for any $n \leq cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$ we thus have $\tau(f,\varepsilon) > cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$. □

As discussed previously, in the proof of Theorem 3, the constraint that f be L'-Lipschitz with $L' < L$ arose from the fact that we added a small bump function $\pm g_{\tilde{\varepsilon}}$ to f, with the requirement that the new function $f \pm g_{\tilde{\varepsilon}}$ be globally L-Lipschitz. We treat the one-dimensional case differently. For a given algorithm, a given function f and a number of query points smaller than $cS_{\mathrm{C}}(f,\varepsilon)/(1+m_\varepsilon)$, we show the existence of a segment unvisited by the algorithm and containing three close-to-optimal points that are separated enough. We then replace the function f on this segment with an upward or downward hat function which makes the algorithm fail to be ε-optimal. By replacing f with a new function on the segment, rather than adding a bump function to f, we can allow f to have Lipschitz constant arbitrarily close to L.

A.2 The Piyavskii-Shubert Algorithm and Proof of Proposition 6

In this section, we recall the definition of the certified Piyavskii-Shubert algorithm (Algorithm 2, Piyavskii 1972, Shubert 1972) and we show that its sample complexity, in dimension $d \geq 2$, is not lower bounded by $S_{\mathrm{C}}(f,\varepsilon)$ for some functions in the boundary case $L' = L$.

In contrast to the one-dimensional case $d = 1$, the next result shows that there are special cases with $L' = L$ and $d \geq 2$ for which the upper bound $S_{\mathrm{C}}(f,\varepsilon)$ is far from tight. Quantifying the optimal certified sample complexity for dimensions $d \geq 2$ in the boundary case $L' = L$ (or when L' is arbitrarily close to L) is left as an open problem.

Algorithm 2: Certified Piyavskii-Shubert algorithm

input: accuracy $\varepsilon > 0$, Lipschitz constant $L > 0$, norm $\|\cdot\|$, initial guess $\boldsymbol{x}_1 \in \mathcal{X}$
for $i = 1, 2, \ldots$ **do**
pick the next query point $\boldsymbol{x}_i$
observe the value $f(\boldsymbol{x}_i)$
output the recommendation $\boldsymbol{x}_i^\star \leftarrow \operatorname{argmax}_{\boldsymbol{x} \in \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_i\}} f(\boldsymbol{x})$
output the certificate $\gamma_i \leftarrow \mathbb{I}\{\widehat{f}_i^\star - f_i^\star \leq \varepsilon\}$, where $\widehat{f}_i(\cdot) \leftarrow \min_{j \in [i]} \{f(\boldsymbol{x}_j) + L\,\|\boldsymbol{x}_j - (\cdot)\|\}$,
$\widehat{f}_i^\star \leftarrow \max_{\boldsymbol{x} \in \mathcal{X}} \widehat{f}_i(\boldsymbol{x})$, $f_i^\star \leftarrow \max_{j \in [i]} f(\boldsymbol{x}_j)$, and let $\boldsymbol{x}_{i+1} \in \operatorname{argmax}_{\boldsymbol{x} \in \mathcal{X}} \widehat{f}_i(\boldsymbol{x})$

Proposition 6. *Let $d \geq 2$, $\mathcal{X} := B_1$, and $\|\cdot\|$ be a norm. The certified Piyavskii-Shubert algorithm (Algorithm 2) with initial guess $\boldsymbol{x}_1 := \mathbf{0}$ is a certified algorithm for L-Lipschitz functions satisfying, for the globally L-Lipschitz function $f := L\,\|\cdot\|$ and any $\varepsilon \in (0, \varepsilon_0)$, σ(Piyavskii-Shubert, $f, \varepsilon) = 2 \ll c/\varepsilon^{d-1} \leq S_{\mathrm{C}}(f, \varepsilon)$, where $c > 0$ is a constant independent of ε.*

Proof. Fix any $\varepsilon \in (0, \varepsilon_0)$. When f is at least L-Lipschitz around a maximizer (Assumption 1), then $\max_{\boldsymbol{x} \in \mathcal{X}} \widehat{f}_i(\boldsymbol{x}) \geq \max_{\boldsymbol{x} \in \mathcal{X}} f(\boldsymbol{x})$ for all $i \in \mathbb{N}^*$ (for a proof of this fact, see Bouttier et al. 2020). Hence if $\gamma_i = 1$, then $\max_{\boldsymbol{x} \in \mathcal{X}} f(x) - f(\boldsymbol{x}_i^\star) \leq \widehat{f}_i(\boldsymbol{x}) - f_i^\star \leq \varepsilon$, and $\boldsymbol{x}_i^\star$ is necessarily ε-optimal. This shows that the certified Piyavskii-Shubert algorithm is indeed a certified algorithm for L-Lipschitz functions. Then, since $\varepsilon_0 = L$, we have that $\widehat{f}_1 = f$, $\gamma_1 = 0$, and $\boldsymbol{x}_2$ belongs to the the unit sphere, i.e., $\boldsymbol{x}_2$ is a maximizer of f. Since $\widehat{f}_2 = f$, we have that $\gamma_2 = 1$, hence σ(Piyavskii-Shubert, $f, \varepsilon) = 2$. Finally, by definition (4), we have $S_{\mathrm{C}}(f, \varepsilon) \geq \mathcal{N}(\operatorname{argmax}_{\mathcal{X}} f, \varepsilon/L)$. Since $\operatorname{argmax}_{\mathcal{X}} f$ is the unit sphere, there exists a constant c, only depending on d, $\|\cdot\|$ and L, such that $S_{\mathrm{C}}(f, \varepsilon) \geq c/\varepsilon^{d-1}$. □

We give some intuition on the previous proposition. Consider a function f that has Lipschitz constant exactly L, and a pair of points in $\mathcal{X}$ whose respective values of f are maximally distant, that is the difference of values of f is exactly L times the norm of the input difference. This configuration provides strong information on the value of the global maximum of f, as is illustrated in the proof of Proposition 6. Another interpretation is that when f has Lipschitz constant exactly L, there is less flexibility for the L-Lipschitz function g that yields the maximal optimization error in (6).

B Missing proofs of Section 5

In this section, we provide the missing part of the proof of Theorem 9.

Fix any $\varepsilon \in (0, \varepsilon_0]$ and recall that $m_\varepsilon := \lceil \log_2(\varepsilon_0/\varepsilon) \rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$, and for all $k \leq m_\varepsilon - 1$, $\varepsilon_k := \varepsilon_0 2^{-k}$. We begin

by proving the first inequality:

$$
\begin{aligned}
&\int_{\varepsilon}^{\infty} \frac{\mathcal{N}\left(\mathcal{X}_{\left(\max\left(\varepsilon,\frac{u}{4}\right),u\right]},\ \frac{u}{L}\right)}{u}\,\mathrm{d}u = \sum_{k=-1}^{m_\varepsilon}\int_{\varepsilon_k}^{\varepsilon_{k-1}} \frac{\mathcal{N}\left(\mathcal{X}_{\left(\max\left(\varepsilon,\frac{u}{4}\right),u\right]},\ \frac{u}{L}\right)}{u}\,\mathrm{d}u \\
&\leq \sum_{k=-1}^{m_\varepsilon}\int_{\varepsilon_k}^{\varepsilon_{k-1}} \frac{\mathcal{N}\left(\mathcal{X}_{\left(\max\left(\varepsilon,\frac{\varepsilon_k}{4}\right),\varepsilon_{k-1}\right]},\ \frac{\varepsilon_k}{L}\right)}{\varepsilon_k}\,\mathrm{d}u \leq \sum_{k=-1}^{m_\varepsilon}\mathcal{N}\left(\mathcal{X}_{\left(\max\left(\varepsilon,\frac{\varepsilon_k}{4}\right),\varepsilon_{k-1}\right]},\ \frac{\varepsilon_k}{L}\right) \\
&= \sum_{k=-1}^{m_\varepsilon}\left(\mathcal{N}\left(\mathcal{X}_{\left(\frac{\varepsilon_k}{4},\,\varepsilon_{k-1}\right]},\ \frac{\varepsilon_k}{L}\right)\mathbb{I}_{k\in\{-1,\dots,m_\varepsilon-3\}} + \mathcal{N}\left(\mathcal{X}_{(\varepsilon,\,\varepsilon_{k-1}]},\ \frac{\varepsilon_k}{L}\right)\mathbb{I}_{k\in\{m_\varepsilon-2,\dots,m_\varepsilon\}}\right) \\
&\leq \sum_{k=-1}^{m_\varepsilon}\left(\sum_{h=k}^{k+2}\mathcal{N}\left(\mathcal{X}_{(\varepsilon_h,\,\varepsilon_{h-1}]},\ \frac{\varepsilon_h}{L}\right)\mathbb{I}_{k\in\{-1,\dots,m_\varepsilon-3\}} + \sum_{h=k}^{m_\varepsilon}\mathcal{N}\left(\mathcal{X}_{(\varepsilon_h,\,\varepsilon_{h-1}]},\ \frac{\varepsilon_h}{L}\right)\mathbb{I}_{k\in\{m_\varepsilon-2,\dots,m_\varepsilon\}}\right) \\
&= \sum_{k=-1}^{0}(\dots) + \sum_{k=1}^{m_\varepsilon}(\dots) \leq 5S_{\mathrm{NC}}(f,\varepsilon)\ ,
\end{aligned}
$$

where the first and third inequalities follow by the monotonicity properties of packing numbers, the second one by $\varepsilon_{k-1}-\varepsilon_k \leq \varepsilon_k$ for all $k\in[m_\varepsilon]$, and the last one by noting that both brackets[7] $(\dots)$ contain at most three non-zero terms $\mathcal{N}\left(\mathcal{X}_{(\varepsilon_h,\,\varepsilon_{h-1}]},\ \frac{\varepsilon_h}{L}\right)$, with $h\in[m_\varepsilon]$, that overlap at most $2+3$ times, and nothing else. This proves the first part of the theorem.

C Packing vs Volume

The two prototypical sample complexity statements involving assumptions on volumes (8)–(9) (see e.g., Perevozchikov 1990) and packing numbers (10)–(11) (e.g., Munos 2014) for algorithms applied to functions f that are L-Lipschitz around a maximizer and accuracies $\varepsilon\in(0,\varepsilon_0]$ are:

$$
\begin{aligned}
&\exists r\in(0,1],\ \exists C>0,\ \forall u\in(0,\varepsilon_0],\ \mathrm{vol}(\mathcal{X}_u)\leq Cu^{dr} && (8)\\
&\qquad\qquad \implies \exists c_2>0,\ \forall n\geq c_2\big(\ln(1/\varepsilon)\mathbb{I}_{r=1}+(1/\varepsilon)^{d(1-r)}\mathbb{I}_{r\neq 1}\big),\ \boldsymbol{x}_n^\star\in\mathcal{X}_\varepsilon\ , && (9)\\
&\exists d^\star\in[0,d],\ \exists C^\star>0,\ \forall u\in(0,\varepsilon_0],\ \mathcal{N}(\mathcal{X}_u,u/L)\leq C^\star(\varepsilon_0/\varepsilon)^{d^\star} && (10)\\
&\qquad\qquad \implies \exists c_2>0,\ \forall n\geq c_2\big(\ln(1/\varepsilon)\mathbb{I}_{d^\star=0}+(1/\varepsilon)^{d^\star}\mathbb{I}_{r\neq 1}\big),\ \boldsymbol{x}_n^\star\in\mathcal{X}_\varepsilon\ . && (11)
\end{aligned}
$$

In the two following sections, we will study the relationship between sample complexity statements expressed under assumptions on volumes (8) and packing numbers (10).

C.1 Which Assumption is Stronger

The first result suggests that packing assumptions (10) are stronger than volume's (8).

Proposition 10. *If $E\subset\mathbb{R}^d$ is bounded and Lebesgue-measurable and L-Lipschitz around a maximizer, letting $c:=v_{1/L}$, for all $u\in(0,\varepsilon_0]$,*

$$\mathrm{vol}(\mathcal{X}_u)\leq c\,\mathcal{N}\left(\mathcal{X}_u,\ \frac{u}{L}\right)u^d\ .$$

Proof. For all $u\in(0,\varepsilon_0]$, $\mathrm{vol}(\mathcal{X}_u)\leq \mathcal{M}\left(\mathcal{X}_u,\ \frac{u}{L}\right)\mathrm{vol}\left(\frac{u}{L}B_1\right)\leq \mathcal{N}\left(\mathcal{X}_u,\ \frac{u}{L}\right)u^d v_{1/L}$, where in the first inequality we upper bounded the volume of a set with the volume of the balls of a smallest (u/L)-cover and in the second one we used the fact that covering numbers are always smaller than packing numbers (14). □

[7] We wrote $(\dots)$ to avoid copying twice the long expression on the previous line.

The previous proposition implies in particular that if $d^\star \in [0, d)$ is a near-optimality dimension of f for a constant and $C^\star > 0$, i.e., if for all $u \in (0, \varepsilon_0]$, $\mathcal{N}(\mathcal{X}_u,\ u/L) \leq C^\star \left(\varepsilon_0/u\right)^{d^\star}$ (as in (10)), then $\mathrm{vol}(\mathcal{X}_u) \leq Cu^{dr}$, where $C := v_{1/L} C^\star \varepsilon_0^{d^\star}$ and $r = 1 - d^\star/d$, recovering the assumption on volume (8). The assumption on packing looks therefore stronger. However, we show now that if f is globally Lipschitz (and its domain is not too pathological), the converse is also true, i.e., that assumptions on packings and volumes bounds are essentially equivalent.

Proposition 11. *If f is globally L-Lipschitz and $\mathcal{X}$ satisfies Assumption 3 with $r_0 > 0, \gamma \in (0, 1]$, then, for all $0 < w < u < 2Lr_0$,*

$$\mathcal{N}\left(\mathcal{X}_u,\ \frac{u}{L}\right) \leq \frac{1}{\gamma} \frac{\mathrm{vol}\big(\mathcal{X}_{(3/2)u}\big)}{\mathrm{vol}\left(\frac{u}{2L} B_1\right)} \qquad \textit{and} \qquad \mathcal{N}\left(\mathcal{X}_{(w,u]},\ \frac{w}{L}\right) \leq \frac{1}{\gamma} \frac{\mathrm{vol}\big(\mathcal{X}_{(w/2,\, 3u/2]}\big)}{\mathrm{vol}\left(\frac{w}{2L} B_1\right)} \,.$$

Proof. Fix any $u > w > 0$. Let $\eta_1 := \frac{u}{L}$, $\eta_2 := \frac{w}{L}$, $E_1 := \mathcal{X}$, $E_2 := \mathcal{X}_w^c$, and $i \in [2]$. Note that for any $\eta > 0$ and $A, E \subset \mathbb{R}^d$, the balls of radius $\eta/2$ centered at the elements of an η-packing of A intersected with E are all disjoint and included in $(A + B_{\eta/2}) \cap E$. Thus, $\mathcal{N}\left(\mathcal{X}_u \cap E_i,\ \eta_i\right) \leq \mathrm{vol}\big((\mathcal{X}_u \cap E_i + B_{\eta_i/2}) \cap \mathcal{X}\big) / \mathrm{vol}(B_{\eta_i/2} \cap \mathcal{X})$. To lower bound the denominator, simply apply Assumption 3 to obtain $\mathrm{vol}\big(B_{\eta_i/2}(\boldsymbol{x}) \cap \mathcal{X}\big) \geq \gamma\, \mathrm{vol}(B_{\eta_i/2})$. To upper bound the numerator, take an arbitrary point $\boldsymbol{x}_i \in (\mathcal{X}_u \cap E_i + B_{\eta_i/2}) \cap \mathcal{X}$. By definition of Minkowski sum, there exists $\boldsymbol{x}_i' \in \mathcal{X}_u \cap E_i$ such that $\|\boldsymbol{x}_i - \boldsymbol{x}_i'\| \leq \eta_i/2$. Hence $f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_i) \leq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_i') + \big|f(\boldsymbol{x}_i') - f(\boldsymbol{x}_i)\big| \leq u + L(\eta_i/2) \leq (3/2)u$. This implies that $\boldsymbol{x}_i \in \mathcal{X}_{(3/2)u}$, which proves the first inequality. For the second one, note that $\boldsymbol{x}_2$ satisfies $f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_2) \geq f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_2') - \big|f(\boldsymbol{x}_2') - f(\boldsymbol{x}_2)\big| \geq w - L(\eta_2/2) = (1/2)w$. ☐

The previous proposition implies in particular that if there exist $r \in (0, 1]$ and $C > 0$ such that, for all $u \in (0, \varepsilon_0]$, $\mathrm{vol}(\mathcal{X}_u) \leq Cu^{dr}$ (as in (8)), then $d^\star = d(1-r)$ is a near-optimality dimension of f for $C^\star := \big(C(3/2)^{dr}\big)/(\gamma v_{1/2L} \varepsilon_0^{d^\star})$, i.e., for all $u \in (0, \varepsilon_0]$, $\mathcal{N}(\mathcal{X}_u,\ u/L) \leq C^\star \left(\varepsilon_0/\varepsilon\right)^{d^\star}$, recovering the assumption on packing numbers (10). Therefore, for globally L-Lipschitz functions, the two assumptions are essentially equivalent.

C.2 Which Statement is More General

We recall the prototypical volume-base statement presented at the beginning of Appendix C.

$$\exists r \in (0, 1],\, \exists C > 0,\, \forall u \in (0, \varepsilon_0],\, \mathrm{vol}(\mathcal{X}_u) \leq Cu^{dr} \tag{12}$$

$$\implies \exists c_2 > 0,\, \forall n \geq c_2\big(\ln(1/\varepsilon)\mathbb{I}_{r=1} + (1/\varepsilon)^{d(1-r)}\mathbb{I}_{r\neq 1}\big),\ \boldsymbol{x}_n^\star \in \mathcal{X}_\varepsilon \,. \tag{13}$$

In the previous section we showed that assumptions on packing numbers imply the above assumptions (12) on volumes when f is Lipschitz around a maximizer (and measurable), but for the converse to be true f has to be globally Lipschitz. Although one might believe that this would make volume assumption better (because they are weaker), it turns out that they are in fact *too weak* in general when the function is only Lipschitz around a maximizer to imply the conclusion (13). We prove this in the following proposition.

Proposition 12. *Let $L > 0$ and $r \in (0, 1]$. Then, for any deterministic algorithm and all multiplicative constants $c_2 > 0$, there exist a function $f\colon \mathcal{X} \to \mathbb{R}$ and an accuracy $\varepsilon' > 0$ such that:*

1. *f is L-Lipschitz around a maximizer with respect to the Euclidean norm $\|\cdot\|_2$;*
2. *there exists $c_1 > 0$ such that $\mathrm{vol}(\mathcal{X}_u) \leq c_1 u^{dr}$ for all $u \in (0, \varepsilon_0]$;*
3. *for any $\varepsilon \in (0, \varepsilon')$ and all $n \leq \widetilde{n}_\varepsilon$, $\boldsymbol{x}_n^\star \notin \mathcal{X}_\varepsilon$, where $\widetilde{n}_\varepsilon := c_2\big(\ln(1/\varepsilon)\mathbb{I}_{r=1} + (1/\varepsilon)^{d(1-r)}\mathbb{I}_{r\neq 1}\big)$.*

Proof. Fix any deterministic algorithm and $c_2 > 0$. Denote by $\boldsymbol{x}_1(g), \boldsymbol{x}_2(g), \ldots$ the queries and by $\boldsymbol{x}_1^\star(g), \boldsymbol{x}_2^\star(g), \ldots$ the recommendations of the algorithm when applied to the constant function $g = 0$. Assume first that $r \neq 1$. Let

$$\varepsilon' \in \big(0,\, \min\{(L/2)^{1/r}/(8v_1c_2)^{1/rd},\, c_2^{1/d(1-r)},\, 1\}\big],$$

then fix any $\varepsilon \in (0, \varepsilon')$ and define $n_\varepsilon := \big\lceil c_2/\varepsilon^{d(1-r)} \big\rceil \geq \widetilde{n}_\varepsilon$. Let

$$E := \big\{\boldsymbol{x}_1(g), \boldsymbol{x}_1^\star(g), \ldots, \boldsymbol{x}_{n_\varepsilon}(g), \boldsymbol{x}_{n_\varepsilon}^\star(g)\big\}.$$

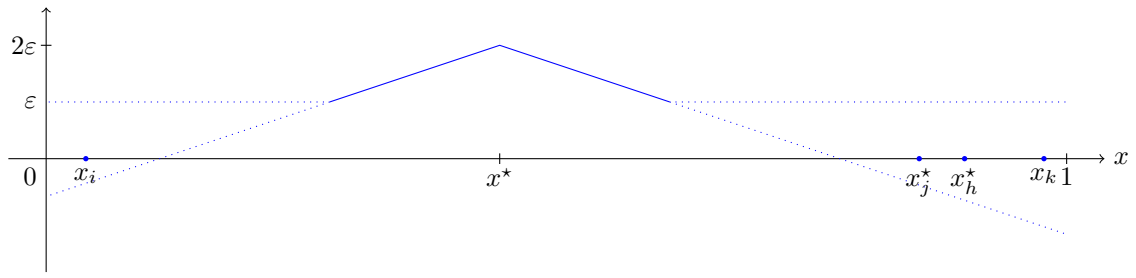

Figure 1: The graph of a function f for which volume-based results do not apply.

Note that there exists $\boldsymbol{x}^\star \in \mathcal{X}$ such that $\min_{\boldsymbol{x}\in E} \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \geq 1/(4v_1 n_\varepsilon)^{1/d} \geq \varepsilon^{1-r}/(8v_1 c_2)^{1/d}$. Indeed, if the first inequality did not hold, than the spheres with radius $1/(4v_1 n_\varepsilon)^{1/d}$ centered at the points in E would cover $\mathcal{X}$, but this could not happen since the largest volume that they could cover is $(2n_\varepsilon)\, v_{1/(4v_1 n_\varepsilon)^{1/d}} = 1/2 < 1$. Fix any such $\boldsymbol{x}^\star$, note that $B_{2\varepsilon/L}(\boldsymbol{x}^\star) \cap E = \varnothing$, and let $Q := \mathbb{Q}^d \cap \big(\mathcal{X} \setminus \big(B_{\varepsilon/L}(\boldsymbol{x}^\star) \cup E\big)\big)$. We can now define a function f that attains its maximum at $\boldsymbol{x}^\star$, has value ε in the zero-volume dense set Q, and satisfies $f(\boldsymbol{x}) = 0$ for all $\boldsymbol{x} \in E$. Concretely, consider the function $f\colon \mathcal{X} \to \mathbb{R}$ defined for all $\boldsymbol{x} \in \mathcal{X}$, by (see Fig. 1)

$$f(\boldsymbol{x}) = \big(2\varepsilon - L\,\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big)\, \mathbb{I}_{\mathcal{X}\setminus(Q\cup E)}(\boldsymbol{x}) + \varepsilon\, \mathbb{I}_Q(\boldsymbol{x})\ .$$

Then f is L-Lipschitz around $\boldsymbol{x}^\star$. Moreover, letting $c_1 := v_{1/L}\, \max\big\{\varepsilon_0^{d(1-r)}, 1\big\}$, we have

$$\mathrm{vol}(\mathcal{X}_u) = \mathrm{vol}\big(B_{u/L}(\boldsymbol{x}^\star)\big) = v_{1/L}\, u^d = v_{1/L}\, u^{d(1-r)}\, u^{dr} \leq c_1 u^{dr}\ ,$$

for all $u \in (0, \varepsilon_0]$, because f is almost everywhere equal to $\boldsymbol{x} \mapsto 2\varepsilon - L\,\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2$. Finally, by definition, for all $n \leq n_\varepsilon$, the recommendation $\boldsymbol{x}_n^\star$ of the algorithm after n evaluations of f satisfies $f(\boldsymbol{x}_n^\star) = 0$, hence $\boldsymbol{x}_n^\star \notin \mathcal{X}_\varepsilon$. The case $r = 1$ can be treated similarly, by letting $n_\varepsilon := \big\lceil c_2/\varepsilon^{d(1-\rho)} \big\rceil$ for some $\rho \in (0,1)$ and noting that $n_\varepsilon \geq \widetilde{n}_\varepsilon$ for all sufficiently small ε. □

The result is not surprising since for ε, f as in the proof, $\mathcal{N}\big(\mathcal{X}_\varepsilon, \varepsilon/2L\big) \approx 1/\varepsilon^d$ because of the dense subset Q. Packing numbers of ε-optimizers better capture the difficulty of the problem. In conclusion, packing assumptions are be the best between the two, as they imply convergence result also in the more general case of functions f that are only Lipschitz around a maximizer.

D Useful Results on Packing, Covering, and Volume

For the sake of completeness, we recall the definitions of packing and covering numbers, as well as several useful results on packing, covering, and volume.

D.1 Definitions of Packing and Covering Numbers

For any bounded set $A \subset \mathbb{R}^d$ and any real number $r > 0$, the *r-packing number* of A is the largest cardinality of an r-packing of A, that is,

$$\mathcal{N}(A, r) := \sup\left\{k \in \mathbb{N}^* : \exists \boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in A, \min_{i\neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\| > r\right\}$$

if A is nonempty, zero otherwise.

The *r-covering number* of A is the smallest cardinality of an r-covering of A, that is,

$$\mathcal{M}(A, r) := \min\left\{k \in \mathbb{N}^* : \exists \boldsymbol{x}_1, \dots, \boldsymbol{x}_k \in \mathbb{R}^d, \forall \boldsymbol{x} \in A, \exists i \in \{1, \dots, k\}, \|\boldsymbol{x} - \boldsymbol{x}_i\| \leq r\right\}$$

if A is nonempty, zero otherwise.

D.2 Useful Inequalities

Covering numbers and packing numbers are closely related. In particular, the following well-known inequalities hold—see, e.g., Wainwright [2019, Lemmas 5.5 and 5.7, with permuted notation of $\mathcal{M}$ and $\mathcal{N}$].[8]

Lemma 13. *Fix any norm $\|\cdot\|$. For any bounded set $A \subset \mathbb{R}^d$ and any real number $r > 0$,*

$$\mathcal{N}(A, 2r) \leq \mathcal{M}(A, r) \leq \mathcal{N}(A, r) \ . \tag{14}$$

Furthermore, for all $\delta > 0$ and all $r > 0$,

$$\mathcal{M}\left(B_\delta, r\right) \leq \left(1 + 2\frac{\delta}{r}\mathbb{I}_{r<\delta}\right)^d \ . \tag{15}$$

We now state a known lemma about packing numbers at different scales. This is the go-to result for rescaling packing numbers.

Lemma 14. *For any $E \subset \mathcal{X}$ and $0 < r_1 \leq r_2 < \infty$, we have*

$$\mathcal{N}\left(E, r_1\right) \leq \left(\frac{4r_2}{r_1}\right)^d \mathcal{N}\left(E, r_2\right).$$

Proof. Consider a r_1-packing of E with cardinality $N_1 = \mathcal{N}\left(E, r_1\right)$, written $F = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{N_1}\}$. Consider then the following iterative procedure. Let $F_0 = F$ and initialize $k = 1$. While F_{k-1} is non-empty, let $\boldsymbol{y}_k$ be any point in F_{k-1}, let F_k be obtained from F_{k-1} by removing the points at $||.||$-distance less or equal to r_2 from $\boldsymbol{y}_k$ and increase k by one. Then this procedure yields an r_2 packing of E with cardinality equal to the number of steps (the final value of k). At each step, to the set of removed points we can associate a set of disjoint balls with radius $r_1/2$ in a ball with radius $2r_2$. Hence the number of removed points is smaller or equal to $v_{2r_2}/v_{r_1/2} = (4r_2/r_1)^d$. Hence the number of steps is larger or equal to $\mathcal{N}\left(E, r_1\right)(r_1/4r_2)^d$. This concludes the proof since $\mathcal{N}\left(E, r_2\right)$ is larger than this number of steps. □

We now prove a result on the sum of volumes of overlapping layers that is used in the proof of Theorem 7.

Lemma 15. *If f is L-Lipschitz around a maximizer (Assumption 1) and Lebesgue-measurable, fix $\varepsilon > 0$ and recall that $m_\varepsilon := \left\lceil \log_2(\varepsilon_0/\varepsilon) \right\rceil$, $\varepsilon_{m_\varepsilon} := \varepsilon$, and for all $k \leq m_\varepsilon - 1$, $\varepsilon_k =: \varepsilon_0 2^{-k}$. Then, there exist $c, C > 0$ (e.g., $c = 1/5$ and $C := 2 \cdot 16^d$) such that, for all $\varepsilon \in (0, \varepsilon_0]$,*

$$\frac{\mathrm{vol}\big(\mathcal{X}_{2\varepsilon}\big)}{\varepsilon^d} + \sum_{k=1}^{m} \frac{\mathrm{vol}\left(\mathcal{X}_{\left(\frac{1}{2}\varepsilon_k,\ 2\varepsilon_{k-1}\right]}\right)}{\varepsilon_{k-1}^d} \leq 8^d \left(\frac{\mathrm{vol}(\mathcal{X}_\varepsilon)}{\varepsilon^d} + \sum_{i=1}^{m} \frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_k,\ \varepsilon_{k-1}]}\big)}{\varepsilon_{k-1}^d}\right) .$$

[8] The definition of r-covering number of a subset A of $\mathbb{R}^d$ implied by Wainwright [2019, Definition 5.1] is slightly stronger than the one used in our paper, because elements $x_1, \ldots, x_N$ of r-covers belong to A rather than just $\mathbb{R}^d$. Even if we do not need it for our analysis, Inequality (15) holds also in this stronger sense.

Proof. To avoid clutter, we denote m_ε simply by m. The left hand side can be upper bounded by

$$\begin{aligned}
&\frac{\mathrm{vol}\big(\mathcal{X}_\varepsilon\big) + \mathrm{vol}\big(\mathcal{X}_{(\varepsilon_m,\,\varepsilon_{m-1}]}\big) + \mathrm{vol}\big(\mathcal{X}_{(\varepsilon_{m-1},\,\varepsilon_{m-2}]}\big)}{\varepsilon^d} \\
&\qquad + \sum_{k=1}^{m-2} \frac{\mathrm{vol}\left(\mathcal{X}_{(\varepsilon_{k+1},\,\varepsilon_k]}\right) + \mathrm{vol}\left(\mathcal{X}_{(\varepsilon_k,\,\varepsilon_{k-1}]}\right) \mathrm{vol}\left(\mathcal{X}_{(\varepsilon_{k-1},\,\varepsilon_{k-2}]}\right)}{\varepsilon_{k-1}^d} \\
&\qquad + \frac{\mathrm{vol}\left(\mathcal{X}_\varepsilon\right) + \mathrm{vol}\big(\mathcal{X}_{(\varepsilon_m,\,\varepsilon_{m-1}]}\big) + \mathrm{vol}\big(\mathcal{X}_{(\varepsilon_{m-1},\,\varepsilon_{m-2}]}\big) + \mathrm{vol}\big(\mathcal{X}_{(\varepsilon_{m-2},\,\varepsilon_{m-3}]}\big)}{\varepsilon_{m-2}^d} \\
&\qquad + \frac{\mathrm{vol}\left(\mathcal{X}_\varepsilon\right) + \mathrm{vol}\big(\mathcal{X}_{(\varepsilon_m,\,\varepsilon_{m-1}]}\big) + \mathrm{vol}\big(\mathcal{X}_{(\varepsilon_{m-1},\,\varepsilon_{m-2}]}\big)}{\varepsilon_{m-1}^d} \\
&\leq 3\frac{\mathrm{vol}\big(\mathcal{X}_\varepsilon\big)}{\varepsilon^d} + (2^d+2)\frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_m,\,\varepsilon_{m-1}]}\big)}{\varepsilon_{m-1}^d} + (4^d+2^d+1)\frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_{m-1},\,\varepsilon_{m-2}]}\big)}{\varepsilon_{m-2}^d} \\
&\qquad + \frac{1}{2^d}\sum_{k=2}^{m-1} \frac{\mathrm{vol}\left(\mathcal{X}_{(\varepsilon_k,\,\varepsilon_{k-1}]}\right)}{\varepsilon_{k-1}^d} + \sum_{k=1}^{m-2} \frac{\mathrm{vol}\left(\mathcal{X}_{(\varepsilon_k,\,\varepsilon_{k-1}]}\right)}{\varepsilon_{k-1}^d} + 2^d \sum_{k=1}^{m-3} \frac{\mathrm{vol}\left(\mathcal{X}_{(\varepsilon_k,\,\varepsilon_{k-1}]}\right)}{\varepsilon_{k-1}^d} \\
&= 3\frac{\mathrm{vol}\big(\mathcal{X}_\varepsilon\big)}{\varepsilon^d} + \frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_m,\,\varepsilon_{m-1}]}\big)}{\varepsilon_{m-1}^d} + 4^d\frac{\mathrm{vol}\big(\mathcal{X}_{(\varepsilon_{m-1},\,\varepsilon_{m-2}]}\big)}{\varepsilon_{m-2}^d} \\
&\qquad + \frac{1}{2^d}\sum_{k=2}^{m-1} \frac{\mathrm{vol}\left(\mathcal{X}_{(\varepsilon_k,\,\varepsilon_{k-1}]}\right)}{\varepsilon_{k-1}^d} + \sum_{k=1}^{m} \frac{\mathrm{vol}\left(\mathcal{X}_{(\varepsilon_k,\,\varepsilon_{k-1}]}\right)}{\varepsilon_{k-1}^d} + 2^d \sum_{k=1}^{m} \frac{\mathrm{vol}\left(\mathcal{X}_{(\varepsilon_k,\,\varepsilon_{k-1}]}\right)}{\varepsilon_{k-1}^d}
\end{aligned}$$

where we applied several times the definition of the ε_k's, the inequality follows by $1/\varepsilon^d + 1/\varepsilon_{m-1}^d + 1/\varepsilon_{m-2}^d \leq \min\big\{3(1/\varepsilon^d),\ (2^d+2)(1/\varepsilon_{m-1}^d),\ (4^d+2^d+1)(1/\varepsilon_{m-2}^d)\big\}$, and the proof is concluded observing that $\max(3, 1, 4^d) = 4^d$ and $4^d + 1/2^d + 1 + 2^d \leq 8^d$. □

The previous result can be generalized to arbitrary partitions and rescaling of the layers. We prove it here for the interested reader.

Proposition 16. *If f is L-Lipschitz around a maximizer (Assumption 1) and Lebesgue-measurable, fix any $m \in \mathbb{N}$ and let $0 := \xi_{m+1} < \xi_m \leq \xi_{m-1} \leq \ldots \leq \xi_1 \leq \xi_0 := \varepsilon_0$. For any $k \in \{0, \ldots, m\}$, let also $a_k \in (0, 1]$, $b_k \geq 1$,*

$$\begin{aligned}
i_k &:= \min\big\{ i \in \{k, \ldots, m+1\} : \xi_i \leq a_k \xi_k \big\}\ , \\
j_k &:= \begin{cases} 0\ , & \text{if } b_k\xi_k > \xi_0\ , \\ \max\big\{ j \in \{0, \ldots, k\} : b_k\xi_k \leq \xi_j \big\}\ , & \text{otherwise}\ . \end{cases}
\end{aligned}$$

For all $h \in [m+1]$, define $n_{a,h}$ and $n_{b,h}$ as the following numbers of overlaps with $(\xi_h,\, \xi_{h-1}]$:

$$\begin{aligned}
n_{a,h} &:= \#\big\{ k \in \{2, \ldots, m+1\} : i_{k-1} \geq k \text{ and } k \leq h \leq i_{k-1} \big\}\ , \\
n_{b,h} &:= \#\big\{ k \in [m] : j_k \leq k-1 \text{ and } j_k + 1 \leq h \leq k \big\}\ .
\end{aligned}$$

Then, let $c_k := (\xi_{j_k+1}/\xi_{k+1})^d$ for all $k \in [m-1]$, $c_m := (\xi_{j_m+1}/\xi_m)^d$, and $c := \max_{k\in[m]} c_k$. If $C := 1 + c\max\big(n_{a,m+1},\, \max_{h\in[m]}(n_{a,h} + n_{b,h})\big)$, we have, for any function $f\colon \mathcal{X} \to \mathbb{R}$,

$$\frac{\mathrm{vol}(\mathcal{X}_{b_m\xi_m})}{\xi_m^d} + \sum_{k=1}^{m} \frac{\mathrm{vol}\big(\mathcal{X}_{(a_k\xi_k,\, b_{k-1}\xi_{k-1}]}\big)}{\xi_k^d} \leq C\left(\frac{\mathrm{vol}(\mathcal{X}_{\xi_m})}{\xi_m^d} + \sum_{k=1}^{m} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_k,\,\xi_{k-1}]}\big)}{\xi_k^d} \right)\ .$$

Proof. Letting $V := \mathrm{vol}(\mathcal{X}_{\xi_m})/\xi_m^d + \sum_{k=1}^m \mathrm{vol}\big(\mathcal{X}_{(\xi_k,\,\xi_{k-1}]}\big)/\xi_k^d$, we have

$$
\begin{aligned}
&\frac{\mathrm{vol}\big(\mathcal{X}_{b_m\xi_m}\big)}{\xi_m^d} + \sum_{k=1}^m \frac{\mathrm{vol}\big(\mathcal{X}_{(a_k\xi_k,\,b_{k-1}\xi_{k-1}]}\big)}{\xi_k^d} \\
&= \frac{\mathrm{vol}\big(\mathcal{X}_{\xi_m}\big)}{\xi_m^d} + \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_m,\,b_m\xi_m]}\big)}{\xi_m^d} \\
&\quad + \sum_{k=1}^m \frac{\mathrm{vol}\big(\mathcal{X}_{(a_k\xi_k,\,\xi_k]}\big)}{\xi_k^d} + \sum_{k=1}^m \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_k,\,\xi_{k-1}]}\big)}{\xi_k^d} + \sum_{k=1}^m \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{k-1},\,b_{k-1}\xi_{k-1}]}\big)}{\xi_k^d} \\
&\le V + \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_m,\,\xi_{j_m}]}\big)}{\xi_m^d} + \sum_{k=1}^m \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{i_k},\,\xi_k]}\big)}{\xi_k^d} + \sum_{k=1}^m \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{k-1},\,\xi_{j_{k-1}}]}\big)}{\xi_k^d} \\
&=: V + (\mathrm{I}) + (\mathrm{II}) + (\mathrm{III})\ .
\end{aligned}
$$

We upper bound separately the addends (II) and (I) + (III). We have

$$
\begin{aligned}
(\mathrm{II}) &= \sum_{k=1}^m \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{i_k},\,\xi_k]}\big)}{\xi_k^d} = \sum_{k=2}^{m+1} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{i_{k-1}},\,\xi_{k-1}]}\big)}{\xi_{k-1}^d} = \sum_{\substack{k=2 \\ i_{k-1}\ge k}}^{m+1} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{i_{k-1}},\,\xi_{k-1}]}\big)}{\xi_{k-1}^d} \\
&= \sum_{\substack{k=2 \\ i_{k-1}\ge k}}^{m+1} \sum_{i=k}^{i_{k-1}} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_i,\,\xi_{i-1}]}\big)}{\xi_{k-1}^d} \le \sum_{\substack{k=2 \\ i_{k-1}\ge k}}^{m+1} \sum_{i=k}^{i_{k-1}} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_i,\,\xi_{i-1}]}\big)}{\xi_i^d \mathbb{I}_{i\neq m+1} + \xi_{i-1}^d \mathbb{I}_{i=m+1}} \\
&= n_{a,m+1} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{m+1},\,\xi_m]}\big)}{\xi_m^d} + \sum_{h=1}^m n_{a,h} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_h,\,\xi_{h-1}]}\big)}{\xi_h^d} \\
&\le n_{a,m+1} \frac{\mathrm{vol}\big(\mathcal{X}_{\xi_m}\big)}{\xi_m^d} + \sum_{h=1}^m n_{a,h} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_h,\,\xi_{h-1}]}\big)}{\xi_h^d}\ .
\end{aligned}
$$

Next, we have

$$
\begin{aligned}
\text{(I)} + \text{(III)} &= \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_m\,,\,\xi_{j_m}]}\big)}{\xi_m^d} + \sum_{k=1}^{m} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{k-1}\,,\,\xi_{j_{k-1}}]}\big)}{\xi_k^d} \\
&= \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_m\,,\,\xi_{j_m}]}\big)}{\xi_m^d} \mathbb{I}_{j_m \le m-1} + \sum_{\substack{k=1 \\ j_{k-1} \le k-2}}^{m} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{k-1}\,,\,\xi_{j_{k-1}}]}\big)}{\xi_k^d} \\
&= \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_m\,,\,\xi_{j_m}]}\big)}{\xi_m^d} \mathbb{I}_{j_m \le m-1} + \sum_{\substack{k=2 \\ j_{k-1} \le k-2}}^{m} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_{k-1}\,,\,\xi_{j_{k-1}}]}\big)}{\xi_k^d} \\
&= \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_m\,,\,\xi_{j_m}]}\big)}{\xi_m^d} \mathbb{I}_{j_m \le m-1} + \sum_{\substack{k=1 \\ j_k \le k-1}}^{m-1} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_k\,,\,\xi_{j_k}]}\big)}{\xi_{k+1}^d} \\
&= \sum_{\substack{j=j_m+1 \\ j_m \le m-1}}^{m} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_m^d} + \sum_{\substack{k=1 \\ j_k \le k-1}}^{m-1} \sum_{j=j_k+1}^{k} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_{k+1}^d} \\
&= \sum_{\substack{j=j_m+1 \\ j_m \le m-1}}^{m} \frac{\xi_j^d}{\xi_m^d} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_j^d} + \sum_{\substack{k=1 \\ j_k \le k-1}}^{m-1} \sum_{j=j_k+1}^{k} \frac{\xi_j^d}{\xi_{k+1}^d} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_j^d} \\
&\le \sum_{\substack{j=j_m+1 \\ j_m \le m-1}}^{m} \frac{\xi_{j_m+1}^d}{\xi_m^d} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_j^d} + \sum_{\substack{k=1 \\ j_k \le k-1}}^{m-1} \sum_{j=j_k+1}^{k} \frac{\xi_{j_k+1}^d}{\xi_{k+1}^d} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_j^d} \\
&= \sum_{\substack{k=1 \\ j_k \le k-1}}^{m} \sum_{j=j_k+1}^{k} c_k \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_j^d} \le c \sum_{\substack{k=1 \\ j_k \le k-1}}^{m} \sum_{j=j_k+1}^{k} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_j\,,\,\xi_{j-1}]}\big)}{\xi_j^d} \\
&= c \sum_{h=1}^{m} n_{b,h} \frac{\mathrm{vol}\big(\mathcal{X}_{(\xi_h\,,\,\xi_{h-1}]}\big)}{\xi_h^d} \,.
\end{aligned}
$$

Putting everything together and noting that $c \ge 1$ gives the result. □

E The DOO algorithm: definition and proofs

E.1 The DOO algorithm

Let us present the (non-certified and certified) DOO algorithm and its underlying assumptions. The non-certified algorithm and the assumptions will essentially be the same as in the original reference Munos [2011], with minor adjustments to our framework, for ease of exposition. The notion of certificate is not present in Munos [2011].

The algorithm is defined for a fixed $K \in \mathbb{N}^\star$, by an infinite sequence of subsets of $\mathcal{X}$ (cells) of the form $(X_{h,i})_{h\in\mathbb{N},i=0,\dots,K^h-1}$. For $h \in \mathbb{N}$, the sets $X_{h,0}, \dots, X_{h,K^h-1}$ are non-empty, pairwise disjoint, and their union contains $\mathcal{X}$. Furthermore, $(X_{h,i})_{h\in\mathbb{N},i=0,\dots,K^h-1}$ is associated with a K-ary tree, meaning that for any $h \in \mathbb{N}$ and $j \in \{0, \dots, K^h - 1\}$, there exist K distinct $i_1, \dots, i_K \in \{0, \dots, K^{h+1} - 1\}$ such that $X_{h+1,i_1}, \dots, X_{h+1,i_K}$ form a partition of $X_{h,j}$. We call $(h+1, i_1), \dots, (h+1, i_K)$ the children of (h, j).

For each cell $X_{h,i}$, $h \in \mathbb{N}, i = 0, \dots, K^h - 1$, there is a representative $\boldsymbol{x}_{h,i} \in X_{h,i}$, which can be thought of, e.g., as the center of the cell. We assume that feasible cells have feasible representatives, that is: if $X_{h,i} \cap \mathcal{X} \neq \varnothing$, then $\boldsymbol{x}_{h,i} \in \mathcal{X}$.

The sequence of cells and representatives is well-behaved in the sense of the following two assumptions.

Assumption 4. *There are fixed* $0 < \delta < 1$ *and* $R < +\infty$ *such that for any* $h \in \mathbb{N}$, $i = 0, \dots, K^h - 1$ *and* $u, v \in X_{h,i}$,

$$||u - v|| \leq R\delta^h.$$

Assumption 5. *There is a fixed* $\nu > 0$ *such that, with* δ *as in Assumption 4, for any* $h \in \mathbb{N}$, $i = 0, \dots, K^h - 1$, $h' \in \mathbb{N}$, $i' = 0, \dots, K^{h'} - 1$, *with* $(h, i) \neq (h', i')$,

$$||\boldsymbol{x}_{h,i} - \boldsymbol{x}_{h',i'}|| \geq \nu\delta^{\max(h,h')}.$$

Assumption 4 is very classical. Note that Assumption 5, which is key for our improved analysis, is slightly stronger than in Munos [2011], yet easy to satisfy. For a compact $\mathcal{X}$, a sequence $(X_{h,i})_{h\in\mathbb{N},i=0,\dots,K^h-1}$ satisfying Assumptions 4 and 5 indeed exists. For instance, when $\mathcal{X}$ is the unit hypercube $[0,1]^d$ and $||.||$ is the supremum norm $||.||_\infty$, we may use bisections with $K = 2^d$, letting $X_{h,i}$ be an hypercube of edge length 2^{-h} and $\boldsymbol{x}_{h,i}$ be its center, for $h \in \mathbb{N}$ and $i = 0, \dots, 2^{dh} - 1$. In this case we have $R = 1$, $\delta = 1/2$ and $\nu = 1/2$ in Assumptions 4 and 5.

The DOO algorithm is defined in Algorithm 3. For both non-certified and certified versions, the most promising cell among a set of active cells $\mathcal{L}_n$ is selected at every iteration k (see (16)) before being split into its K children. For the certified version, the algorithm declares its recommendations $\boldsymbol{x}_n^\star$ to be ε-optimal (i.e., $\gamma_n = 1$) whenever the condition (17) is met.

Algorithm 3: DOO (non-certified and certified versions)

input: input set $\mathcal{X}$, Lipschitz bound L, cells $(X_{h,i})_{h\in\mathbb{N},i=0,\dots,K^h-1}$, representatives $(\boldsymbol{x}_{h,i})_{h\in\mathbb{N},i=0,\dots,K^h-1}$.
initialization: let $n = 1$. Let $\boldsymbol{x}_1^\star = \boldsymbol{x}_1 = \boldsymbol{x}_{0,0}$. Let $\mathcal{L}_1 = \{(0,0)\}$. Evaluate $f(\boldsymbol{x}_1)$.
certified version only: Let $\gamma = 0$.
for *iteration* $k = 1, 2, \dots$ **do**
let (breaking ties arbitrarily)

$$(h^\star, i^\star) \in \operatorname*{argmax}_{(h,i)\in\mathcal{L}_n} \left\{ f(\boldsymbol{x}_{h,i}) + LR\delta^h \right\} . \tag{16}$$

certified version only: If

$$f(\boldsymbol{x}_{h^\star,i^\star}) + LR\delta^{h^\star} \leq \max(f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)) + \varepsilon, \tag{17}$$

then let $\gamma \leftarrow 1$ and $\gamma_n \leftarrow \gamma$.
let $\mathcal{L}_+$ be the set of the K children of $(h^\star, i^\star)$.
for $(h^\star + 1, j) \in \mathcal{L}_+$ **do**
if $X_{h^\star+1,j} \cap \mathcal{X} \neq \varnothing$ then, let $n \leftarrow n + 1$, let $\boldsymbol{x}_n = \boldsymbol{x}_{h^\star+1,j}$, evaluate $f(\boldsymbol{x}_n)$, let $\boldsymbol{x}_n^\star \in \operatorname{argmax}_{\boldsymbol{x}\in\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_n\}} f(\boldsymbol{x})$ and let $\mathcal{L}_n = \mathcal{L}_{n-1} \cup \{(h^\star + 1, j)\}$.
certified version only: let $\gamma_n = \gamma$.
remove $(h^\star, i^\star)$ from $\mathcal{L}_n$.

E.2 Proof of Theorem 1

We start by proving Theorem 1, and will then comment on where our analysis improves over that of Munos [2011, Theorem 1], as well as on a straightforward extension.

Proof. We first recall the guarantee (19) below, which is classical (e.g., Munos 2011). By induction, it is straightforward to show that the union of the cells in $\mathcal{L}_n$ contains $\mathcal{X}$ at all steps $n \in \mathbb{N}^\star$. Therefore, the global maximizer $\boldsymbol{x}^\star$ (for which the inequality in Assumption 1 holds) belongs to a cell $X_{\bar{h},\bar{i}}$ with $(\bar{h}, \bar{i}) \in \mathcal{L}_n$. Consider now the cell $X_{h^\star,i^\star}$ in

(16), at step n. We have, using first (16), and then Assumptions 1 and 4,

$$\begin{aligned} f(\boldsymbol{x}_{h^\star,i^\star}) + LR\delta^{h^\star} &\geq f(\boldsymbol{x}_{\bar{h},\bar{i}}) + LR\delta^{\bar{h}} \\ &\geq f(\boldsymbol{x}^\star) - LR\delta^{\bar{h}} + LR\delta^{\bar{h}} \\ &= f(\boldsymbol{x}^\star) \; . \end{aligned} \tag{18}$$

This implies that, for $(h^\star, i^\star)$ given by (16),

$$f(\boldsymbol{x}_{h^\star,i^\star}) \in \mathcal{X}_{LR\delta^{h^\star}} . \tag{19}$$

We now proceed in a slightly different way than in the proof of Munos [2011, Theorem 1]. Consider the first time at which the DOO algorithm reaches step (16) with $f(\boldsymbol{x}_{h^\star,i^\star}) \geq f(\boldsymbol{x}^\star) - \varepsilon$. Then let I_ε be the number of times the DOO algorithm went through step (16) strictly before that time, and denote by n_ε the total number of evaluations of f strictly before that same time. Then we have

$$n_\varepsilon \leq 1 + KI_\varepsilon.$$

Furthermore, after n_ε evaluations of f, we have, by definitions of the recommendation $\boldsymbol{x}^\star_{n_\varepsilon}$ and n_ε,

$$f(\boldsymbol{x}^\star_{n_\varepsilon}) = \max_{\boldsymbol{x}\in\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_{n_\varepsilon}\}} f(\boldsymbol{x}) \geq \max_{\boldsymbol{x}\in\mathcal{L}_{n_\varepsilon}} f(\boldsymbol{x}) \geq f(\boldsymbol{x}_{h^\star,i^\star}) \geq f(\boldsymbol{x}^\star) - \varepsilon \; .$$

Since the optimization error $f(\boldsymbol{x}^\star) - f(\boldsymbol{x}^\star_n)$ can only decrease for $n \geq n_\varepsilon$, the last inequality above entails that the non-certified sample complexity of DOO (non-certified version) is bounded by n_ε and thus

$$\zeta(\text{non-certified DOO}, f, \varepsilon) \leq 1 + KI_\varepsilon. \tag{20}$$

Consider now the sequence $(h^\star_1, i^\star_1), \ldots, (h^\star_{I_\varepsilon}, i^\star_{I_\varepsilon})$ corresponding to the first I_ε times the DOO algorithm A went through step (16). Let $\mathcal{E}_\varepsilon$ be the corresponding finite set $\{\boldsymbol{x}_{h^\star_1,i^\star_1}, \ldots, \boldsymbol{x}_{h^\star_{I_\varepsilon},i^\star_{I_\varepsilon}}\}$. By definition of I_ε, we have $\mathcal{E}_\varepsilon \subseteq \mathcal{X}_{(\varepsilon,\varepsilon_0]}$. Since $\varepsilon = \varepsilon_{m_\varepsilon} \leq \varepsilon_{m_\varepsilon - 1} \leq \ldots \leq \varepsilon_0$, we have $\mathcal{E}_\varepsilon \subseteq \bigcup_{i=1}^{m_\varepsilon} \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}$, so that the cardinality I_ε of $\mathcal{E}_\varepsilon$ (a leaf can never be selected twice) satisfies

$$I_\varepsilon = \mathrm{card}(\mathcal{E}_\varepsilon) \leq \sum_{i=1}^{m_\varepsilon} \mathrm{card}\left(\mathcal{E}_\varepsilon \cap \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}\right) \; . \tag{21}$$

Let $N_{\varepsilon,i}$ be the cardinality of $\mathcal{E}_\varepsilon \cap \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}$. For $\boldsymbol{x}_{h,j} \in \mathcal{E}_\varepsilon \cap \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}$, from (19), we have $\varepsilon_i < LR\delta^h$ and thus

$$\delta^h > \frac{\varepsilon_i}{LR} \; .$$

Consider two distinct $\boldsymbol{x}_{h,j}, \boldsymbol{x}_{h',j'} \in \mathcal{E}_\varepsilon \cap \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}$. Then, from Assumption 5, we obtain

$$||\boldsymbol{x}_{h,j} - \boldsymbol{x}_{h',j'}|| \geq \nu\delta^{\max(h,h')} > \frac{\nu\varepsilon_i}{LR} \; .$$

Hence, by definition of packing numbers, we have

$$N_{\varepsilon,i} \leq \mathcal{N}\left(\mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}, \frac{\nu\varepsilon_i}{LR}\right).$$

Using now Lemma 14, we obtain

$$N_{\varepsilon,i} \leq \left(\mathbf{1}_{\nu/R\geq 1} + \mathbf{1}_{\nu/R<1}\left(\frac{4R}{\nu}\right)^d\right)\mathcal{N}\left(\mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}, \frac{\varepsilon_i}{L}\right). \tag{22}$$

Combining the last inequality with (20) and (21) concludes the proof. □

Remark 17. *The analysis of the DOO algorithm in Munos [2011, Theorem 1] yields a bound on the non-certified sample complexity than can be expressed in the form* $1 + C \sum_{k=1}^{m_\varepsilon} \mathcal{N}\left(\mathcal{X}_{\varepsilon_{k-1}}, \frac{\varepsilon_k}{L}\right)$*, with a constant* C*. The corresponding proof relies on two main arguments. First, when a cell of the form* $(h^\star, i^\star)$*,* $i^\star \in \{0, \dots, K^{h^\star} - 1\}$*, is selected in* (16)*, then the corresponding cell representative* $\boldsymbol{x}_{i^\star,h^\star}$ *is* $LR\delta^{h^\star}$*-optimal (we also use this argument). Second, as a consequence, for a given fixed value of* $h^\star$*, for the sequence of values of* $i^\star$ *that are selected in* (16)*, the corresponding cell representatives* $\boldsymbol{x}_{h^\star,i^\star}$ *form a packing of* $\mathcal{X}_{[0,LR\delta^{h^\star}]}$*.*

Our analysis in the proof of Theorem 1 stems from the observation that using a packing of $\mathcal{X}_{LR\delta^{h^\star}}$ *yields a suboptimal analysis, since the cell representatives* $\boldsymbol{x}_{h^\star,i^\star}$ *can be much better than* $LR\delta^{h^\star}$*-optimal. Hence, we proceed differently from Munos [2011], by first partitioning all the selected cell representatives (in* (16)*) according to their level of optimality as in* (21) *and then by exhibiting packings of the different layers of input points* $\mathcal{X}_{(\varepsilon,\varepsilon_{m_\varepsilon-1}]}, \mathcal{X}_{(\varepsilon_{m_\varepsilon-1},\varepsilon_{m_\varepsilon-2}]}, \dots, \mathcal{X}_{(\varepsilon_1,\varepsilon_0]}$*. In a word, we partition the values of* f *instead of partitioning the input space when counting the representatives selected at all levels.*

Remark 18. *The bound of Theorem 1, based on* (3)*, is built by partitioning* $(\varepsilon, \varepsilon_0]$ *into the* m_ε *sets*

$$(\varepsilon, \varepsilon_{m_\varepsilon-1}], (\varepsilon_{m_\varepsilon-1}, \varepsilon_{m_\varepsilon-2}], \dots, (\varepsilon_1, \varepsilon_0]$$

whose lengths are sequentially doubled (except from $(\varepsilon_{m_\varepsilon}, \varepsilon_{m_\varepsilon-1}]$ *to* $(\varepsilon_{m_\varepsilon-1}, \varepsilon_{m_\varepsilon-2}]$*). As can be seen from the proof of Theorem 1, more general bounds could be obtained, based on more general partitions of* $(\varepsilon, \varepsilon_0]$*. The benefits of the present partition are the following. First, it considers sets whose upper values are no more than twice the lower values, which controls the magnitude of their corresponding packing numbers in* (3) *(at scale the lower values). Second the number of sets in the partition is logarithmic in* ε *which controls the sum in* (3)*.*

The same generalization could be applied to the bound based on (4) *of Theorem 2, for the certified version of the DOO algorithm. In this latter case, another benefit of choosing the partition* $(\varepsilon, \varepsilon_{m_\varepsilon-1}], (\varepsilon_{m_\varepsilon-1}, \varepsilon_{m_\varepsilon-2}], \dots, (\varepsilon_1, \varepsilon_0]$ *(together with the additional set* $[0, \varepsilon]$*) is that the upper bound is then tight up to a logarithmic factor for most functions* f*, as proved in Section 3.*

E.3 Proof of Theorem 2

Let us first show that Algorithm 3 (certified version) is indeed a certified algorithm. Let f be any function satisfying Assumption 1. For notational simplicity, we set[9]

$$n := \sigma(\text{certified DOO}, f, \varepsilon) = \inf\{i \in \mathbb{N}^* : \gamma_i = 1\}$$

and we show that $f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_n^\star) \leq \varepsilon$. After exactly n evaluations of f, when Algorithm 3 reaches step (16), the condition (17) holds for the first time. From (18) in the proof of Theorem 1, which applies here since the non-certified and certified versions select the same leaves $(h^\star, i^\star)$ and output the same queries $\boldsymbol{x}_m$ and recommendations $\boldsymbol{x}_m^\star$, we know that $f(\boldsymbol{x}^\star) \leq f(\boldsymbol{x}_{h^\star,i^\star}) + LR\delta^{h^\star}$. Since condition (17) also guarantees that

$$f(\boldsymbol{x}_{h^\star,i^\star}) + LR\delta^{h^\star} \leq \max(f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)) + \varepsilon = f(\boldsymbol{x}_n^\star) + \varepsilon$$

this entails that $f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_n^\star) \leq \varepsilon$. Since the optimization error $n' \mapsto f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{n'}^\star)$ can only decrease over time, the requirement $\gamma_{n'} = 1 \Rightarrow f(\boldsymbol{x}^\star) - f(\boldsymbol{x}_{n'}^\star) \leq \varepsilon$ is true for all $n' \geq n$ and thus all $n \in \mathbb{N}^*$. This proves that Algorithm 3 (certified version) is a certified algorithm.

We now show the upper bound on $\sigma(\text{certified DOO}, f, \varepsilon)$. Let I_ε be the number of times the algorithm went through step (16) strictly before the first iteration k where γ_n is set to 1, that is, strictly before (17) holds for the first time. Note that $\sigma(\text{certified DOO}, f, \varepsilon)$ is the total number of evaluations of f before (17) holds for the first time, so that

$$\sigma(\text{certified DOO}, f, \varepsilon) \leq 1 + KI_\varepsilon \ . \tag{23}$$

[9]Note that, in the definition of Algorithm 3, the variable γ_n is sometimes assigned twice, with first the value of 0 and then the value of 1. In that case, we consider that $\gamma_n = 1$.

Consider now the sequence $(h_1^\star, i_1^\star), \dots, (h_{I_\varepsilon}^\star, i_{I_\varepsilon}^\star)$ corresponding to the first I_ε times the DOO algorithm went through step (16). Let $\mathcal{E}_\varepsilon$ be the corresponding finite set $\{\boldsymbol{x}_{h_1^\star, i_1^\star}, \dots, \boldsymbol{x}_{h_{I_\varepsilon}^\star, i_{I_\varepsilon}^\star}\}$. Of course we have $\mathcal{E}_\varepsilon \subset \mathcal{X}_\varepsilon \bigcup \left(\bigcup_{i=1}^{m_\varepsilon} \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}\right)$, so that

$$I_\varepsilon = \mathrm{card}(\mathcal{E}_\varepsilon) \leq \mathrm{card}\left(\mathcal{E}_\varepsilon \cap \mathcal{X}_\varepsilon\right) + \sum_{i=1}^{m_\varepsilon} \mathrm{card}\left(\mathcal{E}_\varepsilon \cap \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}\right) \ . \tag{24}$$

Let $N_{\varepsilon, m_\varepsilon+1}$ be the cardinality of $\mathcal{E}_\varepsilon \cap \mathcal{X}_\varepsilon$. For $i = 1, \dots, m_\varepsilon$, let $N_{\varepsilon,i}$ be the cardinality of $\mathcal{E}_\varepsilon \cap \mathcal{X}_{(\varepsilon_i,\, \varepsilon_{i-1}]}$. With exactly the same arguments as in the proof of Theorem 1, we show, for $i = 1, \dots, m_\varepsilon$, that (22) holds.

Let now $\boldsymbol{x}_{h_l^\star, i_l^\star} \in \mathcal{E}_\varepsilon \cap \mathcal{X}_\varepsilon$, with $\ell \in \{1, \dots I_\varepsilon\}$. The pair $(h_l^\star, i_l^\star)$ was selected when the algorithm went through step (16) for the ℓ-th time. By definition of I_ε, (17) does not hold at this time, which implies, with m being the number of evaluations of f at this time,

$$f(\boldsymbol{x}_{h_\ell^\star, i_\ell^\star}) + LR\delta^{h_\ell^\star} > \max(f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_m)) + \varepsilon \geq f(\boldsymbol{x}_{h_\ell^\star, i_\ell^\star}) + \varepsilon \ .$$

This implies that $LR\delta^{h_\ell^\star} > \varepsilon$ and thus

$$\delta^{h_\ell^\star} > \frac{\varepsilon}{LR} \ .$$

Now consider two distinct $\boldsymbol{x}_{h,j}, \boldsymbol{x}_{h',j'} \in \mathcal{E}_\varepsilon \cap \mathcal{X}_\varepsilon$. Then, from Assumption 5, we obtain

$$||\boldsymbol{x}_{h,j} - \boldsymbol{x}_{h',j'}|| \geq \nu\delta^{\max(h,h')} > \frac{\nu\varepsilon}{LR} \ .$$

Hence, we have

$$N_{\varepsilon, m_\varepsilon+1} \leq \mathcal{N}\left(\mathcal{X}_\varepsilon, \frac{\nu\varepsilon}{LR}\right).$$

Using now Lemma 14, we obtain

$$N_{\varepsilon, m_\varepsilon+1} \leq \left(\mathbf{1}_{\nu/R \geq 1} + \mathbf{1}_{\nu/R<1}\left(\frac{4R}{\nu}\right)^d\right) \mathcal{N}\left(\mathcal{X}_\varepsilon, \frac{\varepsilon}{L}\right).$$

Combining (23) and (24) with (22) and the last inequality concludes the proof.