The sample complexity of level set approximation

François Bachoc, Tommaso R Cesari, Sébastien Gerchinovitz

The sample complexity of level set approximation

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

October 23, 2020

Abstract

We study the problem of approximating the level set of an unknown function by sequentially querying its values. We introduce a family of algorithms called Bisect and Approximate through which we reduce the level set approximation problem to a local function approximation problem. We then show how this approach leads to rate-optimal sample complexity guarantees for Hölder functions, and we investigate how such rates improve when additional smoothness or other structural assumptions hold true.

1 INTRODUCTION

Let $f\colon [0,1]^d \to \mathbb{R}$ be any function. For $a \in \mathbb{R}$, we consider the problem of finding the level set

$$\{f = a\} \stackrel{\text{def}}{=} \big\{\boldsymbol{x} \in [0,1]^d : f(\boldsymbol{x}) = a\big\} \,.$$

Setting: Sequential Black-Box Evaluation. 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 (Online Protocol 1). At every round $n \geq 1$, the query point $\boldsymbol{x}_n$ can be chosen as a deterministic function of the values $f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_{n-1})$ observed so far. At the end of round n, the learner outputs a subset S_n of $[0,1]^d$ with the goal of approximating the level set $\{f = a\}$.

Online Protocol 1: Deterministic Scheme

for $n = 1, 2, \ldots$ **do**
  pick the next query point $\boldsymbol{x}_n \in [0,1]^d$
  observe the value $f(\boldsymbol{x}_n)$
  output an approximating set $S_n \subset [0,1]^d$

The problem of identifying the level set $\{f = a\}$ of a black-box function arises often in practice. In particular, this problem is closely related to excursion set estimation (also called failure domain estimation), where the goal is to estimate $\{f \geq a\}$.[1] Level-set identification and failure domain estimation are relevant to the field of computer experiments and uncertainty quantification, where $f(x)$ provides the output of a complex computer model for some input parameter x (Sacks et al., 1989; Santner et al., 2003). Typical fields of applications are nuclear engineering (Chevalier et al., 2014), coastal flooding

[1]As it will become apparent later, our techniques for estimating level sets can be adapted for sub/superlevel set approximation straightforwardly, whilst retaining the same sample complexity guarantees (see Footnote 4).

(Azzimonti et al., 2020) and network systems (Ranjan et al., 2008). Level set identification is also relevant when $f(x)$ corresponds to natural data (Rahimi et al., 2004; Galland et al., 2004). In many such situations, f is so complex that it is considered black-box.

Learning Goal. There exist several ways to compare the estimators S_n and the level set $\{f = a\}$. A first possibility is to use metrics or pseudometrics $\rho(A, B)$ between sets $A, B \subset [0,1]^d$, such as the Hausdorff distance or the volume of the symmetric difference (e.g., Tsybakov 1997). However a small value of $\rho(S_n, \{f = a\})$ does not imply that S_n contains the whole set $\{f = a\}$, nor—in the case of the volume of the symmetric difference—that $f(x) \approx a$ for all $\boldsymbol{x} \in S_n$. In practice, we might fail to identify *all* critical states of a given system, or raise unnecessary false alarms.

In this paper, we therefore consider an alternative (new) way of quantifying our performance. For any accuracy $\varepsilon > 0$, denote by $\{|f - a| \leq \varepsilon\} \stackrel{\text{def}}{=} \{\boldsymbol{x} \in [0,1]^d : |f(\boldsymbol{x}) - a| \leq \varepsilon\}$ the *inflated* level set at scale ε. We will focus on algorithms whose outputs S_n are ε-approximations of $\{f = a\}$, as defined below.

Definition 1 (ε-approximation of a level set)**.** We say that a set $S \subset [0,1]^d$ is an *ε-approximation* of the level set $\{f = a\}$ if and only if it contains $\{f = a\}$ while consisting only of points at which f is at most ε-away from a, i.e.,

$$\{f = a\} \subset S \subset \{|f - a| \leq \varepsilon\} . \tag{1}$$

The main mathematical problem we address is that of determining the *sample complexity* of level set approximation, that is, the minimum number of evaluations of f after which S_n is an ε-approximation of $\{f = a\}$ (see Section A of the Supplementary Material for a formal definition). We are interested in algorithms with rate-optimal worst-case sample complexity over classical function classes, as well as (slightly) improved sample complexity bounds in more favorable cases.

Main Contributions and Outline of the Paper.

- We define a new learning goal for level set approximation (see above and Section A of the Supplementary Material) similar in spirit to that of Gotovos et al. (2013).
- In Section 2 we briefly discuss the inherent hardness of the level set approximation problem (Theorem 1) and the role played by smoothness or structural assumptions on f.
- In Section 3 we design a family of algorithms called *Bisect and Approximate* through which we reduce the level set approximation problem to a local function approximation problem.
- In Sections 4 and 5 we instantiate Bisect and Approximate to the cases of Hölder or gradient-Hölder functions. We derive upper and lower bounds showing that the sample complexity for level set approximation is of the order of $1/\varepsilon^{d/\beta}$ in the worst-case, where $\beta \in (0, 2]$ is a smoothness parameter.
- In Section 5.2 we also show that Bisect and Approximate algorithms adapt to more favorable functions f by featuring a slightly improved sample complexity in such cases.

Some lemmas and proofs are deferred to the Supplementary Material.

Related Works. Sequential learning (sometimes referred to as sequential design of experiments) for level set and sublevel set identification is an active field of research. Many algorithms are based on Gaussian process priors over the black box function f (Ranjan et al., 2008; Vazquez and Bect, 2009; Picheny et al., 2010; Bect et al., 2012; Chevalier et al., 2014; Ginsbourger et al., 2014; Wang et al., 2016; Bect et al., 2017; Gotovos et al., 2013). In contrast with this large number of algorithms, few theoretical guarantees exist on the consistency or rate of convergence. Moreover, the majority of these guarantees are probabilistic. This means that consistency results state that an error goes to zero almost surely with respect to the Gaussian process prior measure over the unknown function f, and that the rates of convergence hold in probability, with respect to the same prior measure. In this probabilistic setting, Bect et al. (2019) provide a consistency result for a class of methods called Stepwise Uncertainty Reduction. Gotovos et al. (2013) provide rates of convergence, with noisy observations and for a classification-based loss function.

The loss function of Gotovos et al. (2013), given for sublevel set estimation, is similar in spirit to the notion of ε-approximation studied here for level set approximation, since we both aim at making decisions that are approximately correct for all x in the input space. The main difference is that Gotovos

et al. (2013) assume that f is a realization of a Gaussian process and thus provide guarantees that are probabilistic, while we prove deterministic bounds (for a fixed function). On the other hand, they consider noisy observations, while we assume f can be evaluated perfectly.

A related problem studied in statistics is density level set estimation, in which the superlevel set of a density f is estimated by looking at i.i.d. draws of random variables with density f. For this problem, several different performance measures are considered, such as the Hausdorff distance (Cadre et al., 2013; Singh et al., 2009; Tsybakov, 1997) or a measure of the symmetric difference (Cadre, 2006; Rigollet and Vert, 2009; Tsybakov, 1997).

When the function f is convex, our problem is also related to that of approximating a convex compact body with a simpler set (e.g., a polytope) in Hausdorff distance. This has been studied extensively in convex geometry and several sequential and non-sequential algorithms have been proposed (see, e.g., the two surveys Kamenev 2019; Gruber 1993 and references therein).

The closest connections with our work are within the bandit optimization literature. More precisely, our Bisect and Approximate algorithm and its analysis are inspired from the branch-and-bound algorithm of Locatelli and Carpentier (2018, Appendix A.2) and from the earlier methods of Perevozchikov (1990), Bubeck et al. (2011, HOO algorithm), and Munos et al. (2014, DOO algorithm). All these algorithms address the problem of finding a global extremum of f, while we are interested in finding level sets. However the idea of using a 2^d-ary tree to construct refined partitions of the input domain, and sequential methods to select which branch to explore next, are key in this paper.

There are also algorithmic connections with the nonparametric statistics literature. In particular, the idea of locally approximating a target function has been used many times for different purposes (e.g., Györfi et al. 2002; Tsybakov 2009).

Additional Notation. We denote the set of all positive integers $\{1, 2, \ldots\}$ by $\mathbb{N}^*$. For all $x \in \mathbb{R}$, we denote by $\lceil x \rceil$ (resp., $\lfloor x \rfloor$) the ceiling (resp., floor) function at x, i.e., the smallest (resp., largest) integer larger (resp., smaller) or equal to x.

2 INHERENT HARDNESS

In this section we show that level sets are typically $(d-1)$-dimensional, and discuss the consequences of this fact in terms of the inherent hardness of the level set approximation problem.

We evaluate the dimension through the growth rate of packing numbers, one of the classical ways to measure the size of a set. In the case of the unit hypercube and the sup-norm, recall that packing numbers are defined as follows.

Definition 2 (Packing number)**.** For all $r > 0$, the *r-packing number* $\mathcal{N}(E, r)$ of a subset E of $[0,1]^d$ (with respect to the sup-norm $\|\cdot\|_\infty$) is the largest number of r-separated points contained in E, i.e.,

$$\mathcal{N}(E, r) := \sup\big\{k \in \mathbb{N}^* : \exists\, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in E,\, \min_{i \neq j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_\infty > r\big\} \quad (2)$$

if E is nonempty, zero otherwise.

The next theorem indicates that, with the exceptions of sets of minimizers or maximizers, ε-packing numbers of level sets $\{f = a\}$ of continuous functions f are at least $(d-1)$-dimensional. This result is very natural since $\{f = a\}$ is the solution set of one equation with d unknowns.

Theorem 1. *Let $f\colon [0,1]^d \to \mathbb{R}$ be a non-constant continuous function, and $a \in \mathbb{R}$ be any level such that $\min_{\boldsymbol{x} \in [0,1]^d} f(\boldsymbol{x}) < a < \max_{\boldsymbol{x} \in [0,1]^d} f(\boldsymbol{x})$. Then, there exists $\kappa > 0$ such that, for all $\varepsilon > 0$,*

$$\mathcal{N}\big(\{f = a\},\, \varepsilon\big) \geq \kappa \frac{1}{\varepsilon^{d-1}}\,.$$

We restate and prove this result in the Supplementary Material (Theorem 6, Section F.1).

We note an important difference with the global optimization problem. Indeed, the set of global maximizers (or minimizers) of a function f is typically finite and thus 0-dimensional. This implies that, depending on the shape of f around a global optimum, global optimization algorithms feature a sample complexity ranging roughly between $\log(1/\varepsilon)$ and $(1/\varepsilon)^d$ (see, e.g., Perevozchikov 1990; Munos et al. 2014).

In our case, by Theorem 1, level sets are large, so that we can expect the sample complexity to depend heavily on the input dimension d. This is however not the end of the story. Indeed, as in nonparametric statistics (e.g., Györfi et al. 2002; Tsybakov 2009)

or in convex optimization (e.g., Nesterov 2004; Boyd and Vandenberghe 2004; Bubeck 2015), additional smoothness or structural assumptions like convexity of f play a role in the hardness of the level set approximation problem. Since this problem is important in practice, designing algorithms that best exploit such additional assumptions is an important question. This is what we address in this paper.

3 BA ALGORITHMS & ANALYSIS

In this section, we introduce and analyze a family of algorithms designed for the problem of approximating the level set of an unknown function. They are based on an iterative refinement of the domain $[0,1]^d$, as made precise in the following definition.

Definition 3 (Bisection of a family of hypercubes)**.** Let $\mathcal{C}$ be a family of n hypercubes included in $[0,1]^d$. We say that bisect($\mathcal{C}$) is the *bisection* of $\mathcal{C}$ if it contains exactly the $2^d\,n$ hypercubes obtained by subdividing each $C = [a_1, b_1] \times \cdots [a_d, b_d] \in \mathcal{C}$ into the 2^d equally-sized smaller hypercubes of the form $C' = I_1 \times \cdots \times I_d$ with I_j being either $\big[a_j, (a_j+b_j)/2\big]$ or $\big[(a_j+b_j)/2, b_j\big]$.

Our algorithm is of the branch-and-bound type, similarly to other bandit algorithms for global optimization such as that of Locatelli and Carpentier (2018, Appendix A.2) and earlier methods (Perevozchikov, 1990; Bubeck et al., 2011; Munos et al., 2014).

Our Bisect and Approximate algorithms[2] (BA, Algorithm 2) maintain, at all iterations i, a collection $\mathcal{C}_i$ of hypercubes on which the target function f is determined to take values close to the target level a. A BA algorithm takes as input the level a, a common number of queries k (to be performed in each hypercube at all iterations), and a pair of tolerance parameters $b, \beta > 0$, related to the smoothness of f and the approximation power of the approximators used by the algorithm. At the beginning of each iteration i, the collection of hypercubes $\mathcal{C}_{i-1}$ determined at the end of the last iteration is bisected (line 3), so that all new hypercubes have diameter 2^{-i} (in the sup-norm). Then, the values of the target function f at k points of each newly created hypercube are queried (lines 7–8). The output set S_n after n queries to f is a subset of the union of all hypercubes in $\mathcal{C}_{i-1}$, i.e., the collection of all hypercubes determined during to the latest *completed* iteration. The precise definition of S_n depends on the approximators used during the last completed iteration and the two tolerance parameters b, β (lines 9 and 2).[3] After all k values of f are queried from a hypercube C', this information is used to determine a local approximator $g_{C'}$ of f (line 10). Finally, the collection of hypercubes $\mathcal{C}_i$ is updated using $g_{C'}$ as a proxy for f (line 11) for all hypercubes C'. In this step, all hypercubes C' in which the proxy $g_{C'}$ is too far from the target level a are discarded, where the tightness of the rejection rule increases with the passing of the iterations i and it is further regulated by the two tolerance parameters b, β.

Algorithm 2: Bisect and Approximate (BA)

input: level $a \in \mathbb{R}$, queries $k \in \mathbb{N}^*$, tol. $b, \beta > 0$
init: $D \leftarrow [0,1]^d$, $\mathcal{C}_0 \leftarrow \{D\}$, $g_D \equiv a$, $n \leftarrow 0$
1 **for** *iteration* $i = 1, 2, \ldots$ **do**
2 $S(i) \leftarrow \bigcup_{C \in \mathcal{C}_{i-1}} \big\{\boldsymbol{x} \in C : \big|g_C(\boldsymbol{x}) - a\big| \le b\,2^{-\beta(i-1)}\big\}$
3 $\mathcal{C}'_i \leftarrow \text{bisect}(\mathcal{C}_{i-1})$
4 **for each** *hypercube* $C' \in \mathcal{C}'_i$ **do**
5 **for** $j = 1, \ldots, k$ **do**
6 update $n \leftarrow n + 1$
7 pick a query point $\boldsymbol{x}_n \in C'$
8 observe $f(\boldsymbol{x}_n)$
9 output $S_n \leftarrow S(i)$
10 pick a local approximator $g_{C'} : C' \to \mathbb{R}$
11 $\mathcal{C}_i \leftarrow \big\{C' \in \mathcal{C}'_i : \exists \boldsymbol{x} \in C', \big|g_{C'}(\boldsymbol{x}) - a\big| \le b\,2^{-\beta i}\big\}$

Our analysis of BA algorithms (Theorem 2) hinges on the accuracy of the approximators $g_{C'}$ selected at line 10, as formalized in the following definition.

[2]We refer to Bisect and Approximate algorithms in the plural form because different BA algorithms can be defined with the same input, depending on which rules are used to pick points at line 7 and approximators at line 10. E.g., BAH (Section 4) only looks at the center of each hypercube and uses constant approximators, while BAG (Section 5) queries the value of f at all vertices of each hypercube and builds higher-order polynomial approximators.

[3]Notably, the output set $S(i)$ at iteration i can be represented succinctly and testing if $\boldsymbol{x} \in S(i)$ can be done efficiently in all our BA instances in Sections 4, 5.

Definition 4 (Accurate approximation)**.** Let $b, \beta > 0$ and $C \subset [0,1]^d$. We say that a function $g\colon C \to \mathbb{R}$ is a (b,β)-*accurate approximation* of another function $f\colon [0,1]^d \to \mathbb{R}$ (on C) if the distance (in the sup-norm on C) between f and g can be controlled with the diameter (in the sup-norm) of C as

$$\sup_{\boldsymbol{x}\in C} \big|g(\boldsymbol{x}) - f(\boldsymbol{x})\big| \le b \Big(\sup_{\boldsymbol{x},\boldsymbol{y}\in C} \|\boldsymbol{x}-\boldsymbol{y}\|_\infty \Big)^\beta .$$

We now present one of our main results, which states that BA algorithms run with accurate approximations of the target function return ε-approximations of the target level set after a number of queries that depends on the packing number (Definition 2) of the inflated level set at decreasing scales.[4]

Theorem 2. *Consider a Bisect and Approximate algorithm (Algorithm 2) run with input a, k, b, β. Let $f\colon [0,1]^d \to \mathbb{R}$ be an arbitrary function with level set $\{f = a\} \neq \varnothing$. Assume that the approximators $g_{C'}$ selected at line 10 are (b,β)-accurate approximations of f (Definition 4).*

Fix any accuracy $\varepsilon > 0$, let $i(\varepsilon) := \lceil (1/\beta)\log_2(2b/\varepsilon) \rceil$, and define $n(\varepsilon)$ by[5]

$$4^d\, k \sum_{i=0}^{i(\varepsilon)-1} \lim_{\delta\to 1^-} \mathcal{N}\Big(\big\{ |f-a| \le 2\, b\, 2^{-\beta i} \big\},\ \delta\, 2^{-i} \Big) . \quad (3)$$

Then, for all $n > n(\varepsilon)$, the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$.

The expression (3) can be simplified by taking $\delta = 1$ and increasing the leading multiplicative constant. However, in the following sections we will see how to upper bound this quantity with simpler functions of $1/\varepsilon$, for which the limit can be computed exactly.

Proof. Fix any $n > n(\varepsilon)$. We begin by proving that

$$\{f = a\} \subset S_n . \quad (4)$$

Recall that $S_n = \bigcup_{C\in\mathcal{C}_{\iota-1}} \big\{ \boldsymbol{x} \in C : \big|g_C(\boldsymbol{x}) - a\big| \le b\, 2^{-\beta(\iota-1)} \big\}$ (line 9), where $\iota = \iota(n)$ is the iteration during which the n-th value of f is queried. To prove (4), we will show the stronger result: for all $i \ge 0$,

$$\{f = a\} \subset \bigcup_{C\in\mathcal{C}_i} \Big\{ \boldsymbol{x} \in C : \big|g_C(\boldsymbol{x}) - a\big| \le b\, 2^{-\beta i} \Big\}, \quad (5)$$

i.e., that the level set $\{f = a\}$ is *always* included in the output set, not only after iteration $\iota(n) - 1$ has been completed. We do so by induction. If $i = 0$, then $\{f = a\} \subset [0,1]^d = \big\{ \boldsymbol{x} \in [0,1]^d : |a - a| \le b\, 2^{-\beta\cdot 0} \big\}$, which is the union in (5) by definition of $D = [0,1]^d$, $\mathcal{C}_0 = \big\{D\big\}$ and $g_D \equiv a$ in the initialization of Algorithm 2. Assume now that the inclusion holds for some $i - 1$: we will show that it keeps holding for the next iteration $i \in \mathbb{N}^*$. Indeed, fix any $\boldsymbol{z} \in \{f = a\}$. By induction, $\boldsymbol{z}$ belongs to some hypercube $C \in \mathcal{C}_{i-1}$. Since $\mathcal{C}'_i = \text{bisect}(\mathcal{C}_{i-1})$ (line 3), by definition of bisection (Definition 3) we have that $\bigcup_{C'\in\mathcal{C}'_i} C' = \bigcup_{C\in\mathcal{C}_{i-1}} C$, which in turns implies that there exists a hypercube $C'_{\boldsymbol{z}} \in \mathcal{C}'_i$ such that $\boldsymbol{z} \in C'_{\boldsymbol{z}}$. We show now that this $C'_{\boldsymbol{z}}$ also belongs to $\mathcal{C}_i$, i.e., that it is not discarded during the update of the algorithm at line 11. Indeed, since $g_{C'_{\boldsymbol{z}}}$ is a (b,β)-accurate approximation of f on $C'_{\boldsymbol{z}}$ (by assumption) and the diameter (in the sup-norm) of $C'_{\boldsymbol{z}}$ is $\sup_{\boldsymbol{x},\boldsymbol{y}\in C'_{\boldsymbol{z}}} \|\boldsymbol{x}-\boldsymbol{y}\|_\infty = 2^{-i}$, we have that $\big|g_{C'_{\boldsymbol{z}}}(\boldsymbol{z}) - a\big| = \big|g_{C'_{\boldsymbol{z}}}(\boldsymbol{z}) - f(\boldsymbol{z})\big| \le b\, 2^{-\beta i}$. This gives both that $\boldsymbol{z} \in C'_{\boldsymbol{z}} \in \mathcal{C}_i$ (by definition of $\mathcal{C}_i$ at line 11) and, consequently, that $\boldsymbol{z} \in \bigcup_{C\in\mathcal{C}_i} \big\{ \boldsymbol{x} \in C : \big|g_C(\boldsymbol{x}) - a\big| \le b\, 2^{-\beta i} \big\}$, which clinches the proof of (5) and in turn yields (4).

We now show the validity of the second inclusion

$$S_n \subset \big\{ |f - a| \le \varepsilon \big\} . \quad (6)$$

As above, let $\iota = \iota(n)$ be the iteration during which the n-th value of f is queried by the algorithm. Fix any $\boldsymbol{z} \in S_n$. We will prove that $\boldsymbol{z} \in \big\{ |f - a| \le \varepsilon \big\}$ or, restated equivalently, that $\big|f(\boldsymbol{z}) - a\big| \le \varepsilon$. By definition of $S_n = \bigcup_{C\in\mathcal{C}_{\iota-1}} \big\{ \boldsymbol{x} \in C : \big|g_C(\boldsymbol{x}) - a\big| \le b\, 2^{-\beta(\iota-1)} \big\}$ (line 9), since $\boldsymbol{z} \in S_n$, then there exists $C_{\boldsymbol{z}} \in \mathcal{C}_{\iota-1}$ such that $\boldsymbol{z} \in C_{\boldsymbol{z}}$ and $\big|g_{C_{\boldsymbol{z}}}(\boldsymbol{z}) - a\big| \le b\, 2^{-\beta(\iota-1)}$. Moreover, since $C_{\boldsymbol{z}} \in \mathcal{C}_{\iota-1} \subset \mathcal{C}'_{\iota-1}$ has diameter $\sup_{\boldsymbol{x},\boldsymbol{y}\in C_{\boldsymbol{z}}} \|\boldsymbol{x}-\boldsymbol{y}\|_\infty = 2^{-(\iota-1)}$ (in the sup-norm) and the approximator $g_{C_{\boldsymbol{z}}}$ is a (b,β)-accurate

[4]Note that our BA algorithm can be used for estimating a sublevel set $\{f \le a\}$ by simply dropping the absolute values in lines 2 and 11 of Algorithm 2. As the reader might realize, the same proof techniques would apply with the corresponding (straightforward) changes. An analogous argument applies to superlevel sets $\{f \ge a\}$.

[5]Letting $\sum_{i=0}^{-m} a_i = 0$ for any $m > 0$ and all $a_j \in \mathbb{R}$.

approximation of f on $C_{\boldsymbol{z}}$ (by assumption), we have that $\big|f(\boldsymbol{z}) - g_{C_{\boldsymbol{z}}}(\boldsymbol{z})\big| \leq b\, 2^{-\beta(\iota-1)}$. Thus

$$\big|f(\boldsymbol{z})-a\big| \leq \big|f(\boldsymbol{z})-g_{C_{\boldsymbol{z}}}(\boldsymbol{z})\big| + \big|g_{C_{\boldsymbol{z}}}(\boldsymbol{z})-a\big| \leq 2b2^{-\beta(\iota-1)}$$

and the right-hand side would be smaller than ε — proving (6)— if either $\varepsilon \geq 2b$ (trivially), or in case $\varepsilon \in (0, 2b)$, if we could guarantee that the iteration $\iota = \iota(n)$ during which the n-th value of f is queried satisfies $\iota - 1 \geq \lceil (1/\beta) \log_2(2b/\varepsilon) \rceil = i(\varepsilon)$. In other words, assuming without loss of generality that $\varepsilon \in (0, 2b)$ (so that $i(\varepsilon) \geq 1$) and recalling that $n > n(\varepsilon)$, in order to prove (6) we only need to check that the $i(\varepsilon)$-th iteration is guaranteed to be concluded after at most $n(\varepsilon)$ queries, where $n(\varepsilon)$ is defined in terms of packing numbers in (3). To see this, note that the total number of values of f that the algorithm queries by the end of iteration $i(\varepsilon)$ is $\sum_{i=1}^{i(\varepsilon)} k\, |\mathcal{C}'_i| = 2^d\, k\, \sum_{i=1}^{i(\varepsilon)} |\mathcal{C}_{i-1}| = 2^d\, k\, \sum_{i=0}^{i(\varepsilon)-1} |\mathcal{C}_i|$. To conclude the proof, it is now sufficient to show that for all iterations $i \geq 0$, the number of hypercubes maintained by the algorithm can be upper bounded by

$$|\mathcal{C}_i| \leq 2^d \lim_{\delta \to 1^-} \mathcal{N}\Big(\big\{|f-a| \leq 2\, b\, 2^{-\beta i}\big\},\ \delta\, 2^{-i}\Big)\ . \quad (7)$$

Fix an arbitrary $\delta \in (0,1)$. If $i = 0$, then $|\mathcal{C}_0| = 1 \leq \mathcal{N}\big(\big\{|f-a| \leq 2\, b\big\}, \delta\big)$ by the definitions of $\mathcal{C}_0 = \big\{[0,1]^d\big\}$ (initialization of Algorithm 2) and δ-packing number (Definition 2) of $\big\{|f-a| \leq 2\, b\big\}$ (which is non-empty because it contains $\{f = a\}$). Fix any iteration $i \in \mathbb{N}^*$. By definition of $\mathcal{C}_i$ (line 11), for all hypercubes $C \in \mathcal{C}_i$ there exists a point $\boldsymbol{x}_C \in C$ such that $|g_C(\boldsymbol{x}_C) - a| \leq b\, 2^{-\beta i}$. Hence, for each hypercube $C \in \mathcal{C}_i$ there exists one of its points $\boldsymbol{x}_C \in C$ such that $|f(\boldsymbol{x}_C) - a|$ can be upper bounded by

$$|f(\boldsymbol{x}_C) - g_C(\boldsymbol{x}_C)| + |g_C(\boldsymbol{x}_C) - a| \leq 2\, b\, 2^{-\beta i}\ , \quad (8)$$

where, recalling that all hypercubes in $\mathcal{C}_i$ have diameter 2^{-i} (in the sup-norm), the bound on the term $|f(\boldsymbol{x}_C) - g_C(\boldsymbol{x}_C)|$ is a consequence of g_C being a (b, β)-accurate approximation of f on C.

Now we claim that the family of hypercubes $\mathcal{C}_i$ can be partitioned into 2^d subfamilies $\mathcal{C}_i(1), \ldots, \mathcal{C}_i(2^d)$ with the property that all distinct hypercubes $C \neq C'$ belonging to the same family $C_i(k)$ are strictly $(\delta\, 2^{-i})$-separated (in the sup-norm), i.e., that for all $k \in \{1, \ldots, 2^d\}$ and all $C, C' \in \mathcal{C}_i(k)$, $C \neq C'$, we have $\inf_{\boldsymbol{x} \in C, \boldsymbol{y} \in C'} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty > \delta\, 2^{-i}$.

We defer the proof of this claim to Section C of the Supplementary Material (for an insightful picture, see Figure 1 in the same section). Assume for now that it is true and fix an arbitrary $k \in \{1, \ldots, 2^d\}$. Then, for all $C \in \mathcal{C}_i(k)$, there exists $\boldsymbol{x}_C$ such that (8) holds. Therefore, we determined the existence of $\big|\mathcal{C}_i(k)\big|$-many $\big(\delta\, 2^{-i}\big)$-separated points that are all included in $\big\{|f-a| \leq 2\, b\, 2^{-\beta i}\big\}$. By definition of $\big(\delta\, 2^{-i}\big)$-packing number of $\big\{|f-a| \leq 2\, b\, 2^{-\beta i}\big\}$ (i.e., the *largest* cardinality of a set of $\big(\delta\, 2^{-i}\big)$-separated points included in $\big\{|f-a| \leq 2\, b\, 2^{-\beta i}\big\}$ —Definition 2), this implies that $\big|\mathcal{C}_i(k)\big| \leq \mathcal{N}\big(\big\{|f-a| \leq 2\, b\, 2^{-\beta i}\big\},\ \delta\, 2^{-i}\big)$. Recalling that $\mathcal{C}_i(1), \ldots, \mathcal{C}_i(2^d)$ is a partition of C_i, we then obtain

$$\big|\mathcal{C}_i\big| = \sum_{k=1}^{2^d} \big|\mathcal{C}_i(k)\big| \leq 2^d \mathcal{N}\Big(\big\{|f-a| \leq 2b2^{-\beta i}\big\}, \delta 2^{-i}\Big)$$

which, after taking the infimum over $\delta \in (0,1)$ and by the monotonicity of the packing number $r \mapsto \mathcal{N}(E, r)$ (for any $E \subset [0,1]^d$), yields

$$\begin{aligned}\big|\mathcal{C}_i\big| &\leq \inf_{\delta \in (0,1)} \Big(2^d \mathcal{N}\Big(\big\{|f-a| \leq 2b2^{-\beta i}\big\}, \delta 2^{-i}\Big)\Big) \\ &= 2^d \lim_{\delta \to 1^-} \mathcal{N}\Big(\big\{|f-a| \leq 2\, b\, 2^{-\beta i}\big\},\ \delta\, 2^{-i}\Big)\ .\end{aligned}$$

This gives (7) and concludes the proof. □

By looking at the end of the proof of the previous result, one could see that the exponential term 4^d in our bound (3) could be lowered to 2^d under the assumption that at any iteration i, Algorithm 2 picks at least one $\boldsymbol{x}_{C'} \in C'$ for each $C' \in \mathcal{C}'_i$ such that $\big\|\boldsymbol{x}_{C'_1} - \boldsymbol{x}_{C'_2}\big\|_\infty \geq 2^{-i}$ for all distinct $C'_1, C'_2 \in \mathcal{C}'_i$. Notably this property is enjoyed by all our BA instances in Sections 4, 5.

4 HÖLDER FUNCTIONS

In this section, we focus on Hölder functions, and we present a BA instance that is rate-optimal for determining their level sets.

Definition 5 (Hölder function)**.** Let $c > 0$, $\gamma \in (0, 1]$, and $E \subset [0,1]^d$. We say that a function $f\colon E \to \mathbb{R}$ is (c, γ)-*Hölder* (with respect to the sup-norm $\|\cdot\|_\infty$) if $\big|f(\boldsymbol{x}) - f(\boldsymbol{y})\big| \leq c\, \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^\gamma$, for all $\boldsymbol{x}, \boldsymbol{y} \in E$.

Our BA instance for Hölder functions (BAH, Algorithm 3) runs Algorithm 2 with $k = 1$, $b = c$, $\beta = \gamma$. The local approximators $g_{C'}$ are constant and equal to the value $f(\boldsymbol{c}_{C'})$ at the center $\boldsymbol{c}_{C'}$ of

C'. In particular, the output set S_n is now the entire union of all hypercubes determined in the latest completed iteration.

Algorithm 3: BA for Hölder Functions (BAH)

input: level $a \in \mathbb{R}$, tol. $c > 0$, $\gamma \in (0, 1]$
init: $D \leftarrow [0,1]^d$, $\mathcal{C}_0 \leftarrow \{D\}$, $n \leftarrow 0$
for *iteration* $i = 1, 2, \ldots$ **do**
  $S(i) \leftarrow \bigcup_{C \in \mathcal{C}_{i-1}} C$
  let $\mathcal{C}'_i \leftarrow \text{bisect}(\mathcal{C}_{i-1})$
  for each *hypercube* $C' \in \mathcal{C}'_i$ **do**
    update $n \leftarrow n + 1$
    pick the center $\boldsymbol{c}_{C'}$ of C' as the next $\boldsymbol{x}_n$
    observe $f(\boldsymbol{x}_n)$
    output $S_n \leftarrow S(i)$
  $\mathcal{C}_i \leftarrow \left\{C' \in \mathcal{C}'_i : \left|f(\boldsymbol{c}_{C'}) - a\right| \leq c\, 2^{-\gamma i}\right\}$

The next result shows that the optimal worst-case sample complexity of the level set approximation of Hölder functions is of order $1/\varepsilon^{d/\gamma}$, and it is attained by BAH (Algorithm 3).

Theorem 3. *Let $a \in \mathbb{R}, c > 0, \gamma \in (0, 1]$, and $f\colon [0,1]^d \to \mathbb{R}$ be any (c, γ)-Hölder function with level set $\{f = a\} \neq \varnothing$. Fix any accuracy $\varepsilon > 0$. Then, there exists $\kappa_1 > 0$ (independent of ε) such that, for all $n > \kappa_1/\varepsilon^{d/\gamma}$, the output S_n returned by BAH after the n-th query is an ε-approximation of $\{f = a\}$.*

Moreover, there exists κ_2 (independent of ε) such that no deterministic algorithm can guarantee to output an ε-approximation of the level set $\{f = a\}$ for all (c, γ)-Hölder functions f, querying less than $\kappa_2/\varepsilon^{d/\gamma}$ of their values.

The proof is deferred to Section D in the Supplementary Material. The upper bound is an application of Theorem 2. The lower bound is proven by showing that no algorithm can distinguish between the function $f \equiv 0$ and a function that is non-zero only on a small ball, on which it attains the value 2ε. We use a classical construction with bump functions that appears, e.g., in Theorem 3.2 of Györfi et al. (2002) for nonparametric regression lower bounds.

We remark that the rate in the previous result is the same as that of a naive uniform grid filling of the space with step-size of order $\varepsilon^{1/\gamma}$. While this rate cannot be improved in the worst case, the leading constant of our sequential algorithm may be better if a large fraction of the input space can be rejected quickly. More importantly, we will see in Section 5.2 that (slightly) better rates can be attained by BA algorithms if the inflated level sets of the target function f are smaller (as it happens, e.g., if f is convex with proper level set $\{f = a\}$).

In the following section, we will also investigate if and to what extent higher smoothness helps. To this end, we will switch our focus to differentiable functions with Hölder gradients.

5 ∇-HÖLDER FUNCTIONS: *BAG* ALGORITHM & ANALYSIS

In this section, we focus on differentiable functions with Hölder gradients, and we present a BA instance that is rate-optimal for determining their level sets.

Definition 6 (Gradient-Hölder/Lipschitz function)**.** Let $c_1 > 0$, $\gamma_1 \in (0, 1]$, and $E \subset [0,1]^d$. We say that a function $f\colon E \to \mathbb{R}$ is (c_1, γ_1)-*gradient-Hölder* (with respect to $\|\cdot\|_\infty$) if it is the restriction[6] (to E) of a continuously differentiable function defined on $\mathbb{R}^d$ such that $\left\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{y})\right\|_\infty \leq c_1 \left\|\boldsymbol{x} - \boldsymbol{y}\right\|_\infty^{\gamma_1}$ for all $\boldsymbol{x}, \boldsymbol{y} \in E$. If $\gamma_1 = 1$, we say that f is c_1-*gradient-Lipschitz*.

The next lemma introduces the polynomial approximators that will be used by our BA instance and it shows that they are $(c_1 d, 1 + \gamma_1)$-accurate approximations of f on all hypercubes.

Lemma 1 (BAG approximators)**.** *Let $f\colon C' \to \mathbb{R}$ be a (c_1, γ_1)-gradient-Hölder function, for some $c_1 > 0$ and $\gamma_1 \in (0, 1]$. Let $C' \subset [0,1]^d$ be a hypercube with diameter $\ell \in (0, 1]$ and set of vertices V', i.e., $C' = \prod_{j=1}^d [u_j, u_j + \ell]$, for some $\boldsymbol{u} := (u_1, \ldots, u_d) \in [0, 1-\ell]^d$, and $V' = \prod_{j=1}^d \{u_j, u_j + \ell\}$. The function*

$$h_{C'}\colon C' \to \mathbb{R}, \quad \boldsymbol{x} \mapsto \sum_{\boldsymbol{v} \in V'} f(\boldsymbol{v}) \prod_{j=1}^d p_{v_j}(x_j)\,, \tag{9}$$

where

$$p_{v_j}(x_j) := \left(1 - \frac{x_j - u_j}{\ell}\right) \mathbb{I}_{v_j = u_j} + \frac{x_j - u_j}{\ell}\, \mathbb{I}_{v_j = u_j + \ell},$$

[6] They are defined as restrictions of continuously differentiable functions in order to have simply and well-defined gradients on the boundary of their domains.

interpolates the 2^d *pairs* $\big\{(\boldsymbol{v}, f(\boldsymbol{v}))\big\}_{\boldsymbol{v}\in V'}$ *and it satisfies*

$$\sup_{\boldsymbol{x}\in C'} \big|h_{C'}(\boldsymbol{x}) - f(\boldsymbol{x})\big| \le c_1 d\, \ell^{1+\gamma_1} \ .$$

The technical proof of the previous lemma is deferred to Section E of the Supplementary Material.

Our Bisect and Approximate instance for gradient-Hölder functions (BAG) runs Algorithm 2 with $k = 2^d$, $b = c_1 d$, and $\beta = 1 + \gamma_1$. The local approximator $h_{C'}$ (defined in (9)) are computed by querying the values of f at all vertices of C'.

Note that line 12 of Algorithm 4 can be carried out efficiently since it is sufficient to check the condition on $\big|h_{C'}(\boldsymbol{x}) - a\big|$ at the vertices $\boldsymbol{x}$ of C'.[7] Also, note that the output set S_n is the union over hypercubes of pre-images of segments from the polynomial functions in (9).

Algorithm 4: BA for ∇-Hölder f (BAG)

input: level $a \in \mathbb{R}$, tol. $c_1 > 0$, $\gamma_1 \in (0, 1]$
init: $D \leftarrow [0,1]^d$, $\mathcal{C}_0 \leftarrow \{D\}$, $h_D \equiv a$, $n \leftarrow 0$
1 **for** *iteration* $i = 1, 2, \dots$ **do**
2 $S(i) \leftarrow \bigcup_{C\in\mathcal{C}_{i-1}} \big\{\boldsymbol{x} \in C : \big|h_C(\boldsymbol{x}) - a\big| \le c_1 d\, 2^{-(1+\gamma_1)(i-1)}\big\}$
3 $\mathcal{C}'_i \leftarrow \mathrm{bisect}(\mathcal{C}_{i-1})$
4 **for each** *hypercube* $C' \in \mathcal{C}'_i$ **do**
5 let $V' \subset C'$ be the set of vertices of C'
6 **for each** *vertex* $\boldsymbol{v} \in V'$ **do**
7 update $n \leftarrow n + 1$
8 pick vertex $\boldsymbol{v} \in V'$ as the next $\boldsymbol{x}_n$
9 observe $f(\boldsymbol{x}_n)$
10 output $S_n \leftarrow S(i)$
11 interpolate the 2^d pairs $\big\{(\boldsymbol{v}, f(\boldsymbol{v}))\big\}_{\boldsymbol{v}\in V'}$ with $h_{C'} : C' \to \mathbb{R}$ given by (9)
12 update $\mathcal{C}_i \leftarrow \big\{C' \in \mathcal{C}'_i : \text{there exists } \boldsymbol{x} \in C' \text{ such that } \big|h_{C'}(\boldsymbol{x}) - a\big| \le c_1 d\, 2^{-(1+\gamma_1)i}\big\}$

5.1 Worst-Case Sample Complexity

The next result shows that the optimal worst-case sample complexity of the level set approximation of gradient-Hölder functions is of order $1/\varepsilon^{d/(1+\gamma_1)}$, and it is attained by BAG (Algorithm 4).

Theorem 4. *Let* $a \in \mathbb{R}, c_1 > 0, \gamma_1 \in (0, 1]$*, and* $f : [0,1]^d \to \mathbb{R}$ *be any* (c_1, γ_1)*-gradient-Hölder function with level set* $\{f = a\} \neq \varnothing$*. Fix any accuracy* $\varepsilon > 0$*. Then, there exists* $\kappa_1 > 0$ *(independent of* ε*) such that, for all* $n > \kappa_1/\varepsilon^{d/(1+\gamma_1)}$*, the output* S_n *returned by BAG after the* n*-th query is an* ε*-approximation of* $\{f = a\}$*.*

Moreover, there exists $\kappa_2 > 0$ *(independent of* ε*) such that no deterministic algorithm can guarantee to output an* ε*-approximation of the level set* $\{f = a\}$ *for all* (c_1, γ_1)*-gradient-Hölder functions* f*, querying less than* $\kappa_2/\varepsilon^{d/(1+\gamma_1)}$ *of their values.*

The proof proceeds similarly to that of Theorem 3. It is deferred to Section E of the Supplementary Material.

Similarly to Section 4, the rate in the previous result could also be achieved by choosing query points on a regular grid with step-size of order $\varepsilon^{1/(1+\gamma_1)}$. However, our sequential algorithm features an improved sample complexity outside of a worst-case scenario, as shown in the following section.

5.2 Adaptivity to Smaller $d^\star$

Our general result (Theorem 2) suggests that the sample complexity can be controlled whenever there exists $d^\star \ge 0$ such that

$$\forall r \in (0,1), \quad \mathcal{N}\Big(\big\{|f - a| \le r\big\},\ r\Big) \le C^\star \left(\frac{1}{r}\right)^{d^\star}.$$

for some $C^\star > 0$. We call such a $d^\star$ a *NLS dimension* of $\{f = a\}$. Note that such a $d^\star$ always exists and $d^\star \le d$ by $\big\{|f - a| \le r\big\} \subset [0,1]^d$. However $d^\star \ge d - 1$ by Theorem 1 for non-degenerate level sets of continuous functions (for more details, see Section F.1 in the Supplementary Material). The definition of NLS dimension leads to the following result.

Corollary 1. *Let* $a \in \mathbb{R}, c_1 > 0, \gamma_1 \in (0, 1]$*, and* $f : [0,1]^d \to \mathbb{R}$ *be any* (c_1, γ_1)*-gradient-Hölder function with level set* $\{f = a\} \neq \varnothing$*. Let* $d^\star \in [d-1, d]$

[7] Indeed, only three cases can occur. We set $\rho = c_1 d\, 2^{-(1+\gamma_1)i}$. Case 1: if one of the vertices $\boldsymbol{x}$ satisfies $\big|h_{C'}(\boldsymbol{x}) - a\big| \le \rho$, then the condition is checked. Case 2: if the values $h_{C'}(\boldsymbol{x})$ at the vertices are all strictly below $a - \rho$ or all strictly above $a + \rho$, then it is also the case for all $\boldsymbol{x} \in C'$, since $h_{C'}(\boldsymbol{x})$ is a convex combination of all values at the vertices; so the condition is not checked. Case 3: if there are two vertices $\boldsymbol{x}$ and $\boldsymbol{y}$ such that $h_{C'}(\boldsymbol{x}) < a - \rho$ and $h_{C'}(\boldsymbol{y}) > a + \rho$, then there exists $\boldsymbol{z} \in C'$ such that $\big|h_{C'}(\boldsymbol{z}) - a\big| \le \rho$ by continuity of $h_{C'}$ on C'; so the condition is checked.

be a NLS dimension of $\{f = a\}$*. Fix any accuracy* $\varepsilon > 0$*. Then, for all* $n > m(\varepsilon)$*, the output* S_n *returned by BAG after the* n*-th query is an* ε*-approximation of* $\{f = a\}$*, where*

$$m(\varepsilon) := \begin{cases} \kappa_1 + \kappa_2 \log_2\left(\frac{1}{\varepsilon^{1/(1+\gamma_1)}}\right)^+ & \text{if } d^\star = 0 \,, \\ \kappa(d^\star)\frac{1}{\varepsilon^{d^\star/(1+\gamma_1)}} & \text{if } d^\star > 0 \,, \end{cases}$$

for $\kappa_1, \kappa_2, \kappa(d^\star) \geq 0$ *independent of* ε*, that depend exponentially on* d*, where* $x^+ = \max\{x, 0\}$*.*

We remark that $d^\star = d - 1$ can be achieved by well-behaved functions. This is typically the case when f is convex or (as a corollary) if it consists of finitely many convex components.[8] This non-trivial claim is proved in Section F.2 of the Supplementary Material for convex functions with a proper level set[9]. The following result, combined with this fact and Corollary 1, shows that BAG is rate-optimal for determining proper level sets of convex gradient-Lipschitz functions.

Theorem 5. *Fix any level* $a \in \mathbb{R}$ *and an arbitrary accuracy* $\varepsilon > 0$*. No deterministic algorithm* A *can guarantee to output an* ε*-approximation of any* Δ*-proper level set* $\{f = a\}$ *of an arbitrary convex* c_1*-gradient-Lipschitz functions* f *with* $c_1 \geq 3$ *and* $\Delta \in (0, 1/4]$*, querying less than* $\kappa/\varepsilon^{(d-1)/2}$ *of their values, where* $\kappa > 0$ *is a constant independent of* ε*.*

We give a complete proof of this result in Section F.3 of the Supplementary Material.

6 CONCLUSION

We studied the problem of determining ε-approximations of the level set of a target function f by only querying its values. After discussing the inherent hardness of the problem (Theorem 1), we designed the class of BA algorithms for which we proved theoretical guarantees under the assumption that accurate local approximations of f can be computed by only looking at its values (Theorem 2).

This provides a general method to reduce our level set approximation problem to a local approximation problem, decoupled from the original one.

Such an approach leads to rate-optimal worst-case sample complexity guarantees for the case of Hölder and gradient-Hölder functions (Theorems 3, 4). At the same time, we show that in some cases our BA algorithms adapt to a natural structural property of f, namely small NLS dimension (Corollary 1) including convexity (Theorem 5 and preceding discussion).

Future Work. Compared to the best achievable rate $1/\varepsilon^{d/\gamma}$ for (c, γ)-Hölder functions, we show that BA algorithms converge at a faster $1/\varepsilon^{d/(1+\gamma_1)}$ rate if f is (c_1, γ_1)-gradient-Hölder. This points at an interesting line of research: the study of general Hölder spaces in which the target function is k times continuously differentiable and the k-th partial derivatives are (c_k, γ_k)-Hölder, for some $k \in \mathbb{N}^*$, $c_k > 0$, and $\gamma_k \in (0, 1]$. We conjecture that a suitable choice of approximators for our BA algorithms would lead to a rate-optimal sample complexity of order $1/\varepsilon^{d/(k+\gamma_k)}$ for this class of functions, making optimal solutions for this problem sample-efficient. Another possible line of research is the design of algorithms that adapt to the smoothness of f when the latter is unknown, similarly to global bandit optimization (Grill et al., 2015; Bartlett et al., 2019). We leave these interesting directions open for future work.

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 (`https://www.deel.ai/`). This work benefited from the support of the project BOLD from the French national research agency (ANR).

[8] More precisely, if for some $a' > a$, we have that the sublevel $\{f \leq a'\}$ is a disjoint union of a finite number of convex sets on which f is convex.

[9] $\{f = a\}$ is Δ-proper for some $\Delta > 0$ if we have $\min_{\boldsymbol{x} \in [0,1]^d} f(\boldsymbol{x}) + \Delta \leq a \leq \min_{\boldsymbol{x} \in \partial[0,1]^d} f(\boldsymbol{x})$.

References

Dario Azzimonti, David Ginsbourger, Clément Chevalier, Julien Bect, and Yann Richet. Adaptive design of experiments for conservative estimation of excursion sets. *Technometrics*, pages 1–14, 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 *Algorithmic Learning Theory*, pages 184–206, 2019.

Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. *Statistics and Computing*, 22(3): 773–793, 2012.

Julien Bect, Ling Li, and Emmanuel Vazquez. Bayesian subset simulation. *SIAM/ASA Journal on Uncertainty Quantification*, 5(1):762–786, 2017.

Julien Bect, François Bachoc, David Ginsbourger, et al. A supermartingale approach to Gaussian process based sequential design of experiments. *Bernoulli*, 25(4A):2883–2919, 2019.

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

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge University Press, 2004.

Sébastien Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8(3-4):231–357, 2015. ISSN 1935-8237.

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

Benoît Cadre. Kernel estimation of density level sets. *Journal of multivariate analysis*, 97(4):999–1023, 2006.

Benoît Cadre, Bruno Pelletier, and Pierre Pudlo. Estimation of density level sets with a given probability content. *Journal of Nonparamentric Statistics*, 25(1):261–272, 2013.

Clément Chevalier, Julien Bect, David Ginsbourger, Emmanuel Vazquez, Victor Picheny, and Yann Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.

Frédéric Galland, Philippe Réfrégier, and Olivier Germain. Synthetic aperture radar oil spill segmentation by stochastic complexity minimization. *IEEE Geoscience and Remote Sensing Letters*, 1(4):295–299, 2004.

David Ginsbourger, Jean Baccou, Clément Chevalier, Frédéric Perales, Nicolas Garland, and Yann Monerie. Bayesian adaptive reconstruction of profile optima and optimizers. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):490–510, 2014.

Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1344–1350. AAAI Press, 2013.

Jean-Bastien Grill, Michal Valko, and Rémi Munos. Black-box optimization of noisy functions with unknown smoothness. In *Advances in Neural Information Processing Systems*, pages 667–675, 2015.

Peter Gruber. Aspects of approximation of convex bodies. In *Handbook of convex geometry*, pages 319–345. Elsevier, 1993.

László Györfi, Adam Krzyżak, Michael Kohler, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

George Kamenev. Optimal non-adaptive approximation of convex bodies by polytopes. In *Numerical Geometry, Grid Generation and Scientific Computing*, pages 157–172. Springer, 2019.

Andrea Locatelli and Alexandra Carpentier. Adaptivity to smoothness in X-armed bandits. In Sébastien Bubeck, Vianney Perchet, and Philippe Rigollet, editors, *Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 1463–1492. PMLR, 06–09 Jul 2018.

Rémi Munos et al. From bandits to Monte-Carlo Tree search: The optimistic principle applied to optimization and planning. *Foundations and Trends® in Machine Learning*, 7(1):1–129, 2014.

Yurii Nesterov. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied Optimization*. Springer US, 2004.

Alexander 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.

Victor Picheny, David Ginsbourger, Olivier Roustant, Raphael Haftka, and Nam-Ho Kim. Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, 132(7), 2010.

Mohammad Rahimi, Richard Pon, William Kaiser, Gaurav Sukhatme, Deborah Estrin, and Mani Srivastava. Adaptive sampling for environmental robotics. In *IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA'04. 2004*, volume 4, pages 3537–3544. IEEE, 2004.

Pritam Ranjan, Derek Bingham, and George Michailidis. Sequential experiment design for contour estimation from complex computer codes. *Technometrics*, 50(4):527–541, 2008.

Philippe Rigollet and Regis Vert. Optimal rates for plug-in estimators of density level sets. *Bernoulli*, 15 (4):1154–1178, 2009.

Jerome Sacks, William Welch, Toby Mitchell, and Henry Wynn. Design and analysis of computer experiments. *Statistical science*, pages 409–423, 1989.

Thomas Santner, Brian Williams, William Notz, and Brain Williams. *The design and analysis of computer experiments*, volume 1. Springer, 2003.

Aarti Singh, Clayton Scott, Robert Nowak, et al. Adaptive Hausdorff estimation of density level sets. *The Annals of Statistics*, 37(5B):2760–2782, 2009.

Alexandre B. Tsybakov. On nonparametric estimation of density level sets. *The Annals of Statistics*, 25(3): 948–969, 1997.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer, 2009.

Emmanuel Vazquez and Julien Bect. A sequential Bayesian algorithm to estimate a probability of failure. *IFAC Proceedings Volumes*, 42(10):546–550, 2009.

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

Hongqiao Wang, Guang Lin, and Jinglai Li. Gaussian process surrogates for failure detection: A Bayesian experimental design approach. *Journal of Computational Physics*, 313:247–259, 2016.

A SAMPLE COMPLEXITY: FORMAL DEFINITIONS

We provide formal definitions for the notions of deterministic algorithm, sample complexity, and rate-optimal algorithm.

We first precisely define deterministic algorithms that query values of functions sequentially and rely only on this information to build approximations of their level sets (sketched in Online Protocol 1). The behavior of any such algorithm is completely determined by a pair (φ, ψ), where $\varphi = (\varphi_n)_{n\in\mathbb{N}^*}$ is a sequence of functions $\varphi_n \colon \mathbb{R}^{n-1} \to [0,1]^d$ mapping the $n-1$ previously observed values $f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_{n-1})$ to the next query point $\boldsymbol{x}_n$, and $\psi = (\psi_n)_{n\in\mathbb{N}^*}$ is a sequence of functions $\psi_n \colon \mathbb{R}^n \to \big\{\text{subsets of } [0,1]^d\big\}$ mapping the n currently known values $f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)$ to an approximation S_n of the target level set.

We can now define the notion of sample complexity, which corresponds to the smallest number of queries after which the outputs S_n of an algorithm are all ε-approximations of the level set $\{f = a\}$ (recall Definition 1 in the Introduction).

Definition 7 (Sample complexity)**.** For all functions $f \colon [0,1]^d \to \mathbb{R}$, all levels $a \in \mathbb{R}$, any deterministic algorithm A, and any accuracy $\varepsilon > 0$, we denote by $\mathfrak{n}(f, A, \varepsilon, a)$ the smallest number of queries to f that A needs in order for its output sets S_n to be ε-approximations of the level set $\{f = a\}$ for all $n \geq \mathfrak{n}(f, A, \varepsilon, a)$, i.e.,

$$\mathfrak{n}(f, A, \varepsilon, a) := \inf\big\{n' \in \mathbb{N}^* : \forall n \geq n', S_n \text{ is an } \varepsilon\text{-approximation of } \{f = a\}\big\} \,. \tag{10}$$

We refer to $\mathfrak{n}(f, A, \varepsilon, a)$ as the *sample complexity* of A (for the ε-approximation of $\{f = a\}$).

We can now define rate-optimal algorithms rigorously. At a high-level, they output the tightest (up to constants) approximations of level sets that can possibly be achieved by deterministic algorithms.

Definition 8 (Rate-optimal algorithm)**.** For any level $a \in \mathbb{R}$ and some given family $\mathcal{F}$ of real-valued functions defined on $[0,1]^d$, we say that a deterministic algorithm A is *rate-optimal* (for level a and family $\mathcal{F}$) if, in the worst-case, it needs the same number of queries (up to constants) of the best deterministic algorithm in order to output approximations of level sets within any given accuracy, i.e., if there exists a constant $\kappa = \kappa(a, \mathcal{F}) \geq 1$, depending only on a and $\mathcal{F}$, such that, for all $\varepsilon > 0$,

$$\sup_{f\in\mathcal{F}} \mathfrak{n}(f, A, \varepsilon, a) \leq \kappa \inf_{A'\in\mathcal{A}} \sup_{f\in\mathcal{F}} \mathfrak{n}(f, A', \varepsilon, a) \,, \tag{11}$$

where $\mathcal{A}$ denotes the set of all deterministic algorithms.

B USEFUL INEQUALITIES ABOUT PACKING AND COVERING NUMBERS

For all $r > 0$, the *r-covering number* $\mathcal{M}(E, r)$ of a bounded subset E of $\mathbb{R}^d$ (with respect to the sup-norm $\|\cdot\|_\infty$) is the smallest cardinality of an r-covering of E, i.e.,

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

if E is nonempty, zero otherwise.

Covering numbers and packing numbers (2) are closely related. In particular, the following well-known inequalities hold—see, e.g., (Wainwright, 2019, Lemma 5.5 and Example 5.8, with permuted notation of $\mathcal{M}$ and $\mathcal{N}$).[10]

[10]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 $x_1, \dots, x_k$ of r-covers belong to A rather than just $\mathbb{R}^d$. Even if we do not need it for our analysis, Inequality (13) holds also in this stronger sense.

Lemma 2. *For any subset E of $[0,1]^d$ and any real number $r > 0$,*

$$\mathcal{N}(E, 2r) \leq \mathcal{M}(E, r) \leq \mathcal{N}(E, r) \ . \tag{12}$$

Furthermore, for all $\delta > 0$ and all $r > 0$, if $B(\delta) = \big\{\boldsymbol{x} \in \mathbb{R} : \|\boldsymbol{x}\|_\infty \leq \delta\big\}$,

$$\mathcal{M}\big(B(\delta), r\big) \leq \left(1 + 2\frac{\delta}{r}\mathbb{I}_{r<\delta}\right)^d \ . \tag{13}$$

We now state a known lemma about packing numbers at different scales.

Lemma 3. *For any subset E of $[0,1]^d$ and any real numbers $r_1, r_2 > 0$,*

$$\mathcal{N}(E, r_1) \leq \left(1 + 4\frac{r_2}{r_1}\mathbb{I}_{r_2>r_1}\right)^d \times \mathcal{N}(E, r_2) \ .$$

Proof. We can assume without loss of generality that E is nonempty and that $r_1 < r_2$. Then,

$$\begin{aligned}
\mathcal{N}(E, r_1) &\leq \mathcal{M}(E, r_1/2) && \text{(by (12))} \\
&\leq \mathcal{M}(E, r_2) \times \mathcal{M}\big(B(r_2), r_1/2\big) && \text{(see below)} \\
&\leq \mathcal{N}(E, r_2) \times \mathcal{M}\big(B(r_2), r_1/2\big) && \text{(by (12))} \\
&\leq \mathcal{N}(E, r_2) \times \left(1 + \frac{4r_2}{r_1}\right)^d \ . && \text{(by (13))}
\end{aligned}$$

The second inequality is obtained by building the $r_1/2$-covering of E in two steps. First, we cover E with balls of radius r_2. Second, we cover each ball of the first cover with balls of radius $r_1/2$. □

The next lemma upper bounds the packing number of the unit hypercube in the sup-norm, at all scales r.

Lemma 4. *For any positive real number $r > 0$, the r-packing number of the unit cube in the sup-norm satisfies*

$$\mathcal{N}\big([0,1]^d, r\big) \leq \left(\left\lfloor \frac{1}{r} \right\rfloor + 1\right)^d \ .$$

Proof. Since the diameter (in the sup-norm $\|\cdot\|_\infty$) of the unit hypercube is 1, if $r \geq 1$, then the packing number is $\mathcal{N}\big([0,1]^d, r\big) = 1 \leq \big(\lfloor 1/r \rfloor + 1\big)^d$. Consider now the case $r < 1$. Let $\rho := 1 - \lfloor 1/r \rfloor r \in [0, r)$ and G be the set of r-equispaced points $\big\{\rho/2,\ \rho/2 + r,\ \rho/2 + 2r,\ \ldots,\ \rho/2 + \lfloor 1/r \rfloor r\big\}^d$. Note that each point in $[0,1]^d$ is at most $(r/2)$-away from a point in G (in the sup-norm), i.e., G is an $(r/2)$-covering of $[0,1]^d$. We can thus use (12) in Lemma 2 at scale $r/2$ so see that $\mathcal{N}\big([0,1]^d, r\big) \leq \mathcal{M}\big([0,1]^d, r/2\big) \leq |G| = \big(\lfloor 1/r \rfloor + 1\big)^d$. □

The next lemma upper bounds the r-packing number (in the sup-norm) of an inflated level set at scale r.

Lemma 5. *For any function $f\colon [0,1]^d \to \mathbb{R}$ and all scales $r \in (0,1)$,*

$$\mathcal{N}\Big(\big\{|f - a| \leq r\big\},\ r\Big) \leq 2^d \left(\frac{1}{r}\right)^d \ .$$

Proof. Let $f\colon [0,1]^d \to \mathbb{R}$ be an arbitrary function and $r \in (0,1)$ any scale. By the monotonicity of the packing number ($E \subset F$ implies $\mathcal{N}(E, r) \leq \mathcal{N}(E, r)$ by definition of packing number —Definition 2) and the previous lemma (Lemma 4), we get

$$\mathcal{N}\Big(\big\{|f - a| \leq r\big\},\ r\Big) \leq \mathcal{N}\big([0,1]^d, r\big) \leq \left(\frac{1}{r} + 1\right)^d \leq \left(\frac{1}{r} + \frac{1}{r}\right)^d \leq 2^d \left(\frac{1}{r}\right)^d \ .$$

□

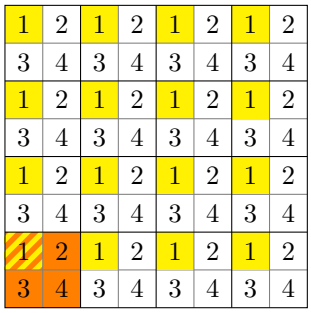

Figure 1: Constructing the partition when $d = 2$. In orange, the original enumeration A_0. In yellow, the family $\mathcal{C}_i(1)$.

C MISSING PROOFS OF SECTION 3

We now provide the missing proof of a claim we made in the proof of Theorem 2.

Claim 1. *Under the assumptions of Theorem 2, let $\delta \in (0,1)$ and $i \in \mathbb{N}^*$. Then, the family of hypercubes $\mathcal{C}_i$ maintained by Algorithm 2 can be partitioned into 2^d subfamilies $\mathcal{C}_i(1), \dots, \mathcal{C}_i(2^d)$ with the property that for all $k \in \{1, \dots, 2^d\}$ and all $C, C' \in \mathcal{C}_i(k)$, $C \neq C'$, we have $\inf_{\boldsymbol{x} \in C, \boldsymbol{y} \in C'} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty > \delta\, 2^{-i}$*

Proof. We build our partition by induction. For a two-dimensional picture, see Figure 1. Denote the elements of the standard basis of $\mathbb{R}^d$ by $\boldsymbol{e}_1, \dots, \boldsymbol{e}_d$. For any $\boldsymbol{x} \in [0,1]^d$ and all $E \subset [0,1]^d$, we denote by $E + \boldsymbol{x}$ the Minkowski sum $\{\boldsymbol{y} + \boldsymbol{x} : \boldsymbol{x} \in E\}$. Let E be the collection of all the hypercubes obtained by partitioning $[0,1]^d$ with a standard uniform grid with step size 2^{-i}, i.e., $E := \{[0, 2^{-i}]^d + \sum_{k=1}^d r_k\, 2^{-i}\, \boldsymbol{e}_k : r_1, \dots, r_k \in \{0, 1, \dots, 2^i - 1\}\}$.

Consider the family A_0 containing the hypercube $[0, 2^{-i}]$ and all other hypercubes of E adjacent to it; formally, $A_0 := \{[0, 2^{-i}]^d + \sum_{k=1}^d r_k\, 2^{-i}\, \boldsymbol{e}_k : r_1, \dots, r_d \in \{0, 1\}\}$. Assign to each of the 2^d hypercubes in A_0 a distinct number between 1 and 2^d. Fix any $k \in \{0, \dots, d-1\}$. For each hypercube $C \in A_k$, proceeding in the positive direction of the x_{k+1} axis, assign the same number as C to every other hypercube in E; formally, assign the same number as C to all hypercubes in $\{C + 2\, r\, 2^{-i}\, \boldsymbol{e}_{k+1} : r \in \{1, \dots 2^{i-1} - 1\}\}$. Denote by A_{k+1} the collection of all hypercubes that have been assigned a number so far. By construction, A_d coincides with the whole E and consists of 2^d distinct subfamilies of hypercubes, each containing only hypercubes that have been assigned the same number. For any number $k \in \{1, \dots, 2^d\}$, we denote by $\mathcal{C}_i(k)$ the subfamily of all hypercubes numbered with k. Fix any $k \in \{1, \dots, 2^d\}$. By construction, each $C \in \mathcal{C}_i(k)$ contains no adjacent hypercubes. Thus, the smallest distance between two distinct hypercubes $C, C' \in \mathcal{C}_i(k)$ is $\inf_{\boldsymbol{x} \in C, \boldsymbol{y} \in C'} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty \geq 2^{-i} > \delta\, 2^{-i}$ for all $\delta \in (0,1)$. ∎

D MISSING PROOFS OF SECTION 4

In this section, we prove Theorem 3 of Section 4. The proof is divided into two parts: one for the upper bound, one for the lower bound. Each time, we restate the corresponding result to ease readability.

D.1 Upper Bound

Proposition 1 (Theorem 3, upper bound)**.** *Consider the BAH algorithm run with input a, c, γ. Let $f\colon [0,1]^d \to \mathbb{R}$ be an arbitrary (c,γ)-Hölder function with level set $\{f = a\} \neq \varnothing$. Fix any accuracy $\varepsilon > 0$. Then, for all*

$$n > \kappa\, \frac{1}{\varepsilon^{d/\gamma}}\,, \qquad \text{where } \kappa := \left(2^{\gamma/d} 8^\gamma\, 2\, c\right)^{d/\gamma}\,,$$

the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$.

Proof. The proof is a simple application of Theorem 2, with $(b, \beta) = (c, \gamma)$. Since we are assuming that the level set $\{f = a\}$ is nonempty, we only need to check that for all iterations i and all hypercubes $C' \in \mathcal{C}'_i$, the constant approximator $g_{C'} \equiv f(\boldsymbol{c}_{C'})$ is a (c, γ)-accurate approximation of f on C'. For any iteration i and all hypercubes $C' \in \mathcal{C}'_i$, we have that

$$\sup_{\boldsymbol{x} \in C'} \big| g_{C'}(\boldsymbol{x}) - f(\boldsymbol{x}) \big| = \sup_{\boldsymbol{x} \in C'} \big| f(\boldsymbol{c}_{C'}) - f(\boldsymbol{x}) \big| \leq c\, 2^{-\gamma i} \,,$$

by definition of $g_{C'}$, the (c, γ)-Hölderness of f, and the fact that the diameter of all hypercubes $C' \in \mathcal{C}'_i$ (in the sup-norm) is 2^{-i}. Thus, Theorem 2 implies that for all $n > n(\varepsilon)$, the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$ where $n(\varepsilon)$ is

$$4^d \sum_{i=0}^{i(\varepsilon)-1} \lim_{\delta \to 1^-} \mathcal{N}\Big(\big\{ |f - a| \leq 2\, c\, 2^{-\gamma i} \big\},\ \delta\, 2^{-i} \Big) \tag{14}$$

and $i(\varepsilon) := \lceil (1/\gamma) \log_2(2c/\varepsilon) \rceil$. If $\varepsilon \geq 2c$, than the sum in (14) ranges from 0 to a *negative* value, thus $n(\varepsilon) = 0$ by definition of sum over an empty set and the result is true with $\kappa = 0$. Assume then that $\varepsilon < 2c$ so that the sum in (14) is not trivially zero. Upper-bounding, for any $\delta \in (0, 1)$ and all $i \geq 0$,

$$\mathcal{N}\Big(\big\{ |f - a| \leq 2\, c\, 2^{-\gamma i} \big\},\ \delta\, 2^{-i} \Big) \leq \mathcal{N}\big([0,1]^d,\ \delta\, 2^{-i} \big) \overset{(\dagger)}{\leq} \big(2^i/\delta + 1 \big)^d \leq (2/\delta)^d\, 2^{di}$$

(for completeness, we include a proof of the known upper bound (†) in Section B, Lemma 4) and recognizing the geometric sum below, we can conclude that

$$\begin{aligned} n(\varepsilon) &\leq 8^d \sum_{i=0}^{\lceil (1/\gamma) \log_2(2c/\varepsilon) \rceil - 1} \big(2^d \big)^i \\ &= 8^d\, \frac{2^{d \lceil (1/\gamma) \log_2(2c/\varepsilon) \rceil} - 1}{2^d - 1} \\ &\leq 8^d\, \frac{2^{d((1/\gamma) \log_2(2c/\varepsilon) + 1)}}{2^d - \big(2^d/2 \big)} = 2\, 8^d\, (2\, c)^{d/\gamma}\, \frac{1}{\varepsilon^{d/\gamma}} \,. \end{aligned}$$

□

D.2 Lower Bound

In this section, we prove our lower bound on the worst-case sample complexity of Hölder functions. We begin by stating a simple known lemma on *bump* functions. Bump functions are a standard tool to build lower bounds in nonparametric regression (see, e.g., (Györfi et al., 2002, Theorem 3.2), whose construction we also adapt for our following result and Proposition 4).

Lemma 6. *Fix any amplitude $\alpha > 0$, a step-size $\eta \in (0, 1/4]$, let $Z := \{0, 2\eta, \ldots, \lfloor 1/2\eta \rfloor 2\eta\}^d \subset [0,1]^d$, and fix an arbitrary $\boldsymbol{z} = (z_1, \ldots, z_d) \in Z$. Consider the* bump *functions*

$$\widetilde{f} \colon \mathbb{R} \to [0,1] \qquad\qquad f_{\alpha,\eta,\boldsymbol{z}} \colon \mathbb{R}^d \to \mathbb{R}$$

$$x \mapsto \widetilde{f}(x) := \begin{cases} \exp\left(\dfrac{-x^2}{1 - x^2} \right) & \text{if } x \in (-1, 1) \\ 0 & \text{otherwise} \,, \end{cases} \qquad\qquad \boldsymbol{x} \mapsto f_{\alpha,\eta,\boldsymbol{z}}(\boldsymbol{x}) := \alpha \prod_{j=1}^{d} \widetilde{f}\left(\frac{x_j - z_j}{\eta} \right) .$$

Then $\widetilde{f}$ is 3-Lipschitz and $f_{\alpha,\eta,\boldsymbol{z}}$ satisfies:

1. *$f_{\alpha,\eta,\boldsymbol{z}}$ is infinitely differentiable;*
2. *$f_{\alpha,\eta,\boldsymbol{z}}(\boldsymbol{x}) \in [0, \alpha)$ for all $\boldsymbol{x} \in \mathbb{R}^d \setminus \{\boldsymbol{z}\}$, and $f_{\alpha,\eta,\boldsymbol{z}}(\boldsymbol{z}) = \alpha$;*

3. $\{f_{\alpha,\eta,\boldsymbol{u}_1} > 0\} \cap \{f_{\alpha,\eta,\boldsymbol{u}_2} > 0\} = \varnothing$ *for any two distinct* $\boldsymbol{u}_1, \boldsymbol{u}_2 \in Z$*;*
4. $\|\boldsymbol{x} - \boldsymbol{y}\|_\infty \le 2\eta$ *for all* $\boldsymbol{x}, \boldsymbol{y}$ *in the closure* $\overline{\{f_{\alpha,\eta,\boldsymbol{z}} > 0\}}$ *of* $\{f_{\alpha,\eta,\boldsymbol{z}} > 0\}$ *and all* $\boldsymbol{z} \in Z$*.*

The proof is a straightforward verification and it is therefore omitted. We now prove our worst-case lower bound for Hölder functions.

Proposition 2 (Theorem 3, lower bound)**.** *Fix any level* $a \in \mathbb{R}$*, any two Hölder constants* $c > 0$*,* $\gamma \in (0, 1]$*, and an arbitrary accuracy* $\varepsilon \in \big(0,\, c/(3d2^\gamma)\big)$*. Let* $n < \kappa/\varepsilon^{d/\gamma}$ *be a positive integer, where* $\kappa := (c/12d)^{d/\gamma}$*. For each deterministic algorithm* A *there is a* (c,γ)*-Hölder function* f *such that, if* A *queries* n *values of* f*, then its output set* S_n *is not an* ε*-approximation of* $\{f = a\}$*. This implies in particular that (recall Definition 7),*

$$\inf_A \sup_f \mathfrak{n}(f, A, \varepsilon, a) \ge \kappa \, \frac{1}{\varepsilon^{d/\gamma}} \,,$$

where the inf *is over all deterministic algorithms* A *and the* sup *is over all* (c,γ)*-Hölder functions* f*.*

Note that the leading constant $\kappa = (c/12d)^{d/\gamma}$ in our lower bound decreases quickly with the dimension d. Though we keep our focus on sample complexity rates, there are ways to improve the multiplicative constants appearing in our lower bounds. For instance, in the proof below, a larger constant $\kappa := (1/4\,(c/2)^{1/\gamma})^d$ can be obtained by replacing bump functions with *spike* functions $\boldsymbol{x} \mapsto \big[2\,\varepsilon - c\,\|\boldsymbol{x} - \boldsymbol{z}\|_\infty^\gamma\big]^+$, where $x \mapsto [x]^+ := \max\{x, 0\}$ denotes the positive part of x. We choose to use bump functions instead because they are well-suited for any smoothness (e.g., in Section E.2, we will apply the same argument to gradient-Hölder functions).

Proof. The following construction is a standard way to prove lower bounds on sample complexity (for a similar example, see Györfi et al. 2002, Theorem 3.2). Consider the set of bump functions $\{f_{\boldsymbol{z}}\}_{\boldsymbol{z} \in Z}$, where Z and $f_{\boldsymbol{z}} := f_{\alpha,\eta,\boldsymbol{z}}$ are defined as in Lemma 6,[11] for $\alpha := 2\varepsilon$ and some $\eta \in (0, 1/4]$ to be selected later. Fix an arbitrary $\boldsymbol{u} = (u_1, \dots, u_d) \in Z$. We show now that $f_{\boldsymbol{u}}$ is (c,γ)-Hölder, for a suitable choice of η. For all $\boldsymbol{x}, \boldsymbol{y}$ in the closure $\overline{\{f_{\boldsymbol{u}} > 0\}}$ of $\{f_{\boldsymbol{u}} > 0\}$, Lemma 6 gives

$$\begin{aligned}
\big|f_{\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{u}}(\boldsymbol{y})\big| &\le 2\varepsilon \sum_{j=1}^d \left|\widetilde{f}\left(\frac{x_j - u_j}{\eta}\right) - \widetilde{f}\left(\frac{y_j - u_j}{\eta}\right)\right| \le 2\varepsilon \sum_{j=1}^d 3\left|\frac{x_j - u_j}{\eta} - \frac{y_j - u_j}{\eta}\right| \le \frac{6\varepsilon d}{\eta}\|\boldsymbol{x} - \boldsymbol{y}\|_\infty \\
&= \frac{6\varepsilon d}{\eta}\|\boldsymbol{x} - \boldsymbol{y}\|_\infty^{1-\gamma}\|\boldsymbol{x} - \boldsymbol{y}\|_\infty^\gamma \le \frac{6\varepsilon d}{\eta}(2\eta)^{1-\gamma}\|\boldsymbol{x} - \boldsymbol{y}\|_\infty^\gamma = \frac{6\varepsilon d 2^{1-\gamma}}{\eta^\gamma}\|\boldsymbol{x} - \boldsymbol{y}\|_\infty^\gamma \,,
\end{aligned}$$

where the first inequality follows by applying d times the elementary consequence of the triangular inequality $\big|g_1(\boldsymbol{x}_1)g_2(\boldsymbol{x}_2) - g_1(\boldsymbol{y}_1)g_2(\boldsymbol{y}_2)\big| \le \max\big\{\|g_1\|_\infty, \|g_2\|_\infty\big\}\big(|g_1(\boldsymbol{x}_1) - g_1(\boldsymbol{y}_1)| + |g_2(\boldsymbol{x}_2) - g_2(\boldsymbol{y}_2)|\big)$, which holds for any two bounded functions $g_i \colon E_i \subset \mathbb{R}^{d_i} \to \mathbb{R}$ ($d_i \in \mathbb{N}^*, i \in \{1, 2\}$). If $\boldsymbol{x}', \boldsymbol{y}' \notin \overline{\{f_{\boldsymbol{u}} > 0\}}$, then $f_{\boldsymbol{u}}(\boldsymbol{x}') = 0 = f_{\boldsymbol{u}}(\boldsymbol{y}')$, hence $\big|f_{\boldsymbol{u}}(\boldsymbol{x}') - f_{\boldsymbol{u}}(\boldsymbol{y}')\big| = 0$. Finally, if $\boldsymbol{x} \in \{f_{\boldsymbol{u}} > 0\}$ but $\boldsymbol{y}' \notin \overline{\{f_{\boldsymbol{u}} > 0\}}$, let $\boldsymbol{y}$ be the unique[12] point in the intersection of the segment $[\boldsymbol{x}, \boldsymbol{y}']$ and the boundary $\partial\{f_{\boldsymbol{u}} > 0\}$ of $\{f_{\boldsymbol{u}} > 0\}$; since $f_{\boldsymbol{u}}$ vanishes at the boundary of $\{f_{\boldsymbol{u}} > 0\}$, then $f_{\boldsymbol{u}}(\boldsymbol{y}) = f_{\boldsymbol{u}}(\boldsymbol{y}')$, therefore $\big|f_{\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{u}}(\boldsymbol{y}')\big| = \big|f_{\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{u}}(\boldsymbol{y})\big|$ and we can reapply the argument above for $\boldsymbol{x}, \boldsymbol{y}$ now both in $\overline{\{f_{\boldsymbol{u}} > 0\}}$, obtaining

$$\big|f_{\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{u}}(\boldsymbol{y}')\big| = \big|f_{\boldsymbol{u}}(\boldsymbol{x}) - f_{\boldsymbol{u}}(\boldsymbol{y})\big| \le \frac{6\varepsilon d 2^{1-\gamma}}{\eta^\gamma}\|\boldsymbol{x} - \boldsymbol{y}\|_\infty^\gamma \le \frac{6\varepsilon d 2^{1-\gamma}}{\eta^\gamma}\|\boldsymbol{x} - \boldsymbol{y}'\|_\infty^\gamma \,,$$

where the last inequality follows by $\|\boldsymbol{x} - \boldsymbol{y}\|_\infty \le \|\boldsymbol{x} - \boldsymbol{y}'\|_\infty$ and the monotonocity of $x \mapsto x^\gamma$ on $[0, \infty)$. Thus, selecting $\eta = (6\varepsilon d 2^{1-\gamma}/c)^{1/\gamma}$ so that $6\varepsilon d 2^{1-\gamma}/\eta^\gamma = c$, we obtain that $f_{\boldsymbol{z}}$ is (c,γ)-Hölder for all $\boldsymbol{z} \in Z$.

[11] More precisely, $f_{\boldsymbol{z}}$ is the restriction of $f_{\alpha,\eta,\boldsymbol{z}}$ to $[0, 1]^d$.

[12] This follows from two simple observations. First, since $f_{\boldsymbol{u}}$ is continuous, the set $\{f_{\boldsymbol{u}} > 0\}$ is open, hence $\boldsymbol{x}$ belongs to its interior. Second, $\{f_{\boldsymbol{u}} > 0\}$ is (the interior of) a hypercube, therefore it is convex.

Moreover, by definition of Z (Lemma 6) and κ, we have that

$$|Z| = \left\lfloor \frac{1}{2\eta} + 1 \right\rfloor^d \geq \left(\frac{1}{2\eta}\right)^d = \left(\frac{1}{2(6\varepsilon d 2^{1-\gamma}/c)^{1/\gamma}}\right)^d = \left(\frac{c}{12d}\right)^{d/\gamma} \frac{1}{\varepsilon^{d/\gamma}} = \kappa \frac{1}{\varepsilon^{d/\gamma}} \,.$$

Recall that the sets $\{f_{\boldsymbol{z}_1} > 0\}$ and $\{f_{\boldsymbol{z}_2} > 0\}$ are disjoint for distinct $\boldsymbol{z}_1, \boldsymbol{z}_2 \in Z$ (Lemma 6). Thus, consider an arbitrary deterministic algorithm and assume that only $n < \kappa/\varepsilon^{d/\gamma}$ values are queried. By construction, there exists at least a $\boldsymbol{z} \in \mathcal{P}$ such that, if the algorithm is run for the level set $\{f = 0\}$ of the constant function $f \equiv 0$, no points are queried inside $\{f_{\boldsymbol{z}} > 0\}$ (and being f constant, the algorithm always observes 0 as feedback for the n evaluations). Being deterministic, if the algorithm is run for the level set $\{f_{\boldsymbol{z}} = 0\}$ of $f_{\boldsymbol{z}}$ it will also query no points inside $\{f_{\boldsymbol{z}} > 0\}$, observing only zeros for all the n evaluations. Since either way, only zeros are observed, using again the fact that the algorithm is deterministic, it returns the same output set S_n in both cases. This set cannot be simultaneously an ε-approximation of both $\{f = 0\}$ and $\{f_{\boldsymbol{z}} = 0\}$. Indeed, for the first set we have that $\{f = 0\} = [0,1]^d = \{f \leq \varepsilon\}$. Thus, if S_n is an ε-approximation of $\{f = 0\}$ it has to satisfy $\{f = 0\} \subset S_n \subset \{f \leq \varepsilon\}$, which in turn gives $S_n = [0,1]^d$. On the other hand, $\max_{\boldsymbol{x} \in [0,1]^d} f_{\boldsymbol{z}}(\boldsymbol{x}) = 2\,\varepsilon$, which implies that $\{f_{\boldsymbol{z}} \leq \varepsilon\}$ is *properly* included in $[0,1]^d$. Hence, if $S_n = [0,1]^d$ were also an ε-approximation of $\{f_{\boldsymbol{z}} = 0\}$, we would have that $[0,1]^d = S_n \subset \{f_{\boldsymbol{z}} \leq \varepsilon\} \neq [0,1]^d$, which yields a contradiction. This concludes the proof of the first claim. The second claim follows directly from the first part and Definition 7. □

E MISSING PROOFS OF SECTION 5

In this section, we present all missing proofs of our results in Section 4. We restate them to ease readability.

E.1 Upper Bound

Lemma (Lemma 1)**.** *Let $f\colon C' \to \mathbb{R}$ be a (c_1, γ_1)-gradient-Hölder function, for some $c_1 > 0$ and $\gamma_1 \in (0,1]$. Let $C' \subset [0,1]^d$ be a hypercube with diameter $\ell \in (0,1]$ and set of vertices V', i.e., $C' = \prod_{j=1}^d [u_j, u_j + \ell]$, for some $\boldsymbol{u} := (u_1, \ldots, u_d) \in [0, 1-\ell]^d$, and $V' = \prod_{j=1}^d \{u_j, u_j + \ell\}$. The function*

$$\begin{aligned} h_{C'}\colon C' &\to \mathbb{R} \\ \boldsymbol{x} &\mapsto \sum_{\boldsymbol{v} \in V'} f(\boldsymbol{v}) \prod_{j=1}^d p_{v_j}(x_j) \,, \end{aligned}$$

where

$$p_{v_j}(x_j) := \left(1 - \frac{x_j - u_j}{\ell}\right) \mathbb{I}_{v_j = u_j} + \frac{x_j - u_j}{\ell}\, \mathbb{I}_{v_j = u_j + \ell},$$

interpolates the 2^d pairs $\big\{(\boldsymbol{v}, f(\boldsymbol{v}))\big\}_{\boldsymbol{v} \in V'}$ and it satisfies

$$\sup_{\boldsymbol{x} \in C'} \big| h_{C'}(\boldsymbol{x}) - f(\boldsymbol{x}) \big| \leq c_1 d\, \ell^{1+\gamma_1} \,.$$

Proof. Up to applying the translation $\boldsymbol{x} \mapsto \boldsymbol{x} + \boldsymbol{u}$, we can (and do) assume without loss of generality that $\boldsymbol{u} = \boldsymbol{0}$. The hypercube and its set of vertices then become $C' = [0,\ell]^d$ and $V' = \{0,\ell\}^d$ respectively. To verify that $h_{C'}$ interpolates the 2^d pairs $\big\{(\boldsymbol{v}, f(\boldsymbol{v}))\big\}_{\boldsymbol{v} \in V'}$, note that by definition of $h_{C'}$, for any vertex $\boldsymbol{w} \in V' = \{0,\ell\}^d$, we have

$$h_{C'}(\boldsymbol{w}) = \sum_{\boldsymbol{v} \in V'} f(\boldsymbol{v}) \prod_{j=1}^d p_{v_j}(w_j) = \sum_{\boldsymbol{v} \in V'} f(\boldsymbol{v}) \prod_{j=1}^d \mathbb{I}_{w_j = v_j} = f(\boldsymbol{w}) \,.$$

To prove the inequality, for all $k \in \{0, \ldots, d\}$, let (P_k) be the property: if an $\boldsymbol{x} \in C'$ has at most k components which are not in $\{0,\ell\}$, then it holds that $\big|h_{C'}(\boldsymbol{x}) - f(\boldsymbol{x})\big| \leq c_1 k\, \ell^{1+\gamma_1}$. To show that $\big|h_{C'}(\boldsymbol{x}) - f(\boldsymbol{x})\big| \leq$

$c_1 d\,\ell^{1+\gamma_1}$ for all $\boldsymbol{x} \in C'$ (therefore concluding the proof) we then only need to check that the property (P_d) is true. We do so by induction. If $k = 0$, then (P_0) follows by $h_{C'}$ being an approximator for $\big\{(\boldsymbol{v}, f(\boldsymbol{v}))\big\}_{\boldsymbol{v}\in V'}$. Assume now that ($\mathrm{P}_k$) holds for $k \in \{0, \dots, d-1\}$. To prove (P_{k+1}), fix an arbitrary $\boldsymbol{x} := (x_1, \dots, x_d) \in C'$, assume that $k+1$ components of $\boldsymbol{x}$ are not in $\{0, \ell\}$ and let $i \in \{1, \dots, d\}$ be any one of them (i.e., $x_i \in (0, \ell)$). Consider the two univariate functions

$$\begin{aligned} h_i \colon [0,\ell] &\to \mathbb{R} \\ t &\mapsto h_{C'}(x_1, \dots, x_{i-1}, t, x_{i+1}, \dots, x_d)\ , \\ f_i \colon [0,\ell] &\to \mathbb{R} \\ t &\mapsto f(x_1, \dots, x_{i-1}, t, x_{i+1}, \dots, x_d)\ . \end{aligned}$$

Being h_i linear (by definition of $h_{C'}$), we get

$$h_i(x_i) = \frac{\ell - x_i}{\ell} h_i(0) + \frac{x_i}{\ell} h_i(\ell)\ . \tag{15}$$

Being f_i continuous on $[0, \ell]$ and derivable on $(0, \ell)$ (by our assumptions on f), the mean value theorem applied to f_i on $[0, x_i]$ and $[0, \ell]$ respectively yields the existence of $\xi_1 \in (0, x_i)$ and $\xi_2 \in (0, \ell)$ such that

$$f_i(x_i) = f_i(0) + f_i'(\xi_1)\, x_i\ , \tag{16}$$

$$f_i(\ell) = f_i(0) + f_i'(\xi_2)\, \ell\ . \tag{17}$$

Putting everything together, we get that

$$\big|h_{C'}(\boldsymbol{x}) - f(\boldsymbol{x})\big| = \big|h_i(x_i) - f_i(x_i)\big|$$

(by definition of h_i and f_i). By (15) and (16), the right-hand side is equal to

$$\left|\frac{\ell - x_i}{\ell} h_i(0) + \frac{x_i}{\ell} h_i(\ell) - f_i(0) - f_i'(\xi_1)\, x_i\right|,$$

which by the triangular inequality is at most

$$\left|\frac{\ell - x_i}{\ell} h_i(0) + \frac{x_i}{\ell} h_i(\ell) - f_i(0) - x_i \frac{f_i(\ell) - f_i(0)}{\ell}\right| + \left|x_i \frac{f_i(\ell) - f_i(0)}{\ell} - f_i'(\xi_1)\, x_i\right|.$$

By (17), this is equal to

$$\left|\frac{\ell - x_i}{\ell}\big(h_i(0) - f_i(0)\big) + \frac{x_i}{\ell}\big(h_i(\ell) - f_i(\ell)\big)\right| + \big|x_i\, f_i'(\xi_2) - f_i'(\xi_1)\, x_i\big|\ .$$

Finally, using again the triangular inequality, we can further upper bound with

$$\frac{\ell - x_i}{\ell} \underbrace{\Big|\big(h_i(0) - f_i(0)\big)\Big|}_{\leq c_1 k \ell^{1+\gamma_1}} + \frac{x_i}{\ell} \underbrace{\Big|\big(h_i(\ell) - f_i(\ell)\big)\Big|}_{\leq c_1 k \ell^{1+\gamma_1}} + \underbrace{x_i}_{\leq \ell} \underbrace{\big|f_i'(\xi_2) - f_i'(\xi_1)\big|}_{\leq c_1 \ell^{\gamma_1}} \leq c_1 (k+1) \ell^{1+\gamma_1}\ ,$$

where on the last line, we applied property (P_k) to the first two terms and we upper bounded the last one leveraging the (c_1, γ_1)-Hölderness of the gradients of f. This proves (P_{k+1}) and concludes the proof. □

Proposition 3 (Theorem 4, upper bound)**.** *Consider the BAG algorithm (Algorithm 4) run with input a, c_1, γ_1. Let $f \colon [0,1]^d \to \mathbb{R}$ be an arbitrary (c_1, γ_1)-gradient-Hölder function with level set $\{f = a\} \neq \varnothing$. Fix any accuracy $\varepsilon > 0$. Then, for all*

$$n > \kappa \frac{1}{\varepsilon^{d/(1+\gamma_1)}}\ ,\ \text{if } \kappa := \big(2^{5+4\gamma_1+(1+\gamma_1)/d}\, c_1 d\big)^{d/\gamma}\ ,$$

the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$.

Proof. We proceed as in the proof of Theorem 3. Theorem 2 implies that for all $n > n(\varepsilon)$, where $i(\varepsilon) := \lceil (1/(1+\gamma_1)) \log_2(2c_1 d/\varepsilon) \rceil$ and $n(\varepsilon)$ is

$$8^d \sum_{i=0}^{i(\varepsilon)-1} \lim_{\delta \to 1^-} \mathcal{N}\Big(\big\{ |f-a| \le 2c_1 d 2^{-(1+\gamma_1)i} \big\}, \delta 2^{-i} \Big) ,$$

the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$. If $\varepsilon \ge 2c_1 d$, than the sum in the definition of $n(\varepsilon)$ ranges from 0 to a *negative* value, thus $n(\varepsilon) = 0$ by definition of sum over an empty set and the result is true with $\kappa = 0$. Assume then that $\varepsilon < 2c_1 d$ so that such sum is not trivially zero. Upper-bounding, for any $\delta \in (0,1)$ and all $i \ge 0$,

$$\mathcal{N}\Big(\big\{ |f-a| \le 2c_1 d\, 2^{-(1+\gamma_1)i} \big\},\ \delta\, 2^{-i} \Big) \le \mathcal{N}\big([0,1]^d,\ \delta\, 2^{-i} \big) \overset{(\dagger)}{\le} \big(2^i/\delta + 1 \big)^d \le (2/\delta)^d\, 2^{di}$$

(for completeness, we include a proof of the known upper bound (†) in Section B, Lemma 4) and recognizing the geometric sum below, we can conclude that

$$\begin{aligned} n(\varepsilon) &\le 16^d \sum_{i=0}^{\lceil (1/(1+\gamma_1)) \log_2(2c_1 d/\varepsilon) \rceil - 1} \big(2^d \big)^i \\ &\le 16^d\, \frac{2^{d((1/(1+\gamma_1)) \log_2(2c_1 d/\varepsilon)+1)}}{2^d - \big(2^d/2\big)} \\ &= 2 \cdot 16^d\, (2c_1 d)^{d/(1+\gamma_1)}\, \frac{1}{\varepsilon^{d/(1+\gamma_1)}} . \end{aligned}$$

□

E.2 Lower Bound

We conclude the section by proving a matching lower bound. Similarly to Proposition 2, we adapt some already known techniques from nonparametric regression (see, e.g., Györfi et al. 2002, Theorem 3.2).

Proposition 4 (Theorem 4, lower bound)**.** *Fix any level $a \in \mathbb{R}$, any two Hölder constants $c_1 > 0$, $\gamma_1 \in (0,1]$, and any accuracy $\varepsilon \in \big(0, c_1/(132 d 2^{3+\gamma_1})\big)$. Let $n < \kappa/\varepsilon^{d/(1+\gamma_1)}$ be a positive integer, where $\kappa := \big(c_1/(528d)\big)^{d/(1+\gamma_1)}$. For each deterministic algorithm A there is a (c_1, γ_1)-gradient-Hölder function f such that, if A queries n values of f, then its output set S_n is not an ε-approximation of $\{f = a\}$. This implies in particular that (recall Definition 7),*

$$\inf_A \sup_f \mathfrak{n}(f, A, \varepsilon, a) \ge \kappa\, \frac{1}{\varepsilon^{d/(1+\gamma_1)}} ,$$

where the inf *is over all deterministic algorithms A and the* sup *is over all (c_1, γ_1)-gradient-Hölder functions f.*

As we pointed out after Proposition 2, the leading constant $\kappa = \big(c_1/(528d)\big)^{d/(1+\gamma_1)}$ in our lower bound is small, and could likely be improved using smoothness-specific perturbations of the zero function, instead of the more universal bump functions.

Proof. The following construction is a standard way to prove lower bounds on sample complexity (for a similar example, see Györfi et al. 2002, Theorem 3.2). Consider the set of bump functions $\{f_{\boldsymbol{z}}\}_{\boldsymbol{z} \in Z}$, where Z and $f_{\boldsymbol{z}} := f_{\alpha,\eta,\boldsymbol{z}}$ are defined as in Lemma 6,[13] for $\alpha := 2\varepsilon$ and some $\eta \in (0, 1/4]$ to be selected later. Fix

[13] More precisely, $f_{\boldsymbol{z}}$ is the restriction of $f_{\alpha,\eta,\boldsymbol{z}}$ to $[0,1]^d$.

an arbitrary $\boldsymbol{u} = (u_1, \dots, u_d) \in Z$. We show now that $f_{\boldsymbol{u}}$ is (c_1, γ_1)-gradient-Hölder, for a suitable choice of η. This is sufficient to prove the result, following the same argument as in the proof of Proposition 2. Note first that for all $i \in \{1, \dots, d\}$ and any $\boldsymbol{x} \in [0,1]^d$, denoting by ∂_i the partial derivative with respect to the i-th variable,

$$\partial_i f_{\boldsymbol{u}}(\boldsymbol{x}) = \frac{2\varepsilon}{\eta} \widetilde{f}'\left(\frac{x_i - u_i}{\eta}\right) \prod_{\substack{j=1 \\ j \neq i}}^{d} \widetilde{f}\left(\frac{x_j - u_j}{\eta}\right) .$$

Hence, using the fact that $\widetilde{f}$ is 3-Lipschitz (Lemma 6) and 22-gradient-Lipschitz (the latter can be done by checking that $\left\|\widetilde{f}''\right\|_\infty \leq 22$), for all $i \in \{1, \dots, d\}$ and any $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^d$, we get

$$\left|\partial_i f_{\boldsymbol{u}}(\boldsymbol{x}) - \partial_i f_{\boldsymbol{u}}(\boldsymbol{y})\right| \leq \frac{2\varepsilon}{\eta} 3 \left(22 \left| \frac{x_i - u_i}{\eta} - \frac{y_i - u_i}{\eta} \right| + 3 \sum_{\substack{j=1 \\ j \neq i}}^{d} \left| \frac{x_j - u_j}{\eta} - \frac{y_j - u_j}{\eta} \right| \right) \leq \frac{132\varepsilon d}{\eta^2} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty , \quad (18)$$

where the first inequality follows by applying d times the elementary consequence of the triangular inequality $\left|g_1(\boldsymbol{x}_1)g_2(\boldsymbol{x}_2) - g_1(\boldsymbol{y}_1)g_2(\boldsymbol{y}_2)\right| \leq \max\left\{\|g_1\|_\infty, \|g_2\|_\infty\right\}\left(|g_1(\boldsymbol{x}_1) - g_1(\boldsymbol{y}_1)| + |g_2(\boldsymbol{x}_2) - g_2(\boldsymbol{y}_2)|\right)$, which holds for any two bounded functions $g_i \colon E_i \subset \mathbb{R}^{d_i} \to \mathbb{R}$ ($d_i \in \mathbb{N}^*, i \in \{1,2\}$), and then using the Lipschitzness of $\widetilde{f}$ and $\widetilde{f}'$. Similarly to Proposition 2, to prove that $f_{\boldsymbol{u}}$ is (c_1, γ_1)-gradient-Hölder, we only need to check that the gradient of $f_{\boldsymbol{u}}$ is (c_1, γ_1)-Hölder on the closure $\overline{\{f_{\boldsymbol{u}} > 0\}}$ of $\{f_{\boldsymbol{u}} > 0\}$. For all $\boldsymbol{x}, \boldsymbol{y} \in \overline{\{f_{\boldsymbol{u}} > 0\}}$, Equation (18) and Lemma 6 yield

$$\begin{aligned}
\left\|\nabla f_{\boldsymbol{u}}(\boldsymbol{x}) - \nabla f_{\boldsymbol{u}}(\boldsymbol{y})\right\|_\infty &\leq \frac{132\varepsilon d}{\eta^2} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty = \frac{132\varepsilon d}{\eta^2} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^{1-\gamma_1} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^{\gamma_1} \\
&\leq \frac{132\varepsilon d}{\eta^2} (2\eta)^{1-\gamma_1} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^{\gamma_1} = \frac{132\varepsilon d 2^{1-\gamma_1}}{\eta^{1+\gamma_1}} \|\boldsymbol{x} - \boldsymbol{y}\|_\infty^{\gamma_1} .
\end{aligned}$$

Therefore, selecting $\eta = (132\varepsilon d 2^{1-\gamma_1}/c_1)^{1/(1+\gamma_1)}$ so that $132\varepsilon d 2^{1-\gamma_1}/\eta^{1+\gamma_1} = c_1$, we obtain that $f_{\boldsymbol{z}}$ is (c_1, γ_1)-Hölder for all $\boldsymbol{z} \in Z$. Moreover, by definition of Z and κ, we have that

$$|Z| \geq \left(\frac{1}{2\eta}\right)^d = \left(\frac{1}{2(132\varepsilon d 2^{1-\gamma_1}/c_1)^{1/(1+\gamma_1)}}\right)^d = \left(\frac{c_1}{528d}\right)^{d/(1+\gamma_1)} \frac{1}{\varepsilon^{d/(1+\gamma_1)}} = \kappa \frac{1}{\varepsilon^{d/(1+\gamma_1)}} .$$

Thus, proceeding as in the proof of Proposition 2, no deterministic algorithm can output a set that is an ε-approximation of the level set $\{f = 0\} = [0,1]^d$ of the constant function $f \equiv 0$ and simultaneously an ε-approximation of the level set $\{f_{\boldsymbol{z}} = 0\}$ of all bump functions $f_{\boldsymbol{z}}$ ($\boldsymbol{z} \in Z$), without querying at least one value in each one of the $|Z| \geq \kappa/\varepsilon^{d/(1+\gamma_1)}$ disjoint sets $\{f_{\boldsymbol{z}} > 0\}$ ($\boldsymbol{z} \in Z$) when applied to $f \equiv 0$. □

F THE BENEFITS OF ADDITIONAL STRUCTURAL ASSUMPTIONS

In this section, we present some examples showing how our general results can be applied to yield (slightly) improved sample complexity bounds when f satisfies additional structural assumptions, such as convexity.

F.1 NLS Dimension

In order to derive more readable bounds on the number of queries needed to return approximations of level sets, we now introduce a quantity that measures the difficulty of finding such approximations.

Definition 9 (NLS dimension)**.** Fix any level $a \in \mathbb{R}$ and a function $f \colon [0,1]^d \to \mathbb{R}$. We say that $d^\star \in [0, d]$ is a *NLS (or Near-Level-Set) dimension* of the level set $\{f = a\}$ if there exists $C^\star > 0$ such that (recalling Definition 2 —packing number)

$$\forall r \in (0,1),\ \mathcal{N}\Big(\big\{|f - a| \leq r\big\},\ r\Big) \leq C^\star \left(\frac{1}{r}\right)^{d^\star} . \quad (19)$$

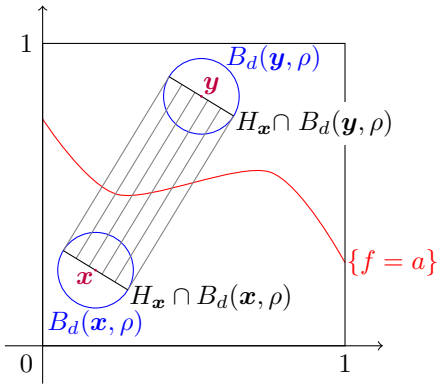

Figure 2: The "dimension" of the level set is at least the same as that of hyperplanes $H_{\boldsymbol{x}}$ and $H_{\boldsymbol{y}}$.

NLS dimensions are a natural generalization of the well-known concept of near-optimality dimension, from the field of non-convex optimization (see, e.g., (Bouttier et al., 2020, Section 2.3 and following discussion in Appendix B)). The idea behind Inequality (19) is that inflated level sets at, say, scale $r \in (0, 1)$, are hard to pinpoint if their complement $\{|f - a| > r\}$ is large. Since for any increasing sequence $r := r_0 < r_1 < r_2 < \ldots$, the set $\{|f - a| > r\}$ of points at which f is more than r-away from a can be decomposed into a union of "layers" $\{r_0 < |f - a| \leq r_1\}, \{r_1 < |f - a| \leq r_2\}, \{r_2 < |f - a| \leq r_3\}, \ldots$, and each of these layers $\{r_{s-1} < |f - a| \leq r_s\}$ is by definition included in $\{|f - a| \leq r_s\}$, by controlling the size of each of these $\{|f - a| \leq r_s\}$ we can control the size of $\{|f - a| > r\}$. Therefore, by controlling how large the inflated level sets $\{|f - a| \leq r\}$ can be at all scales $r \in (0, 1)$, the parameters $C^\star$ and $d^\star$ quantify the difficulty of the level set approximation problem. In contrast, scales $r \geq 1$ are not informative since in this case the packing number in (19) is always 1. To see this, simply note that if $r \geq 1$, no more than 1 strictly r-separated point can be packed in $\{|f - a| \leq r\}$, which is included in $[0, 1]^d$, that has diameter 1 (in the sup-norm).

The dimension d of the domain is always a NLS dimension of any function $f\colon [0, 1]^d \to \mathbb{R}$ (with $C^\star = 2^d$; we add a proof of this claim in Section B, Lemma 5). Hence, it is sufficient to consider NLS dimensions $d^\star \leq d$, as we do in our Definition 9. While (as we will see in Section F.2) $d^\star$ is in general strictly smaller than d, bounds expressed in terms of a NLS dimension should only be considered slight refinements of worst-case bounds expressed in terms of d. Indeed, the following result shows that, with the exceptions of sets of minimizers and maximizers, level sets $\{f = a\}$ of continuous functions f have NLS dimension at least $d - 1$.

Theorem 6 (Theorem 1)**.** *Let $f\colon [0, 1]^d \to \mathbb{R}$ be a non-constant continuous function, and $a \in \mathbb{R}$ be any level such that $\min_{\boldsymbol{x} \in [0,1]^d} f(\boldsymbol{x}) < a < \max_{\boldsymbol{x} \in [0,1]^d} f(\boldsymbol{x})$. Then, there exists $C^\star > 0$ such that, for all $r > 0$,*

$$\mathcal{N}\big(\{f = a\},\ r\big) \geq C^\star \left(\frac{1}{r}\right)^{d-1}.$$

Proof. For all $d_0 \in \mathbb{N}^*$, $\boldsymbol{z} \in \mathbb{R}^{d_0}$, and $\rho > 0$, we denote by $B_{d_0}(\boldsymbol{z}, \rho)$ the closed d_0-dimensional Euclidean ball $\{\boldsymbol{x} \in \mathbb{R}^{d_0} : \|\boldsymbol{x} - \boldsymbol{z}\|_2 \leq \rho\}$ with center $\boldsymbol{z}$ and radius ρ. Since a is neither the maximum nor the minimum of f and f is continuous, then the two sets $\{f < a\}$ and $\{f > a\}$ are non-empty and open. Therefore, we claim that there exist two points $\boldsymbol{x} \in \{f < a\}$, $\boldsymbol{y} \in \{f > a\}$, and a radius $\rho > 0$, such that $B_d(\boldsymbol{x}, \rho) \subset \{f < a\} \cap (0, 1)^d$ and $B_d(\boldsymbol{y}, \rho) \subset \{f > a\} \cap (0, 1)^d$ (Figure 2). To see this, note that if f were identically equal to a on $(0, 1)^d$, then, by continuity, f would be identically equal to a on the whole $[0, 1]^d$, contradicting the assumption that it is non-constant. Hence there exists an $\boldsymbol{x} \in (0, 1)^d$ such that $f(\boldsymbol{x}) \neq a$. Assume that $f(\boldsymbol{x}) < a$ (for the opposite case, proceed analogously). Then, being $\{f < a\} \cap (0, 1)^d$ open, there exists a radius $\rho_1 > 0$ such that $B_d(\boldsymbol{x}, \rho_1) \subset \{f < a\} \cap (0, 1)^d$. Now, if f were lower than or equal to a on $(0, 1)^d$, then, by continuity, f would be lower than or equal to a on the whole $[0, 1]^d$ (and so would be its maximum), contradicting the assumption that $a < \max(f)$. Hence, there exists an $\boldsymbol{y} \in (0, 1)^d$ such that $f(\boldsymbol{y}) > a$. Then, being

$\{f > a\} \cap (0,1)^d$ open, there exists a radius $\rho_2 > 0$ such that $B_d(\boldsymbol{y}, \rho_2) \subset \{f > a\} \cap (0,1)^d$. The claim is therefore proven by letting $\rho := \min(\rho_1, \rho_2)$.

Now, for all $r \geq \rho/\sqrt{d}$, we have

$$\mathcal{N}\big(\{f = a\},\, r\big) \geq 1 \geq \left(\frac{\rho}{\sqrt{d}}\right)^{d-1} \left(\frac{1}{r}\right)^{d-1}$$

and the result is proven with $C^\star = \big(\rho/\sqrt{d}\big)^{d-1}$.

Fix now an arbitrary $r \in \big(0, \rho/\sqrt{d}\big)$. Consider the line $\mathcal{L} := \big\{(1-t)\boldsymbol{x} + t\boldsymbol{y} : t \in \mathbb{R}\big\}$ passing through $\boldsymbol{x}$ and $\boldsymbol{y}$ and the two hyperplanes $H_{\boldsymbol{x}}$ and $H_{\boldsymbol{y}}$ orthogonal to $\mathcal{L}$ and passing through $\boldsymbol{x}$ and $\boldsymbol{y}$ respectively. We denote, for each $\boldsymbol{z} \in \mathbb{R}^d$ and $E \subset \mathbb{R}^d$, the Minkowski sum $\{\boldsymbol{z} + \boldsymbol{u} : \boldsymbol{u} \in E\}$ of $\{\boldsymbol{z}\}$ and E by $\boldsymbol{z} + E$. Note that, by construction, $(\boldsymbol{y} - \boldsymbol{x}) + \big(H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)\big) = H_{\boldsymbol{y}} \cap B_d(\boldsymbol{y}, \rho)$, and there is a rigid transformation $T \colon \mathbb{R}^d \to \mathbb{R}^d$ that maps $H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)$ into the $(d-1)$-dimensional Euclidean ball $B_{d-1}(\boldsymbol{0}, \rho) = \big\{\boldsymbol{z} \in \mathbb{R}^{d-1} : \|\boldsymbol{z}\|_2 \leq \rho\big\}$ of $\mathbb{R}^{d-1}$ (where, with a slight abuse of notation, we identify from here on out $\mathbb{R}^{d-1}$ with the subspace $\big\{(z_1, \ldots, z_d) \in \mathbb{R}^d : z_d = 0\big\}$ of $\mathbb{R}^d$). By the symmetry of the Euclidean balls, for all $\boldsymbol{z}' \in \big(H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)\big)$ and $\rho' > 0$, the transformed through the rigid transformation T of the intersection $B_d(\boldsymbol{z}', \rho') \cap \big(H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)\big)$ of an arbitrary d-dimensional Euclidean ball $B_d(\boldsymbol{z}', \rho')$ centered at $H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)$ and $H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)$ itself is simply the intersection $B_{d-1}(\boldsymbol{z}'', \rho') \cap B_{d-1}(\boldsymbol{0}, \rho)$ between the ball $B_{d-1}(\boldsymbol{0}, \rho)$ and a $(d-1)$-dimensional ball $B_{d-1}(\boldsymbol{z}'', \rho')$ with some center $\boldsymbol{z}'' \in B_{d-1}(\boldsymbol{0}, \rho)$ and the same radius ρ' of $B_d(\boldsymbol{z}', \rho')$. We recall that for any dimension $d_0 \in \mathbb{N}^*$, norm $\|\cdot\|$ on $\mathbb{R}^{d_0}$, scale $r_0 > 0$, and non-empty subset E_0 of $\mathbb{R}^{d_0}$, a set $P \subset E_0$ is an r_0-packing of E_0 in $\mathbb{R}^{d_0}$ with respect to $\|\cdot\|$ if each two distinct points $\boldsymbol{z}_1, \boldsymbol{z}_2 \in P$ satisfy $\|\boldsymbol{z}_1 - \boldsymbol{z}_2\| > r_0$, and a set $C \subset E_0$ is an r_0-covering of E_0 in $\mathbb{R}^{d_0}$ with respect to $\|\cdot\|$ if for all $\boldsymbol{z} \in E_0$ there exists $\boldsymbol{c} \in C$ such that $\|\boldsymbol{z} - \boldsymbol{c}\| \leq r_0$; we denote by $\mathcal{N}_{d_0, \|\cdot\|}(E_0, r_0)$ the largest cardinality of an r_0-packing of E_0 in $\mathbb{R}^{d_0}$ with respect to $\|\cdot\|$, and by $\mathcal{M}_{d_0, \|\cdot\|}(E_0, r_0)$ the smallest cardinality of an r_0-covering of E_0 in $\mathbb{R}^{d_0}$ with respect to $\|\cdot\|$. By the previous observation, then, for all $r_0 > 0$, a set is an r_0-packing (resp., covering) of $H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)$ in $\mathbb{R}^d$ with respect to the d-dimensional Euclidean norm if and only if its transformed under T is an r_0-packing (resp., covering) of $B_{d-1}(\boldsymbol{0}, \rho)$ in $\mathbb{R}^{d-1}$ with respect to the $(d-1)$-dimensional Euclidean norm. Hence

$$\begin{aligned}
&\mathcal{N}_{d, \|\cdot\|_2}\Big(H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho),\, \sqrt{d}\, r\Big) \\
&\quad = \mathcal{N}_{d-1, \|\cdot\|_2}\big(B_{d-1}(\boldsymbol{0}, \rho),\, \sqrt{d}\, r\big) \\
&\quad \geq \mathcal{M}_{d-1, \|\cdot\|_2}\big(B_{d-1}(\boldsymbol{0}, \rho),\, \sqrt{d}\, r\big) \geq \left(\frac{\rho}{\sqrt{d}\, r}\right)^{d-1},
\end{aligned}$$

where the first inequality follows from the fact that each packing that is maximal with respect to the inclusion is also a covering, and the second one is a known lower bound on the number of balls with the same radius that are needed to cover a ball with a bigger radius, expressed in terms of a ratio of volumes (see, e.g., (Wainwright, 2019, Lemma 5.7)). Thus, we determined a $\big(\sqrt{d}\, r\big)$-packing P of $H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho)$ in $\mathbb{R}^d$ with respect to the d-dimensional Euclidean norm consisting of $C^\star(1/r)^{d-1}$ points, where again $C^\star := \big(\rho/\sqrt{d}\big)^{d-1}$. For all $\boldsymbol{p} \in P$, consider the segment $[\boldsymbol{p}, \boldsymbol{p} + \boldsymbol{y} - \boldsymbol{x}]$. By construction, all these segments are parallel, with an endpoint in $H_{\boldsymbol{x}} \cap B_d(\boldsymbol{x}, \rho) \subset \{f < a\}$ and the other in $H_{\boldsymbol{y}} \cap B_d(\boldsymbol{y}, \rho) \subset \{f > a\}$. Thus, the d-dimensional Euclidean distance between any two points belonging to distinct segments is at least equal to the minimum distance between the corresponding lines, which is strictly greater than $\sqrt{d}\, r$ by construction. By the continuity of f, then, for each $\boldsymbol{p} \in P$ there exists a $\boldsymbol{p}_a$ belonging to the segment $[\boldsymbol{p}, \boldsymbol{p} + \boldsymbol{y} - \boldsymbol{x}]$ such that $f(\boldsymbol{p}_a) = a$ which, together with the previous remark, implies that the family $P_a := \bigcup_{\boldsymbol{p} \in P} \boldsymbol{p}_a$ obtained this way is a $\big(\sqrt{d} r\big)$-packing of $\{f = a\}$ in $\mathbb{R}^d$ with respect to the d-dimensional Euclidean norm. Since the two norms $\|\cdot\|_\infty$ and $\|\cdot\|_2$ on $\mathbb{R}^d$ satisfy $\|\cdot\|_\infty \geq \|\cdot\|_2/\sqrt{d}$, then P_a is also an r-packing of $\{f = a\}$ in $\mathbb{R}^d$ with respect to the sup-norm $\|\cdot\|_\infty$, therefore its cardinality $|P_a| = C^\star(1/r)^{d-1}$ is smaller than or equal to the largest cardinality $\mathcal{N}\big(\{f = a\}, r\big)$ of an r-packing of $\{f = a\}$ with respect to the sup-norm $\|\cdot\|_\infty$. This concludes the proof. □

We remark that our definition in Equation (19) could be refined by considering variable $d^\star(r)$ and $C^\star(r)$ at different scales r. This would take into account that at different scales, the inflated level sets could have smaller size. Notably, our general result (Theorem 2) would naturally adapt to this finer definition as they are stated in terms of packing numbers at decreasing scales. For the sake of clarity, in this work we will stick to our worst-case definition of NLS dimension and we begin by showing how Theorem 2 has an immediate corollary in terms of $d^\star$. Note that in the following results, our BA instances are oblivious to the NLS dimension. Also, recall from the comment before Theorem 6 that typical level sets have NLS dimension $d^\star \geq d-1$.

Corollary 2 (of Theorem 2)**.** *Consider a Bisect and Approximate algorithm (Algorithm 2) run with input a, k, b, β. Let $f\colon [0,1]^d \to \mathbb{R}$ be an arbitrary function with level set $\{f = a\} \neq \varnothing$ and let $d^\star \in [0,d]$ be a NLS dimension of $\{f = a\}$ (Definition 9). Assume that the approximators $g_{C'}$ (defined at line 10) are (b,β)-accurate approximations of f (Definition 4), with $\beta \geq 1$. Fix any accuracy $\varepsilon > 0$. Then, for all $n > m(\varepsilon)$, the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$, where*

$$m(\varepsilon) := \begin{cases} \kappa_1 + \kappa_2 \log_2\left(\frac{1}{\varepsilon^{1/\beta}}\right)^+ & \text{if } d^\star = 0\,, \\ \kappa(d^\star)\frac{1}{\varepsilon^{d^\star/\beta}} & \text{if } d^\star > 0\,, \end{cases}$$

for $\kappa_1, \kappa_2, \kappa(d^\star) \geq 0$ independent of ε, that depend exponentially on the dimension d, where $x^+ = \max\{x, 0\}$ for all $x \in \mathbb{R}$.

Proof. Since all the conditions of Theorem 2 are met by assumption, we have that for all $n > n(\varepsilon)$, the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$, where $n(\varepsilon)$ is

$$4^d\, k \sum_{i=0}^{i(\varepsilon)-1} \lim_{\delta \to 1^-} \mathcal{N}\Big(\big\{|f-a| \leq 2\,b\,2^{-\beta i}\big\},\ \delta\, 2^{-i}\Big) \tag{20}$$

and $i(\varepsilon) := \left\lceil (1/\beta)\log_2(2b/\varepsilon)\right\rceil$. If $\varepsilon \geq 2b$, than the sum in the definition of $n(\varepsilon)$ ranges from 0 to a *negative* value, thus $n(\varepsilon) = 0$ by definition of sum over an empty set and the result is true with $\kappa_1 = \kappa_2 = \kappa(d^\star) = 0$. Assume then that $\varepsilon < 2b$ so that such sum is not trivially zero. Being $\beta \geq 1$, we can further upper bound $n(\varepsilon)$ by

$$4^d\, k \sum_{i=0}^{i(\varepsilon)-1} \lim_{\delta \to 1^-} \mathcal{N}\Big(\big\{|f-a| \leq 2b\,2^{-i}\big\},\ \delta\, 2^{-i}\Big)\,.$$

By Lemma 3, the packing number is at most

$$\left(1 + 4\,\frac{2b}{\delta}\,\mathbb{I}_{2b>\delta}\right)^d \mathcal{N}\Big(\big\{|f-a| \leq 2b\,2^{-i}\big\},\ 2b\,2^{-i}\Big)\,.$$

Taking the limit for $\delta \to 1^-$, the first term becomes $\left(1 + 8b\,\mathbb{I}_{2b\geq 1}\right)^d$, while our NLS assumption (19) implies that the packing number is smaller than, or equal to

$$\mathbb{I}_{2b2^{-i}\geq 1} + C^\star\left(\frac{1}{2b\,2^{-i}}\right)^{d^\star} \mathbb{I}_{2b2^{-i}<1} \tag{21}$$

A direct computation shows that the sum over i of the first term in (21) is

$$\sum_{i=0}^{i(\varepsilon)-1} \mathbb{I}_{2b2^{-i}\geq 1} \leq \log_2(4b)\,\mathbb{I}_{2b\geq 1}\,. \tag{22}$$

For the sum over i of second term in (21), we upper bound the indicator function $\mathbb{I}_{2b2^{-i}<1}$ with 1 for all i and study separately the two cases $d^\star = 0$ and $d^\star > 0$. If $d^\star = 0$, then, by definition of $i(\varepsilon)$,

$$\sum_{i=0}^{i(\varepsilon)-1} \left(2^{d^\star}\right)^i = i(\varepsilon) \leq \log_2\left(\frac{1}{\varepsilon^{1/\beta}}\right) + \log_2\left(2\,(2b)^{1/\beta}\right) .$$

Hence, the result follows by defining the additive and multiplicative terms κ_1 and κ_2, respectively, by $\kappa' k \left(\log_2(4b)\,\mathbb{I}_{2b\geq 1} + C^\star \log_2\left(2\,(2b)^{1/\beta}\right)\right)$ and $\kappa' k \frac{C^\star}{(2b)^{d^\star}}$, where $\kappa' := \left(4 + 32\,b\,\mathbb{I}_{2b\geq 1}\right)^d$.

If on the other hand, $d^\star > 0$, recognizing the geometric sum below, we have, by definition of $i(\varepsilon)$

$$\sum_{i=0}^{i(\varepsilon)-1} \left(2^{d^\star}\right)^i = \frac{\left(2^{d^\star}\right)^{i(\varepsilon)} - 1}{2^{d^\star} - 1} \leq \frac{2^{d^\star}\,(2\,b)^{d^\star/\beta}}{2^{d^\star} - 1} \frac{1}{\varepsilon^{d^\star/\beta}} .$$

Thus, if $2b < 1$ or if simultaneously $2b \geq 1$ and $\varepsilon \leq 1/\left(\log_2(4b)\right)^{\beta/d^\star}$ —so that the term $\log_2(4b)\,\mathbb{I}_{2b\geq 1}$ in (22) can be upper bounded by $1/\varepsilon^{d^\star/\beta}\,\mathbb{I}_{2b\geq 1}$— the result follows by defining $\kappa(d^\star)$ as

$$\left(4 + 32\,b\,\mathbb{I}_{2b\geq 1}\right)^d k\,\frac{C^\star}{(2b)^{d^\star}} \left(\mathbb{I}_{2b\geq 1} + \frac{2^{d^\star}\,(2\,b)^{d^\star/\beta}}{2^{d^\star} - 1}\right) .$$

Finally, we consider the case in which $2b \geq 1$ and $\varepsilon > 1/\left(\log_2(4b)\right)^{\beta/d^\star}$. In this simpler instance, we upper bound $n(\varepsilon)$ as in the proofs of Theorems 3 and 4. Look back at Equation (20). Upper-bounding, for any $\delta \in (0,1)$ and all $i \geq 0$,

$$\mathcal{N}\Big(\left\{|f-a| \leq 2\,b\,2^{-\beta i}\right\},\ \delta\,2^{-i}\Big) \leq \mathcal{N}\left([0,1]^d,\ \delta\,2^{-i}\right) \overset{(\dagger)}{\leq} \left(2^i/\delta + 1\right)^d \leq (2/\delta)^d\,2^{di}$$

(for completeness, we include a proof of the known upper bound (†) in Section B, Lemma 4) and recognizing the geometric sum below, we have

$$\begin{aligned} n(\varepsilon) &\leq 8^d\,k \sum_{i=0}^{\lceil (1/\beta)\log_2(2b/\varepsilon)\rceil - 1} \left(2^d\right)^i \\ &= 8^d k \frac{2^{d\lceil (1/\beta)\log_2(2b/\varepsilon)\rceil} - 1}{2^d - 1} \leq 8^d k \frac{2^{d((1/\beta)\log_2(2b/\varepsilon)+1)}}{2^d - \left(2^d/2\right)} \\ &= 2\,8^d\,k\,(2\,b)^{d/\beta}\,\frac{1}{\varepsilon^{d/\beta}} . \end{aligned}$$

Finally, using the assumption $\varepsilon > 1/\left(\log_2(4b)\right)^{\beta/d^\star}$, we can upper bound the term $1/\varepsilon^{d/\beta}$ with

$$\begin{aligned} \frac{1}{\varepsilon^{d/\beta}} &= \left(\frac{1}{\varepsilon}\right)^{(d-d^\star)/\beta} \frac{1}{\varepsilon^{d^\star/\beta}} \\ &< \left(\left(\log_2(4b)\right)^{\beta/d^\star}\right)^{(d-d^\star)/\beta} \frac{1}{\varepsilon^{d^\star/\beta}} \\ &= \left(\log_2(4b)\right)^{(d-d^\star)/d^\star} \frac{1}{\varepsilon^{d^\star/\beta}} , \end{aligned}$$

and the results follows after defining the constant $\kappa(d^\star) := 2\,8^d\,k\,(2\,b)^{d/\beta}\left(\log_2(4b)\right)^{(d-d^\star)/d^\star}$. □

The previous result has the following immediate consequence for BAG algorithms. Recall from the comment before Theorem 6 that typical level sets have NLS dimension $d^\star \geq d-1$.

Corollary (Corollary 1)**.** *Consider the BAG algorithm (Algorithm 4) run with input a, c_1, γ_1. Let $f\colon [0,1]^d \to \mathbb{R}$ be an arbitrary (c_1,γ_1)-gradient-Hölder function with level set $\{f = a\} \neq \varnothing$ and let $d^\star \in [0,d]$ be a NLS dimension of $\{f = a\}$ (Definition 9). Fix any accuracy $\varepsilon > 0$. Then, for all $n > m(\varepsilon)$, the output S_n returned after the n-th query is an ε-approximation of $\{f = a\}$, where*

$$m(\varepsilon) := \begin{cases} \kappa_1 + \kappa_2 \log_2\left(\frac{1}{\varepsilon^{1/(1+\gamma_1)}}\right) & \text{if } d^\star = 0\ , \\ \kappa(d^\star)\frac{1}{\varepsilon^{d^\star/(1+\gamma_1)}} & \text{if } d^\star > 0\ , \end{cases}$$

for $\kappa_1, \kappa_2, \kappa(d^\star) \geq 0$ independent of ε, that depend exponentially on the dimension d.

Proof. The result follows immediately from Corollary 2 and Lemma 1. □

The two previous corollaries suggest a general method for solving the level set approximation problem for a given class $\mathcal{F}$, obtaining bounds that are slightly more refined than the worst-case ones that we saw in Sections 4 and 5. First, determine a family of approximators that accurately approximate the functions in $\mathcal{F}$. Second, obtain for the resulting choice of BA algorithm a sample complexity bound in terms of packing numbers of inflated level sets (as in Theorem 2). Third, find a NLS dimension of an arbitrary $f \in \mathcal{F}$. Importantly, both steps one and three of this process are decoupled from the task of determining approximations of level sets, and as such, they can be investigated independently. For step two, we can simply plug in Theorem 2.

In the next section, we will discuss the notable convex case, in which the estimation of the NLS dimension is non-trivial. As it turns out, this also leads to a rate-optimal sample complexity for BA algorithms.

F.2 Upper Bound for Convex gradient-Hölder Functions

In this section we show a non trivial application of the theory presented so far. We will prove that our BAG algorithm is rate-optimal for approximating the level set of convex gradient-Lipschitz functions.

For the sake of simplicity, we will focus on the approximation of what we call *proper* level sets (Definition 10, below). Informally, a level set is proper if it is non-empty, bounded away from the set of minimizers (where the problem collapses into a simpler, standard minimization problem) and it is not cropped by the boundary of $[0,1]^d$.

Definition 10 (Proper level sets)**.** Fix any level $a \in \mathbb{R}$, a function $f\colon [0,1]^d \to \mathbb{R}$, and a margin $\Delta > 0$. We say that $\{f = a\}$ is a Δ-*proper* level set (for f), if $\{f = a\} \neq \varnothing$ and

$$\min_{\boldsymbol{x} \in [0,1]^d} f(\boldsymbol{x}) + \Delta \leq a \leq \min_{\boldsymbol{x} \in \partial[0,1]^d} f(\boldsymbol{x})\ ,$$

where we denoted by $\partial[0,1]^d$ the boundary of $[0,1]^d$. When we need not explicitly refer to the margin Δ, we simply say that $\{f = a\}$ is a *proper* level set.

In this section, we present an upper bound on the number of samples that our BAG algorithm needs in order to guarantee that its output is an approximation of the target level set of a convex gradient-Hölder function. As we discussed in Section F, now that we established a method on how to get these types of results, we only need to determine a NLS dimension $d^\star$ (Definition 9) of the level set of an arbitrary convex gradient-Hölder function. The following results shows that $d^\star = d - 1$.

Proposition 5. *Fix any level $a \in \mathbb{R}$, two Hölder constants $c > 0, \gamma \in (0,1]$, and an arbitrary convex (c,γ)-Hölder function $f\colon [0,1]^d \to \mathbb{R}$ with proper level set $\{f = a\}$. Then, there exists a constant $C^\star > 0$ such that*

$$\forall r \in (0,1)\ ,\ \mathcal{N}\Big(\big\{|f - a| \leq r\big\},\ r\Big) \leq C^\star\left(\frac{1}{r}\right)^{d-1}\ .$$

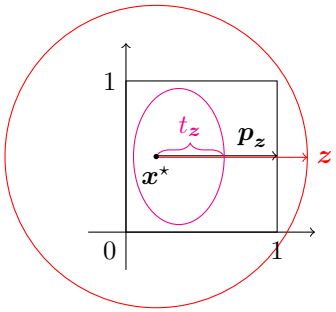

Figure 3: In black, the unit hypercube $[0,1]^d$; in red, the unit sphere centered at the minimizer $\boldsymbol{x}^\star$; in magenta, the level set $\{f = a\}$.

Proof. Let $\Delta > 0$ be a margin such that $\{f = a\}$ is Δ-proper. Fix any $r \in (0,1)$. If $r > \Delta/2$, we can simply apply Lemma 5 in Section B and use the lower bound on r to obtain

$$\mathcal{N}\Big(\big\{|f-a| \le r\big\},\ r\Big) \le 2^d \left(\frac{1}{r}\right)^d \le \frac{2^{d+1}}{\Delta}\left(\frac{1}{r}\right)^{d-1} .$$

Hence, without loss of generality, we can (and do) assume that $r \in (0, \Delta/2)$. In the following, we denote by $\mathcal{S}^{d-1}$ the $(d-1)$-dimensional unit sphere $\big\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \le 1\big\}$ with respect to the Euclidean norm $\|\cdot\|_2$. Let $\boldsymbol{x}^\star$ be a minimizer of f. Note that, being $\{f = a\}$ a proper level set (Definition 10), we have that $f(\boldsymbol{x}^\star) < a \le \min_{\boldsymbol{x}\in\partial[0,1]^d} f(\boldsymbol{x})$, therefore $\boldsymbol{x}^\star$ belongs to the interior $(0,1)^d$ of $[0,1]^d$. Now, for each $\boldsymbol{z} \in \mathcal{S}^{d-1}$, let $\boldsymbol{p}_{\boldsymbol{z}}$ be the unique element of $\partial[0,1]^d$ such that $(\boldsymbol{p}_{\boldsymbol{z}} - \boldsymbol{x}^\star)/\|\boldsymbol{p}_{\boldsymbol{z}} - \boldsymbol{x}_m\|_2 = \boldsymbol{z}$ (Figure 3) and define the convex univariate function

$$\begin{aligned} f_{\boldsymbol{z}} \colon \big[0, \|\boldsymbol{x}^\star - \boldsymbol{p}_{\boldsymbol{z}}\|_2\big] &\to \mathbb{R} \\ t &\mapsto f(\boldsymbol{x}^\star + t\boldsymbol{z}) . \end{aligned}$$

Being $\{f = a\}$ a proper level set, for all $\boldsymbol{z} \in \mathcal{S}^{d-1}$, the function $f_{\boldsymbol{z}}$ satisfies

$$\min_{\boldsymbol{x}\in[0,1]^d} f(\boldsymbol{x}) = f_{\boldsymbol{z}}(0) < a \le \min_{\boldsymbol{x}\in\partial[0,1]^d} f(\boldsymbol{x}) \le f_{\boldsymbol{z}}\big(\|\boldsymbol{x}^\star - \boldsymbol{p}_{\boldsymbol{z}}\|_2\big) .$$

Thus, for each $\boldsymbol{z} \in \mathcal{S}^{d-1}$, by the convexity and continuity of $f_{\boldsymbol{z}}$, there exists a unique value $t_{\boldsymbol{z}} \in \big[0, \|\boldsymbol{x}^\star - \boldsymbol{p}_{\boldsymbol{z}}\|_2\big]$ such that $f_{\boldsymbol{z}}(t_{\boldsymbol{z}}) = a$ (Figure 3), which we use to define the following function on the unit sphere

$$\begin{aligned} s \colon \mathcal{S}^{d-1} &\to \mathbb{R} \\ \boldsymbol{z} &\mapsto s(\boldsymbol{z}) := t_{\boldsymbol{z}} . \end{aligned}$$

In words, $t_{\boldsymbol{z}}$ is the distance between the minimizer $\boldsymbol{x}^\star$ and the level set $\{f = a\}$ in the direction of $\boldsymbol{z}$. We show now that s is Lipschitz with respect to the geodesic distance θ on $\mathcal{S}^{d-1}$, i.e., that there exists a constant $\ell > 0$ such that, for all $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{S}^{d-1}$,

$$\big|s(\boldsymbol{z}_1) - s(\boldsymbol{z}_2)\big| \le \ell\, \theta(\boldsymbol{z}_1, \boldsymbol{z}_2) ,$$

where $\theta(\boldsymbol{z}_1, \boldsymbol{z}_2) = \arccos\big(\langle \boldsymbol{z}_1, \boldsymbol{z}_2\rangle\big)$ is the angle between the two unit vectors $\boldsymbol{z}_1, \boldsymbol{z}_2$. Fix two arbitrary $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{S}^{d-1}$ with geodesic distance $\theta := \theta(\boldsymbol{z}_1, \boldsymbol{z}_2) \in (0, \pi]$. If $\theta \ge \pi/6$, we have

$$\frac{\big|s(\boldsymbol{z}_1) - s(\boldsymbol{z}_2)\big|}{\theta} \le \frac{6\sqrt{d}}{\pi}.$$

Assume now that $\theta < \pi/6$. Consider the two-dimensional plane containing the triangle with vertices $\boldsymbol{x}^\star$, $\boldsymbol{v}_1 := \boldsymbol{x}^\star + s(\boldsymbol{z}_1)\boldsymbol{z}_1$ and $\boldsymbol{v}_2 := \boldsymbol{x}^\star + s(\boldsymbol{z}_2)\boldsymbol{z}_2$ (note that the three points are not aligned). Let $\boldsymbol{v}$ be the

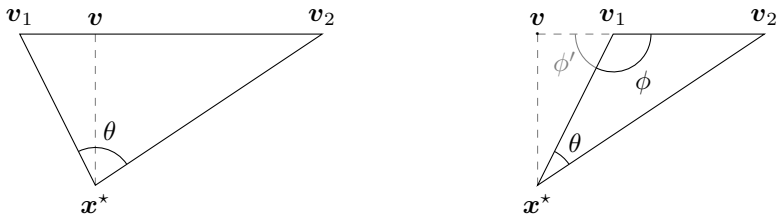

Figure 4: In the left (resp., right) picture, $\boldsymbol{v}$ belongs (resp., does not belong) to the segment $[\boldsymbol{v}_1, \boldsymbol{v}_2]$.

orthogonal projection of $\boldsymbol{x}^\star$ on the line containing $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Assume first that $\boldsymbol{v}$ belongs to the segment $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ (Figure 4, left). Then, the function

$$\begin{aligned} g_{\boldsymbol{v}_1,\boldsymbol{v}_2} \colon \mathbb{R} &\to [0, +\infty) \\ t &\mapsto g_{\boldsymbol{v}_1,\boldsymbol{v}_2}(t) := \Big\| \boldsymbol{x}^\star - \big(\boldsymbol{v}_1 + t(\boldsymbol{v}_2 - \boldsymbol{v}_1)\big) \Big\|_2^2 \end{aligned}$$

has its unique minimum at some $t^\star \in [0, 1]$. For all $t \in \mathbb{R}$, we have

$$\begin{aligned} g_{\boldsymbol{v}_1,\boldsymbol{v}_2}(t) &= \big\|(1-t)(\boldsymbol{v}_1 - \boldsymbol{x}^\star) + t(\boldsymbol{v}_2 - \boldsymbol{x}^\star)\big\|_2^2 \\ &= (1-t)^2 s(\boldsymbol{z}_1)^2 + t^2 s(\boldsymbol{z}_2)^2 + 2t(1-t)s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta) \\ &= t^2\big(s(\boldsymbol{z}_1)^2 + s(\boldsymbol{z}_2)^2 - 2s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta)\big) + t\big(2s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta) - 2s(\boldsymbol{z}_1)^2\big) + s(\boldsymbol{z}_1)^2 \,. \end{aligned}$$

The derivative of this function is given, for all $t \in \mathbb{R}$, by

$$g'_{\boldsymbol{v}_1,\boldsymbol{v}_2}(t) = 2t\big(s(\boldsymbol{z}_1)^2 + s(\boldsymbol{z}_2)^2 - 2s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta)\big) + \big(2s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta) - 2s(\boldsymbol{z}_1)^2\big) \,.$$

Hence we have

$$0 \leq t^\star = \frac{2s(\boldsymbol{z}_1)^2 - 2s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta)}{2\big(s(\boldsymbol{z}_1)^2 + s(\boldsymbol{z}_2)^2 - 2s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta)\big)}.$$

Since the above denominator is strictly positive, we obtain

$$2s(\boldsymbol{z}_1)^2 \geq 2s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\cos(\theta) \,,$$

thus, being $s(\boldsymbol{z}_1)$ and $s(\boldsymbol{z}_2)$ also strictly positive,

$$\cos(\theta) \leq 1 - \frac{s(\boldsymbol{z}_2) - s(\boldsymbol{z}_1)}{s(\boldsymbol{z}_2)}$$

and in turn, since $s(\boldsymbol{z}) \leq \sqrt{d}$ for all $\boldsymbol{z} \in \mathcal{S}^{d-1}$,

$$s(\boldsymbol{z}_2) - s(\boldsymbol{z}_1) \leq \sqrt{d}\big(1 - \cos(\theta)\big) \,.$$

Being $\theta > 0$, we have $1 - \cos(\theta) \leq \theta$ and thus

$$\frac{s(\boldsymbol{z}_2) - s(\boldsymbol{z}_1)}{\theta} \leq \sqrt{d} \,.$$

Swapping the roles of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ (i.e., considering the function $g_{\boldsymbol{v}_2,\boldsymbol{v}_1}$) we obtain similarly

$$\frac{s(\boldsymbol{z}_1) - s(\boldsymbol{z}_2)}{\theta} \leq \sqrt{d} \,.$$

Hence, when $\boldsymbol{v}$ belongs to the segment $[\boldsymbol{v}_1, \boldsymbol{v}_2]$, we obtained

$$\frac{\big|s(\boldsymbol{z}_1) - s(\boldsymbol{z}_2)\big|}{\theta} \leq \sqrt{d}\,.$$

Consider now the last case where $\boldsymbol{v}$ does not belong to the segment $[\boldsymbol{v}_1, \boldsymbol{v}_2]$ (Figure 4, right). Without loss of generality, we can (and do) assume that $s(\boldsymbol{z}_2) > s(\boldsymbol{z}_1)$, and thus that $\boldsymbol{v}$ is closer to $\boldsymbol{v}_1$ than to $\boldsymbol{v}_2$. By convexity of f on the line containing $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, we have $f(\boldsymbol{v}) \geq a$. Using the fact that the level set $\{f = a\}$ is Δ-proper and the (c, γ)-Hölderness of f, we get

$$\Delta \leq a - \min_{\boldsymbol{x} \in [0,1]^d} f(\boldsymbol{x}) \leq f(\boldsymbol{v}) - f(\boldsymbol{x}^\star) \leq c\|\boldsymbol{v} - \boldsymbol{x}^\star\|_\infty^\gamma \leq c\|\boldsymbol{v} - \boldsymbol{x}^\star\|_2^\gamma\,,$$

which in turn implies

$$\|\boldsymbol{v} - \boldsymbol{x}^\star\|_2 \geq \left(\frac{\Delta}{c}\right)^{1/\gamma}. \tag{23}$$

Let ϕ be the angle between $\boldsymbol{x}^\star - \boldsymbol{v}_1$ and $\boldsymbol{v}_2 - \boldsymbol{v}_1$. Applying the sine rule to the triangle $\boldsymbol{x}^\star, \boldsymbol{v}_1, \boldsymbol{v}_2$, we obtain

$$\frac{\sin(\theta)}{\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2} = \frac{\sin(\phi)}{s(\boldsymbol{z}_2)}$$

and thus

$$\sin(\phi) = \frac{s(\boldsymbol{z}_2)\sin(\theta)}{\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2}\,. \tag{24}$$

Let $\phi' = \pi - \phi$ be the angle between $\boldsymbol{x}^\star - \boldsymbol{v}_1$ and $\boldsymbol{v} - \boldsymbol{v}_1$. Note that the angle between $\boldsymbol{x}^\star - \boldsymbol{v}$ and $\boldsymbol{v}_1 - \boldsymbol{v}$ is $\pi/2$, being $\boldsymbol{v}$ the orthogonal projection of $\boldsymbol{x}^\star$ on the line containing $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$. Hence, applying the sine rule to the triangle $\boldsymbol{x}^\star, \boldsymbol{v}_1, \boldsymbol{v}$ we obtain

$$\frac{\sin(\phi')}{\|\boldsymbol{v} - \boldsymbol{x}^\star\|_2} = \frac{\sin(\pi/2)}{s(\boldsymbol{z}_1)}$$

and thus

$$\|\boldsymbol{v} - \boldsymbol{x}^\star\|_2 = s(\boldsymbol{z}_1)\sin(\phi)\,. \tag{25}$$

From (23), (25), and (24), we obtain

$$\frac{s(\boldsymbol{z}_1)s(\boldsymbol{z}_2)\sin(\theta)}{\|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2} \geq \left(\frac{\Delta}{c}\right)^{1/\gamma}.$$

The triangle inequality yields

$$\big|s(\boldsymbol{z}_1) - s(\boldsymbol{z}_2)\big| = \big|\big\|\boldsymbol{v}_1 - \boldsymbol{x}^\star\big\|_2 - \big\|\boldsymbol{v}_2 - \boldsymbol{x}^\star\big\|_2\big| \leq \big\|(\boldsymbol{v}_1 - \boldsymbol{x}^\star) - (\boldsymbol{v}_2 - \boldsymbol{x}^\star)\big\|_2 = \|\boldsymbol{v}_1 - \boldsymbol{v}_2\|_2$$

and thus

$$\big|s(\boldsymbol{z}_1) - s(\boldsymbol{z}_2)\big| \leq \left(\frac{c}{\Delta}\right)^{1/\gamma} d\sin(\theta) \leq \left(\frac{c}{\Delta}\right)^{1/\gamma} d\theta\,,$$

where we used again $s(\boldsymbol{z}) \leq d^{1/2}$ for any $\boldsymbol{z} \in \mathcal{S}^{d-1}$. Putting everything together, we have shown that

$$\frac{\big|s(\boldsymbol{z}_1) - s(\boldsymbol{z}_2)\big|}{\theta} \leq \max\left(\frac{6}{\pi}\sqrt{d},\ \left(\frac{c}{\Delta}\right)^{1/\gamma} d\right) := \ell\,, \tag{26}$$

for all $\theta \in (0, \pi]$, i.e., that s is ℓ-Lipschitz on $\mathcal{S}^{d-1}$ with respect to the geodesic distance.

Consider now a covering of $\mathcal{S}^{d-1}$ with respect to the geodesic distance, with radius βr, where $\beta \in (0, 1]$ will be selected later. This is a set of points $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n \in \mathcal{S}^{d-1}$ such that the union of all the balls (with respect to the geodesic distance θ) with radius βr centered at these points contains the whole $\mathcal{S}^{d-1}$. We show now how

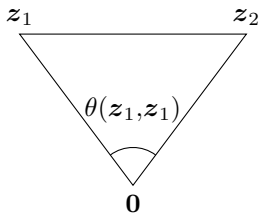

Figure 5: The isosceles triangle $\boldsymbol{z}_1, \mathbf{0}, \boldsymbol{z}_2$.

such a covering can be taken using order of $1/r^{d-1}$ points. Fix any two distinct $\boldsymbol{z}_1, \boldsymbol{z}_2 \in \mathcal{S}^{d-1}$ with geodesic distance $\theta(\boldsymbol{z}_1, \boldsymbol{z}_2) \in (0, \pi/2]$ and consider the isosceles triangle $\boldsymbol{z}_1, \mathbf{0}, \boldsymbol{z}_2$ with angles $\angle(\boldsymbol{z}_1\, \mathbf{0}\, \boldsymbol{z}_2) = \theta(\boldsymbol{z}_1, \boldsymbol{z}_2)$ and $\angle(\mathbf{0}\, \boldsymbol{z}_2\, \boldsymbol{z}_1) = \angle(\boldsymbol{z}_2\, \boldsymbol{z}_1\, \mathbf{0}) = \big(\pi - \theta(\boldsymbol{z}_1, \boldsymbol{z}_2)\big)/2 = \pi/2 - \theta(\boldsymbol{z}_1, \boldsymbol{z}_2)/2$ (Figure 5). The sine rule yields

$$\frac{\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2}{\sin\big(\theta(\boldsymbol{z}_1, \boldsymbol{z}_2)\big)} = \frac{1}{\cos\big(\theta(\boldsymbol{z}_1, \boldsymbol{z}_2)/2\big)}$$

or, equivalently stated,

$$\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 = \frac{\sin\big(\theta(\boldsymbol{z}_1, \boldsymbol{z}_2)\big)}{\cos\big(\theta(\boldsymbol{z}_1, \boldsymbol{z}_2)/2\big)} \,.$$

Using the fact that $\sin(x) \geq (2/\pi)x$, for all $x \in [0, \pi/2]$, the equality above gives

$$\|\boldsymbol{z}_1 - \boldsymbol{z}_2\|_2 \geq \sin\big(\theta(\boldsymbol{z}_1, \boldsymbol{z}_2)\big) \geq \frac{2}{\pi}\, \theta(\boldsymbol{z}_1, \boldsymbol{z}_2) \,.$$

Therefore, if $x \leq \pi/2$, each ball with center $\boldsymbol{c}$ radius $(2/\pi)\rho$ with respect to the Euclidean distance is included in the corresponding ball with center $\boldsymbol{c}$ and radius ρ with respect to the geodesic distance. Thus, being $r < 1 \leq \pi/2$, in order to cover $\mathcal{S}^{d-1}$ with balls with radius βr with respect to the geodesic distance, it is enough to cover $\mathcal{S}^{d-1}$ with balls with radius $(2/\pi)\beta r$ with respect to the Euclidean distance. Moreover, since for any two points $\boldsymbol{x}, \boldsymbol{y} \in \partial[-1,1]^d$ on the boundary of the hypercube $[-1,1]^d$, their Euclidean distance $\|\boldsymbol{x} - \boldsymbol{y}\|_2$ is larger than the Euclidean distance $\big\|\boldsymbol{x}/\|\boldsymbol{x}\|_2 - \boldsymbol{y}/\|\boldsymbol{y}\|_2\big\|_2$ between their projections on the unit sphere, and since any point in the unit sphere can be reached this way, in order to cover the unit sphere with balls with radius $(2/\pi)\beta r$ with respect to the Euclidean distance it is sufficient to cover the boundary $\partial[-1,1]^d$ of $[-1,1]^d$ with balls with radius $(2/\pi)\beta r$ with respect to the Euclidean distance. This is easy to do, as each one of the $2d$ faces $\big\{[-1,1] \times \ldots \times [-1,1] \times \{-1,1\} \times [-1,1] \times \ldots \times [-1,1]\big\}$ of $\partial[-1,1]^d$ can be covered with the same number of balls of radius $(2/\pi)\beta r$ with respect to the $(d-1)$-dimensional Euclidean distance that cover the hypercube $[-1,1]^{d-1}$. This can be done, e.g., by taking a uniform grid of $(2/\pi)\beta r$-spaced points. Projecting these points onto $\mathcal{S}^{d-1}$ gives a covering $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$ of $\mathcal{S}^{d-1}$ with respect to the geodesic distance, with radius βr, and with a number of points n that is at most

$$n \leq 2d\bigg(1 + \bigg\lceil \frac{\pi}{2\beta r} \bigg\rceil\bigg)^{d-1} \leq 2d\bigg(2 + \frac{\pi}{2\beta r}\bigg)^{d-1} \leq 2d\bigg(\frac{3}{2}\pi\bigg)^{d-1}\bigg(\frac{1}{\beta r}\bigg)^{d-1} \,. \tag{27}$$

Fix this covering $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_n$. Fix also an arbitrary $\boldsymbol{x} \in \big\{|f - a| \leq r|\big\}$. Note that, being $r \leq \Delta/2$ and $\{f = a\}$ a Δ-proper level set, then the minimizer $\boldsymbol{x}^\star$ cannot belong to the set $\big\{|f - a| \leq r|\big\}$, hence $\boldsymbol{x} \neq \boldsymbol{x}^\star$. Let $\boldsymbol{z} = (\boldsymbol{x} - \boldsymbol{x}^\star)/\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2$. Similarly as before, define for all $t \in \big[0, \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big]$, the function $f_{\boldsymbol{z}}(t) := f(\boldsymbol{x}^\star + t\boldsymbol{z})$. Then $f_{\boldsymbol{z}}$ is convex, $f_{\boldsymbol{z}}(0) = f(\boldsymbol{x}^\star)$, $f_{\boldsymbol{z}}\big(s(\boldsymbol{z})\big) = a$ and $\big|f_{\boldsymbol{z}}\big(\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big) - a\big| \leq r$. If $f_{\boldsymbol{z}}\big(\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big) < a$, by convexity, we have $s(\boldsymbol{z}) > \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2$, hence

$$\frac{a - r - f(\boldsymbol{x}^\star)}{\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2} \leq \frac{f_{\boldsymbol{z}}\big(\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big) - f_{\boldsymbol{z}}(0)}{\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 - 0} \leq \frac{a - f_{\boldsymbol{z}}\big(\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big)}{s(\boldsymbol{z}) - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2} \leq \frac{r}{s(\boldsymbol{z}) - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2}$$

and recalling that $r \leq \Delta/2$ so that $a - r - f(\boldsymbol{x}^\star) \geq \Delta/2 > 0$, we have

$$s(\boldsymbol{z}) - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 \leq r \frac{\sqrt{d}}{a - r - f(\boldsymbol{x}^\star)} \leq \left(2\frac{\sqrt{d}}{\Delta}\right) r \ ,$$

where we used $\|x - x_m\|_2 \leq \sqrt{d}$. If $f_{\boldsymbol{z}}\big(\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big) \geq a$, proceed similarly. By convexity of $f_{\boldsymbol{z}}$ we have $s(\boldsymbol{z}) \leq \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2$. If $s(\boldsymbol{z}) = \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2$, then trivially $\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 - s(\boldsymbol{z}) = 0 \leq \big(2\sqrt{d}/\Delta\big) r$. If on the other hand, $s(\boldsymbol{z}) < \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2$, using the convexity of $f_{\boldsymbol{z}}$ once again, we get

$$\frac{a - f_{\boldsymbol{z}}(0)}{s(\boldsymbol{z}) - 0} \leq \frac{f_{\boldsymbol{z}}\big(\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big) - a}{\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 - s(\boldsymbol{z})} \leq \frac{r}{\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 - s(\boldsymbol{z})}$$

and using $a - f(\boldsymbol{x}^\star) \geq \Delta > 0$ and $s(\boldsymbol{z}) \leq \sqrt{d}$, yields

$$\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 - s(\boldsymbol{z}) \leq \left(\frac{\sqrt{d}}{\Delta}\right) r \leq \left(2\frac{\sqrt{d}}{\Delta}\right) r \ .$$

Thus we proved that

$$\big|s(\boldsymbol{z}) - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big| \leq \left(2\frac{\sqrt{d}}{\Delta}\right) r \ . \tag{28}$$

Furthermore, there exists $i \in \{1, \ldots, n\}$ such the geodesic distance of $\boldsymbol{z}_i$ and $\boldsymbol{z}$ is smaller than or equal to βr. Therefore we have, from (28), and the ℓ-Lipschitzness of the function r with respect to the geodesic distance,

$$\big|s(\boldsymbol{z}_i) - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big| \leq \big|s(\boldsymbol{z}_i) - s(\boldsymbol{z})\big| + \big|s(\boldsymbol{z}) - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big| \leq (\ell\beta)\, r + \left(2\frac{\sqrt{d}}{\Delta}\right) r \ .$$

Hence, with $\gamma > 0$ to be chosen later, there exists

$$\boldsymbol{x}_i' := \boldsymbol{x}^\star + s(\boldsymbol{z}_i)\boldsymbol{z}_i + k\gamma r \boldsymbol{z}_i \ ,$$

with $k \in \mathbb{Z}$ such that

$$|k| \leq \frac{\ell\beta + \frac{2\sqrt{d}}{\Delta}}{\gamma}$$

and with

$$\big|\|\boldsymbol{x}_i' - \boldsymbol{x}^\star\|_2 - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big| \leq \gamma r \ . \tag{29}$$

This is obtained by covering the segment $[-\ell\beta r - 2rd^{1/2}/\Delta,\ \ell\beta r + 2rd^{1/2}/\Delta]$ with points with equidistance γr. Then, we obtain

$$\begin{aligned}
\|\boldsymbol{x} - \boldsymbol{x}_i'\|_\infty &\leq \|\boldsymbol{x} - \boldsymbol{x}_i'\|_2 \\
&= \Big\| \big(\boldsymbol{x}^\star + \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\, \boldsymbol{z}\big) - \big(\boldsymbol{x}^\star + \|\boldsymbol{x}_i' - \boldsymbol{x}^\star\|_2\, \boldsymbol{z}_i\big) \Big\|_2 \\
&= \big\| \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\, \boldsymbol{z} - \|\boldsymbol{x}_i' - \boldsymbol{x}^\star\|_2\, \boldsymbol{z}_i \big\|_2 \\
&= \Big\| \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 (\boldsymbol{z} - \boldsymbol{z}_i) \\
&\qquad\qquad - \big(\|\boldsymbol{x}_i' - \boldsymbol{x}^\star\|_2 - \|\boldsymbol{x} - \boldsymbol{x}^\star\|_2\big)\boldsymbol{z}_i \Big\|_2 \\
&\leq \big|\|\boldsymbol{x} - \boldsymbol{x}^\star\|_2 - \|\boldsymbol{x}_i' - \boldsymbol{x}^\star\|_2\big| + \sqrt{d}\|\boldsymbol{z} - \boldsymbol{z}_i\|_2 \\
&\leq \big(\gamma + \sqrt{d}\beta\big) r \ ,
\end{aligned}$$

from (29). Hence, with

$$n' \leq 2d\left(\frac{3}{2}\pi\right)^{d-1}\left(\frac{1}{\beta r}\right)^{d-1}\left(1+2\frac{\ell\beta+\frac{2\sqrt{d}}{\Delta}}{\gamma}\right)$$

points, we have obtained a covering of $\big\{|f-a|\leq r\big\}$ with radius $\big(\gamma+\sqrt{d}\beta\big)r$ with respect to the sup-norm $\|\cdot\|_\infty$. Choosing $\beta := 1/\big(4\sqrt{d}\big)$ and $\gamma := 1/4$ so that that $\big(\gamma+\sqrt{d}\beta\big) \leq 1/2$, we therefore determined a covering of $\big\{|f-a|\leq r\big\}$ with radius $r/2$ with respect to the sup-norm consisting of n' elements. Thus, n' is greater than or equal to the smallest cardinality $\mathcal{M}\big(\big\{|f-a|\leq r\big\},\ r/2\big)$ of a covering of $\big\{|f-a|\leq r\big\}$ with radius $r/2$ with respect to the sup-norm. For a known result relating pickings and coverings (we recall it in (12), Section B), we have

$$\mathcal{M}\Big(\big\{|f-a|\leq r\big\},\ r/2\Big) \geq \mathcal{N}\Big(\big\{|f-a|\leq r\big\},\ r\Big)\ ,$$

which concludes the proof. □

Theorem 7. *Consider the BAG algorithm (Algorithm 4) run with input a, c_1, γ_1. Let $f\colon [0,1]^d \to \mathbb{R}$ be an arbitrary convex (c_1,γ_1)-gradient-Hölder function with proper level set $\{f=a\}$. Fix any accuracy $\varepsilon > 0$. Then, for all*

$$n > \kappa\,\frac{1}{\varepsilon^{(d-1)/(1+\gamma_1)}}$$

the output S_n returned after the n-th query is an ε-approximation of $\{f=a\}$, where $\kappa > 0$ is a constant independent of ε that depends exponentially on the dimension d.

Proof. Being f the restriction of a differentiable function defined on on an open set containing $[0,1]^d$, it is Lipschitz on the compact $[0,1]^d$. Thus we can apply Proposition 5 to get a NLS dimension $d^\star = d-1$ for f. The result then follows directly from Corollary 1. □

F.3 Rate-optimal Sample Complexity for Convex Gradient-Lipschitz Functions

Theorem 7 applied to the special case of gradient-Lipschitz functions, states that the BAG algorithm (Algorithm 4) needs order of $1/\varepsilon^{(d-1)/2}$ queries to reliably output an ε-approximation of a gradient-Lipschitz function. The following theorem shows that this rate cannot be improved, i.e., that BAG is rate-optimal (Definition 8) for determining proper level sets of gradient-Lipschitz functions.

Theorem 8. *Fix any level $a \in \mathbb{R}$ and an arbitrary accuracy $\varepsilon > 0$. No deterministic algorithm A can guarantee to output an ε-approximation of any Δ-proper level set $\{f=a\}$ of an arbitrary convex c_1-gradient-Lipschitz functions f with $c_1 \geq 3$ and $\Delta \in (0, 1/4]$, querying less than $\kappa/\varepsilon^{(d-1)/2}$ of their values, where $\kappa > 0$ is a constant independent of ε. This implies in particular that (recall Definition 7),*

$$\inf_A \sup_f \mathfrak{n}(f, A, \varepsilon, a) \geq \kappa\,\frac{1}{\varepsilon^{(d-1)/2}}\ ,$$

where the inf *is over all deterministic algorithms A and the* sup *is over all c_1-gradient-Lipschitz functions f with Δ-proper level set $\{f=a\}$, with $c_1 \geq 3$ and $\Delta \in (0, 1/4]$.*

Proof. We will prove the equivalent statement that no algorithm can output a $(\kappa'\varepsilon)$-approximation of any Δ-proper level set $\{f=a\}$ of an arbitrary c_1-gradient-Lipschitz functions f with $c_1 \geq 3$ and $\Delta \in (0, 1/4]$, querying less than $1/\varepsilon^{(d-1)/2}$ of their values, where $\kappa' > 0$ is a constant independent of ε.

Let $\boldsymbol{o} := (1/2, \dots, 1/2) \in [0,1]^d$, $\boldsymbol{o}_1 := (1/2 + 1/(4d)^{1/2}, \dots, 1/2 + 1/(4d)^{1/2}) \in (0,1)^d$, and

$$\begin{aligned} f_0\colon [0,1]^d &\to \mathbb{R} \\ \boldsymbol{x} &\mapsto f_0(\boldsymbol{x}) := a - \frac{1}{4} + \|\boldsymbol{x}-\boldsymbol{o}\|_2^2\ . \end{aligned}$$

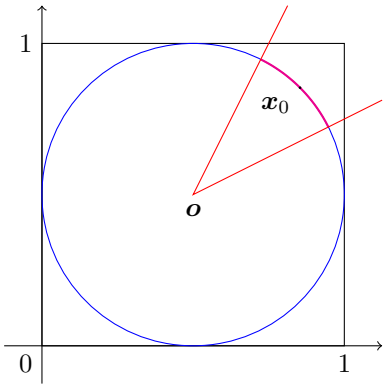

Figure 6: The blue circle is the level set $\mathcal{S}$; the geodesic ball on $\mathcal{S}$ with center $\boldsymbol{x}_0$ and radius $\kappa_1\varepsilon^{1/2}$ is the arc in magenta and the corresponding cone is in red.

Then f_0 is the restriction to $[0,1]^d$ of the differentiable function $\boldsymbol{x} \mapsto a - \frac{1}{4} + \|\boldsymbol{x} - \boldsymbol{o}\|_2^2$ defined on $\mathbb{R}^d$, and it satisfies, for all $\boldsymbol{x}, \boldsymbol{y} \in [0,1]^d$

$$\begin{aligned} \left\|\nabla f_0(\boldsymbol{x}) - \nabla f_0(\boldsymbol{y})\right\|_\infty &= \left\|2(\boldsymbol{x} - \boldsymbol{o}) - 2(\boldsymbol{y} - \boldsymbol{o})\right\|_\infty \\ &= 2\left\|\boldsymbol{x} - \boldsymbol{y}\right\|_\infty , \end{aligned} \tag{30}$$

i.e., it is 2-gradient-Lipschitz. Moreover, f_0 has minimum equal to $a - 1/4$ at $\boldsymbol{o}$ and satisfies $f_0(\boldsymbol{o}_1) = a$. Also, the minimum of f_0 over $\partial[0,1]^d$ is equal to $a + (1/2)^2 - 1/4 = a$. Hence $\{f_0 = a\}$ is a Δ-proper level set, with $\Delta = 1/4$.

Consider an arbitrary deterministic algorithm A applied to the level set $\{f_0 = a\}$ of f_0 and assume that only $n < 1/\varepsilon^{(d-1)/2}$ values are queried before outputting a set S_n.

Let $\mathcal{S}$ be the Euclidean sphere with center $\boldsymbol{o}$ and radius $\|\boldsymbol{o} - \boldsymbol{o}_1\|_2 = 1/2$ (Figure 6). Note that $\{f_0 = a\} = \mathcal{S}$. For any constant $\kappa_1 > 0$ and each point $\boldsymbol{x}_0$ in $\mathcal{S}$, consider the convex cone having origin $\boldsymbol{o}$, and with intersection with $\mathcal{S}$ equal to the geodesic ball on $\mathcal{S}$ with center $\boldsymbol{x}_0$ and radius $\kappa_1\varepsilon^{1/2}$. Then we can choose κ_1 (small enough) and $\boldsymbol{x}_0$ such that this cone does not contain any points of f_0 queried by the algorithm. Fix such a κ_1. If S_n does not contain $\boldsymbol{x}_0$ then, since $f_0(\boldsymbol{x}_0) = a$, we have shown that $\{f_0 = a\} \not\subset S_n$, and the result follows.

Assume now that $\boldsymbol{x}_0 \in S_n$. We will define a function $f_1 \colon [0,1]^d \to \mathbb{R}$ such that the sum $f_0 + f_1$ is convex and 3-gradient-Lipschitz, the level set $\{f_0 + f_1 = a\}$ is Δ-proper, and the algorithm applied to the level set $\{f_0 + f_1 = a\}$ of $f_0 + f_1$ does not return a $(\kappa'\varepsilon)$-approximation of $\{f_0 + f_1 = a\}$. The idea is to carefully design a function f_1 that is non-zero only on the cone that has not been explored by the algorithm. This way, we can make $f_0 + f_1$ a perturbation of f_0 that is not far enough from f_0 so that the algorithm can distinguish the two, but it is different enough so that no $(\kappa'\varepsilon)$-approximation of $\{f_0 = a\}$ can be a $(\kappa'\varepsilon)$-approximation of $\{f_0 + f_1 = a\}$. The subtle part is that by construction, such an f_1 is not convex, but the sum $f_0 + f_1$ has to retain the convexity of f_0.

We begin by defining three non-negative auxiliary functions $\phi_1, \phi_2, \phi_3 \colon [0,1] \to \mathbb{R}$, for all $t \in [0,1]^d$, by

$$\phi_1(t) := \begin{cases} t & \text{if } t \in [0, 1/4] \\ 1/4 - (t - 1/4) & \text{if } t \in [1/4, 3/4] \\ -1/4 + (t - 3/4) & \text{if } t \in [3/4, 1] \end{cases} ,$$

$\phi_2(t) := \int_0^t \mathrm{d}x \int_0^x \phi_1(u)\,\mathrm{d}u$, and $\phi_3(t) := \phi_2(1-t)$. We remark that ϕ_2 is twice differentiable with second derivative ϕ_1. We see that $\phi_2(0) = 0$, $\phi_2'(0) = 0$ and $\phi_2''(0) = 0$. We see that ϕ_2' is strictly positive on $[0,1]$. Hence, $\kappa_2 := \phi_2(1) > 0$. We also see that $\phi_2'(1) = 0$ and $\phi_2''(1) = 0$. Then ϕ_3 is twice differentiable and non-negative on $[0,1]$ and satisfies $\phi_3''(0) = 0$, $\phi_3''(1) = 0$, $\phi_3'(0) = 0$, $\phi_3'(1) = 0$, $\phi_3(0) = \kappa_2 > 0$ and $\phi_3(1) = 0$.

We write $B(\boldsymbol{x}, r)$ for the closed Euclidean ball with center $\boldsymbol{x}$ and radius r intersected with $[0,1]^d$. We define the function $f_1 \colon [0,1]^d \to \mathbb{R}$, for all $\boldsymbol{x} \in [0,1]^d$, by

$$f_1(\boldsymbol{x}) := \begin{cases} \beta\varepsilon\, \phi_3\Big(\dfrac{\|\boldsymbol{x}-\boldsymbol{x}_0\|_2}{\kappa_3\varepsilon^{1/2}}\Big) & \text{if } \boldsymbol{x} \in B(\boldsymbol{x}_0, \kappa_3\varepsilon^{1/2}) \\ 0 & \text{otherwise}\,, \end{cases}$$

with $\kappa_3, \beta > 0$ to be selected later. We can find $\kappa_3 > 0$ small enough such that $B(\boldsymbol{x}_0, \kappa_3\varepsilon^{1/2})$ is included in the cone discussed above (recall Figure 6). Fix such a κ_3. Then, f_0 and $f_0 + f_1$ differ only on this cone which is not explored by the algorithm. As a consequence, the algorithm applied to $f_0 + f_1$ returns the same set S_n, which contains $\boldsymbol{x}_0$. Since $f_0(\boldsymbol{x}_0) + f_1(\boldsymbol{x}_0) = a + \beta\varepsilon\kappa_2$, the proof will be completed (letting $\kappa' := \beta\kappa_2/2$) once we show that we can select $\beta > 0$, independently of ε, such that $f_0 + f_1$ is a convex 3-gradient-Lipschitz function with Δ-proper level set $\{f_0 + f_1 = a\}$, where $\Delta = 1/4$.

Because of the above discussed inclusion of the ball in the cone, we have $f_1(\boldsymbol{o}) = 0$. Hence

$$\min_{\boldsymbol{x}\in[0,1]^d}\big(f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})\big) \le f_0(\boldsymbol{o}) + f_1(\boldsymbol{o}) = a - \frac{1}{4} \le a = \min_{\boldsymbol{x}\in\partial[0,1]^d} f_0(\boldsymbol{x}) \le \min_{\boldsymbol{x}\in\partial[0,1]^d}\big(f_0(\boldsymbol{x}) + f_1(\boldsymbol{x})\big)\,,$$

which proves that the level set $\{f_0 + f_1 = a\}$ is Δ-proper, with $\Delta = 1/4$.

By definition of f_1, its gradient is, for $\boldsymbol{x} \in [0,1]^d$,

$$\nabla f_1(\boldsymbol{x}) = \beta\varepsilon\phi_3'\Big(\frac{\|\boldsymbol{x}-\boldsymbol{x}_0\|_2}{\kappa_3\varepsilon^{1/2}}\Big)\frac{1}{\kappa_3\varepsilon^{1/2}}\frac{\boldsymbol{x}-\boldsymbol{x}_0}{\|\boldsymbol{x}-\boldsymbol{x}_0\|_2}$$

if $\boldsymbol{x} \in B(\boldsymbol{x}_0, \kappa_3\varepsilon^{1/2})$, 0 otherwise. We remark that in the above formula, by convention, $\nabla f_1(\boldsymbol{x}_0) = 0$, which follows from the properties of ϕ_3. Next, we observe that ∇f_1 satisfies

$$\sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\big\|\nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v})\big\|_2}{\|\boldsymbol{u}-\boldsymbol{v}\|_2} \le \sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in B(\boldsymbol{x}_0,\kappa_3\varepsilon^{1/2}) \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\big\|\nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v})\big\|_2}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}\,,$$

Indeed, for $\boldsymbol{u}, \boldsymbol{v} \notin B(\boldsymbol{x}_0, \kappa_3\varepsilon^{1/2})$ the gradient difference is zero while for $\boldsymbol{u} \in B(\boldsymbol{x}_0, \kappa_3\varepsilon^{1/2})$ and $\boldsymbol{v} \notin B(\boldsymbol{x}_0, \kappa_3\varepsilon^{1/2})$ the gradient difference is equal to the difference between the gradient at $\boldsymbol{u}$ and the gradient at the intersection of the segment $[u, v]$ and the boundary $\partial B(\boldsymbol{x}_0, \kappa_3\varepsilon^{1/2})$. Hence,

$$\begin{aligned}
&\sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\big\|\nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v})\big\|_2}{\|\boldsymbol{u}-\boldsymbol{v}\|_2} \\
&\quad\le \sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in B(\boldsymbol{x}_0,\kappa_3\varepsilon^{1/2}) \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\beta\varepsilon}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}\bigg\|\phi_3'\Big(\frac{\|\boldsymbol{u}-\boldsymbol{x}_0\|_2}{\kappa_2\varepsilon^{1/2}}\Big)\frac{1}{\kappa_3\varepsilon^{1/2}}\frac{\boldsymbol{u}-\boldsymbol{x}_0}{\|\boldsymbol{u}-\boldsymbol{x}_0\|_2} - \phi_3'\Big(\frac{\|\boldsymbol{v}-\boldsymbol{x}_0\|_2}{\kappa_2\varepsilon^{1/2}}\Big)\frac{1}{\kappa_3\varepsilon^{1/2}}\frac{\boldsymbol{v}-\boldsymbol{x}_0}{\|\boldsymbol{v}-\boldsymbol{x}_0\|_2}\bigg\|_2 \\
&\quad= \sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in B(\boldsymbol{0},1) \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\frac{\beta}{\kappa_3^2}\Big\|\phi_3'\big(\|\boldsymbol{u}\|_2\big)\frac{\boldsymbol{u}}{\|\boldsymbol{u}\|_2} - \phi_3'\big(\|\boldsymbol{v}\|_2\big)\frac{\boldsymbol{v}}{\|\boldsymbol{v}\|_2}\Big\|_2}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}\,.
\end{aligned}$$

Letting $\tilde{f}_1 \colon B(\boldsymbol{0}, 1) \to \mathbb{R}$ be defined for all $t \in B(\boldsymbol{0}, 1)$, by $\tilde{f}_1(\boldsymbol{x}) = \phi_3\big(\|\boldsymbol{x}\|_2\big)$, we obtain

$$\sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\big\|\nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v})\big\|_2}{\|\boldsymbol{u}-\boldsymbol{v}\|_2} \le \sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in B(\boldsymbol{0},1) \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\frac{\beta}{\kappa_3^2}\big\|\nabla\tilde{f}_1(\boldsymbol{u}) - \nabla\tilde{f}_1(\boldsymbol{v})\big\|_2}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}\,.$$

Since $\tilde{f}_1$ is a fixed twice differentiable function which does not depend on ε, we can choose $\beta > 0$ small enough, independently of ε, such that

$$\sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\left\|\nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v})\right\|_2}{\|\boldsymbol{u}-\boldsymbol{v}\|_2} \leq \frac{1}{\sqrt{d}} \,.$$

This implies that, for all $\boldsymbol{u}, \boldsymbol{v} \in [0,1]^d$,

$$\left\|\nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v})\right\|_\infty \leq \left\|\nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v})\right\|_2 \leq \frac{1}{\sqrt{d}}\|\boldsymbol{u}-\boldsymbol{v}\|_2 \leq \|\boldsymbol{u}-\boldsymbol{v}\|_\infty \,. \tag{31}$$

Thus, the two bounds (30) and (31) yield

$$\sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\left\|\nabla(f_0+f_1)(\boldsymbol{u}) - \nabla(f_0+f_1)(\boldsymbol{v})\right\|_\infty}{\|\boldsymbol{u}-\boldsymbol{v}\|_\infty} \leq 3 \,.$$

Therefore, $f_0 + f_1$ is 3-gradient-Lipschitz. Finally, we have

$$\begin{aligned}
&\inf_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\left\langle \nabla(f_0+f_1)(\boldsymbol{u}) - \nabla(f_0+f_1)(\boldsymbol{v}), \frac{\boldsymbol{u}-\boldsymbol{v}}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}\right\rangle}{\|\boldsymbol{u}-\boldsymbol{v}\|_2} \\
&\qquad \geq \inf_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\left\langle \nabla f_0(\boldsymbol{u}) - \nabla f_0(\boldsymbol{v}), \frac{\boldsymbol{u}-\boldsymbol{v}}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}\right\rangle}{\|\boldsymbol{u}-\boldsymbol{v}\|_2} - \sup_{\substack{\boldsymbol{u},\boldsymbol{v}\in[0,1]^d \\ \boldsymbol{u}\neq\boldsymbol{v}}} \frac{\left\langle \nabla f_1(\boldsymbol{u}) - \nabla f_1(\boldsymbol{v}), \frac{\boldsymbol{u}-\boldsymbol{v}}{\|\boldsymbol{u}-\boldsymbol{v}\|_2}\right\rangle}{\|\boldsymbol{u}-\boldsymbol{v}\|_2} \geq 2 - \frac{1}{\sqrt{d}} \geq 1 \,.
\end{aligned}$$

Hence $f_0 + f_1$ is 1-strongly convex and thus it is convex. In conclusion, we have eventually selected a constant $\beta > 0$, independent of ε, such that $f_0 + f_1$ is a convex 3-gradient-Lipschitz function with Δ-proper level set $\{f_0 + f_1 = a\}$, but S_n is not a $(\kappa'\varepsilon)$-approximation of $\{f_0 + f_1 = a\}$. This concludes the proof. □

We conclude this section by remarking the analogy between the problem of approximating the level set of a convex function and that of determining an approximation of a convex body in Hausdorff distance. The latter problem has been studied extensively in convex geometry. Notably, while the scope of the and the techniques used in this field differ from ours, the sample complexity results for the two problems are similar. For an overview of these results, we refer the reader to the two surveys (Kamenev, 2019; Gruber, 1993).