Adaptive approximation of monotone functions

Pierre Gaillard, Sébastien Gerchinovitz, Étienne de Montbrun

Adaptive approximation of monotone functions

Pierre Gaillard[1*†], Sébastien Gerchinovitz[2,3†] and Étienne de Montbrun[4†]

[1]Univ. Grenoble Alpes, INRIA, CNRS, Grenoble INP, LJK, Grenoble, 38000, France.

[2]DEEL, IRT Saint Exupéry, 3 rue Tarfaya, Toulouse, 31400, France.

[3]UMR5219, Institut Mathématiques de Toulouse, 118 route de Narbonne, Toulouse, 31400, France.

[4] TSE, 1, Esplanade de l'Université, Toulouse, 31000, France.

*Corresponding author(s). E-mail(s): pierre.gaillard@inria.fr;

Contributing authors: sebastien.gerchinovitz@irt-saintexupery.com; edemontb@ens-paris-saclay.fr;

[†]These authors contributed equally to this work.

Abstract

We study the classical problem of approximating a non-decreasing function $f : \mathcal{X} \to \mathcal{Y}$ in $L^p(\mu)$ norm by sequentially querying its values, for known compact real intervals $\mathcal{X}$, $\mathcal{Y}$ and a known probability measure μ on $\mathcal{X}$. For any function f we characterize the minimum number of evaluations of f that algorithms need to guarantee an approximation $\widehat{f}$ with an $L^p(\mu)$ error below ε after stopping. Unlike worst-case results that hold uniformly over all f, our complexity measure is dependent on each specific function f. To address this problem, we introduce GreedyBox, a generalization of an algorithm originally proposed by Novak (1992) for numerical integration. We prove that GreedyBox achieves an optimal sample complexity for any function f, up to logarithmic factors. Additionally, we uncover results regarding piecewise-smooth functions. Perhaps as expected, the $L^p(\mu)$ error of GreedyBox decreases much faster for piecewise-C^2 functions than predicted by the algorithm (without any knowledge on the smoothness of f). A simple modification even achieves optimal minimax approximation rates for such functions, which we compute explicitly. In particular, our findings highlight multiple performance gaps between adaptive and non-adaptive algorithms, smooth

and piecewise-smooth functions, as well as monotone or non-monotone functions. Finally, we provide numerical experiments to support our theoretical results.

Keywords: L^p-approximation, sequential algorithms, numerical integration

1 Introduction

Let $\mathcal{X}, \mathcal{Y}$ be any non-empty compact intervals in $\mathbb{R}$. The problem we consider in this paper is the following. Given any non-decreasing function $f : \mathcal{X} \to \mathcal{Y}$ that is initially unknown but that a learner can sequentially evaluate at points $x_1, x_2, \ldots \in \mathcal{X}$ of their choice, how to best estimate f with as few evaluations of f as possible? We will study the $L^p(\mu)$ error as a performance criterion, for some known integer $p \geqslant 1$ and some known probability measure μ on $\mathcal{X}$. More precisely, we will study algorithms that can guarantee an $L^p(\mu)$ error observably below ε after a finite number of evaluations, and will characterize the minimum number of such evaluations to reach this goal, for any non-decreasing function f. Even though this problem is a classical one, finding the best f-dependent sample complexity and an algorithm that achieves it is still an open question.

To make things more formal, we first describe how the learner interacts with the unknown function f.

Online protocol.

Given an accuracy level $\varepsilon > 0$, the learner first chooses a point $x_1 \in \mathcal{X}$, then observes $f(x_1) \in \mathcal{Y}$, then chooses $x_2 \in \mathcal{X}$, then observes $f(x_2) \in \mathcal{Y}$, etc. At each round $t \geqslant 2$, the point $x_t \in \mathcal{X}$ is chosen as a measurable function of the whole history $h_{t-1} := \big(x_1, f(x_1), \ldots, x_{t-1}, f(x_{t-1})\big)$. The process ends after a finite number $\tau_\varepsilon \geqslant 1$ of rounds whose value may be determined during the observation process (τ_ε is a stopping time).[1] Finally, after observing the whole history h_{τ_ε}, the learner outputs a function $\widehat{f}_{\tau_\varepsilon} : \mathcal{X} \to \mathcal{Y}$ as a candidate for estimating f. We will call *algorithm*[2] any procedure that, given $\varepsilon > 0$ and f, returns a tuple $\big(\tau_\varepsilon, (x_t)_{1 \leqslant t \leqslant \tau_\varepsilon}, \widehat{f}_{\tau_\varepsilon}\big)$ in $\mathbb{N}^* \times \mathcal{X}^{\tau_\varepsilon} \times (\mathcal{X} \to \mathcal{Y})$ satisfying the above online protocol. We only consider deterministic algorithms, except for the integral estimation problem (Section 4.1) for which randomized algorithms achieve better rates in expectation.

[1] This means that, for any integer $t \geqslant 1$, whether the inequality $\tau_\varepsilon \leqslant t$ is true or not is fully known after observing h_t (in a measurable way).

[2] Throughout the paper our definition of algorithms refers in fact to *adaptive algorithms* that adjust their sequence of points $x_1, \ldots, x_t$ to the function f to be approximated based on previous observations. The latter should be contrasted with *non-adaptive algorithms*, for which the sequence of points $(x_t)_{t \geqslant 1}$ is fixed for all functions f's.

Learning goal: small number of evaluations with guaranteed $L^p(\mu)$ error.

Let $p \geqslant 1$ be any positive integer and μ be any probability measure on $\mathcal{X}$. The performance of the learner will be evaluated by its $L^p(\mu)$ *error* defined by

$$\left\| \widehat{f}_{\tau_\varepsilon} - f \right\|_p := \left(\int_{\mathcal{X}} \left| \widehat{f}_{\tau_\varepsilon}(x) - f(x) \right|^p d\mu(x) \right)^{1/p} . \tag{1}$$

In all the sequel, an accuracy level $\varepsilon > 0$ will be initially given to the learner, who will be required to guarantee that $\|\widehat{f}_{\tau_\varepsilon} - f\|_p \leqslant \varepsilon$ after stopping, for any (initially unknown) non-decreasing function $f : \mathcal{X} \to \mathcal{Y}$. Given this constraint, the goal of the learner is to make as few evaluations of f as possible, that is, to minimize the stopping time τ_ε. We will also refer to τ_ε as the *sample complexity* of the algorithm.

Main intuitions and informal presentation of the results.

Before detailing our results in the next sections, we describe the main intuitions in the special case where $p = 1$ and μ is the Lebesgue measure on $\mathcal{X} = \mathcal{Y} = [0, 1]$. The ideas are introduced informally and will be made more precise later.

Imagine that we have already evaluated f at some points $x_1, \ldots, x_t \in [0, 1]$. Since we know that f is non-decreasing and bounded between 0 and 1, we can deduce that the graph of f is contained inside the $t + 1$ adjacent rectangles (or *boxes*) shown in Figure 1.[3] Therefore, estimating f with any function $\widehat{f}_t$ whose graph also lies in these $t+1$ adjacent boxes will guarantee an L^1 error $\|\widehat{f}_t - f\|_1$ of at most the total area ξ_t of these boxes. We can even achieve $\|\widehat{f}_t - f\|_1 \leqslant \xi_t/2$ by estimating f with a piecewise-constant function (on each box $B_j = [c_{j-1}, c_j] \times \left[y_j^-, y_j^+\right]$, choose $\widehat{f}_t(x) = (y_j^- + y_j^+)/2$) or with a piecewise-affine function (choose $\widehat{f}_t$ that linearly interpolates all the observed points $(c_j, f(c_j))$).

Now let $\varepsilon > 0$, and suppose that we want to guarantee an L^1 error below ε as quickly as possible. Given the above comment, it seems that an ideal choice of the sequence $x_1, x_2, \ldots$ is such that the total area ξ_t of the $t + 1$ adjacent boxes at round t falls below 2ε for the smallest value of t possible. We derive lower and upper bounds that support this intuition:

- Lower bound: if after stopping (at time τ_ε) we want to guarantee that $\|\widehat{f}_{\tau_\varepsilon} - f\|_1 \leqslant \varepsilon$ whatever f, then $\tau_\varepsilon + 1$ must be larger than or equal to the minimum number (denoted by $\mathcal{N}_1(f, 2\varepsilon)$) of adjacent boxes that contain the graph of f and whose total area is at most of 2ε (see Theorem 1).
- A nearly optimal greedy algorithm: of course an optimal choice of $x_1, x_2, \ldots$ is impractical (it would require the full knowledge of the function f). However, a natural algorithm is to choose at each round t the next point x_{t+1} in the middle of the box with maximum area, so as to greedily reduce the total area of the boxes. See Figure 1 for an illustration. This algorithm, which we call GreedyBox, was suggested in a similar form by Novak [1]. One of our main contributions is

[3] Two boxes are degenerate (and thus not visible) on Figure 1, since GreedyBox evaluates f at the endpoints 0 and 1.

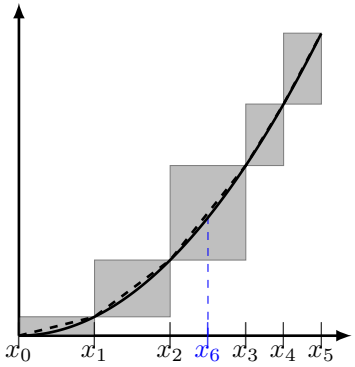


Fig. 1: An illustrative example of the problem on the square function. After 5 iterations, the output of GreedyBox is represented by the gray boxes. The estimated function is represented by the solid line and its approximation by the dashed line. The next evaluation point x_6 divides the box with the largest area in half.

to show that the stopping time τ_ε of GreedyBox is always at most of the order of the lower bound $\mathcal{N}_1(f, 2\varepsilon)$ up to a logarithmic factor in $1/\varepsilon$ (see Theorem 2).

Both the lower and upper bounds are proved in a more general setting, in $L^p(\mu)$ norm, for any integer $p \geqslant 1$ and any probability measure μ on $\mathcal{X}$.

1.1 Contributions and outline of the paper

Our main contribution is to characterize the optimal sample complexity of algorithms with guaranteed $L^p(\mu)$ error after stopping (see after Equation (1)), for any non-decreasing function f, any $p \geqslant 1$ and any probability measure μ. More precisely:

- In Section 2 we prove a general f-dependent lower bound that applies to any algorithm with guaranteed $L^p(\mu)$ error after stopping (see Theorem 1).
- In Section 3 we study GreedyBox (Algorithm 1) and show that its sample complexity matches our lower bound up to logarithmic factors (see Theorem 2). An important practical feature of GreedyBox is that, at each iteration t, it provides a certificate that upper bounds its error and stops as soon as this certificate falls below ε.

All the results are written in the case where $\mathcal{X} = \mathcal{Y} = [0, 1]$ for convenience, but all of them can be rescaled to any non-empty compact intervals $\mathcal{X}$ and $\mathcal{Y}$ of $\mathbb{R}$.[4]

In Section 4 we study consequences (with improved rates) for two specific subproblems:

- *Integral approximation.* For this problem, we show that the deterministic version of GreedyBox (Algorithm 1) is also optimal up to logarithmic factors. However, drawing inspiration from Novak [1], we introduce a randomized version in Section 4.1.2 that improves the accuracy by a factor of $t^{-1/2}$ in expectation after t iterations (Theorem 4).

[4]The case where $\mathcal{X}$ is not closed can be addressed similarly via a simple extension argument and by replacing the values of f at the endpoints of $\mathcal{X}$ by the endpoints of $\mathcal{Y}$.

- *Worst-case function approximation under a smoothness assumption.* In the worst case, the upper bound of Theorem 2 is of the order of ε^{-1} for monotone functions. It is well known that for smooth functions, better rates can be achieved using improved quadrature formulas (e.g., Davis and Rabinowitz [2]). For example, C^2 functions can be ε-approximated in any $L^p(\mu)$ norm after roughly $\varepsilon^{-1/2}$ evaluations (also discussed in Appendix B.8). In Section 4.2, we provide a minimax lower bound showing that $\Omega\big(\varepsilon^{-1+(\frac{1-\alpha}{1+p})_+}\big)$ function evaluations are necessary for any algorithm seeking to approximate a piecewise-affine function with $\varepsilon^{-\alpha}$ singularities (see Proposition 6). We establish that GreedyBox in fact achieves this rate for piecewise-C^2 functions when $\alpha \geqslant 1/2$, but is suboptimal in the regime $0 \leqslant \alpha < 1/2$ (Theorem 5 and Proposition 7). Lastly, we propose a simple modification of GreedyBox that optimally addresses both regimes. These results highlight three significant differences for the L^p-approximation problem:
  - between monotone piecewise-C^k functions with two or more discontinuities, for which a better rate than $\varepsilon^{-1/2}$ is not achievable, and C^k functions (with no singularities) which can be approximated at a rate of $\mathcal{O}(\varepsilon^{-1/k})$;
  - between general piecewise-C^2 functions and monotone piecewise-C^2 functions, with a minimax rate respectively of at least $\Omega(\varepsilon^{-p})$ and at most $o(\varepsilon^{-1})$ for $\alpha < 1$;
  - between non-adaptive algorithms, which need $\Omega(\varepsilon^{-1})$ function evaluations, and adaptive ones, which only require $o(\varepsilon^{-1})$ for $\alpha < 1$.

Finally, in Section 5, we provide numerical experiments that compare GreedyBox to the trapezoidal method on several functions f. Our simulations validate the rates anticipated by our analysis and demonstrate the superiority of our approach compared to the uniform trapezoidal rule in approximating monotone piecewise-smooth functions within the L^p norm.

1.2 Important definitions and notation

We now introduce several definitions and notations that will be useful to present our results more formally. In all the sequel, we work with $\mathcal{X} = \mathcal{Y} = [0,1]$, some fixed integer $p \geqslant 1$ and a probability measure μ on $[0,1]$.

Standard notation.

We denote by $\mathbb{N}^*$ the set of integers greater than or equal to 1. For any $x \in \mathbb{R}$, we denote the floor and ceiling functions at x by $\lfloor x \rfloor := \max\{k \in \mathbb{Z} : k \leqslant x\}$ and $\lceil x \rceil := \min\{k \in \mathbb{Z} : k \geqslant x\}$ respectively. For any measurable function $g : [0,1] \to \mathbb{R}$, we denote the $L^p(\mu)$-norm of g by $\|g\|_p := \big(\int_{[0,1]} |g(x)|^p d\mu(x)\big)^{1/p}$.

Box, width, generalized area.

We call *box* any subset $B = [x^-, x^+] \times [y^-, y^+] \subseteq [0,1]^2$ with $x^- < x^+$ and $y^- \leqslant y^+$. Its *width* is given by

$$\mathrm{width}(B) := \mu\big((x^-, x^+)\big)$$

and its *generalized area* is given by

$$\mathcal{A}_p(B) := \Big((y^+ - y^-)^p \, \mu\big((x^-, x^+)\big) \Big)^{1/p} .$$

Note that the above definitions consider open intervals (x^-, x^+). The generalized area corresponds to the usual notion of area when $p = 1$ and μ is the Lebesgue measure.

Adjacent boxes.

We denote by $\mathcal{B}$ the set of all boxes. We say that a finite sequence of $t \geqslant 1$ boxes $B_1, \dots, B_t \in \mathcal{B}$ are *adjacent* if and only if they are of the form $B_j = [c_{j-1}, c_j] \times [y_j^-, y_j^+]$ for all $j = 1, \dots, t$, for some sequence $0 = c_0 < c_1 < \ldots < c_{t-1} < c_t = 1$.

Box-cover and box-covering number of a function.

Let $f : [0,1] \to [0,1]$ be any non-decreasing function. We call *box-cover of* f any sequence $B_1, \dots, B_t \in \mathcal{B}$ of adjacent boxes that contains the graph of f except maybe at the boxes' endpoints, i.e., writing $B_j = [c_{j-1}, c_j] \times [y_j^-, y_j^+]$ as above, such that $\big\{(x, f(x)) : x \in [0,1] \setminus \{c_0, \dots, c_t\}\big\} \subseteq \cup_{i=1}^t B_i$. Furthermore, given $\varepsilon > 0$, we define two complexity quantities:

(i) $\mathcal{N}_p(f, \varepsilon)$ denotes the minimum cardinality t of a box-cover $B_1, \dots, B_t$ of f whose generalized areas satisfy $\big(\sum_{i=1}^t \mathcal{A}_p(B_i)^p\big)^{1/p} \leqslant \varepsilon$. We call this quantity the *box-covering number of* f *at scale* ε.

(ii) $\mathcal{N}'_p(f, \varepsilon)$ denotes the minimum cardinality t of a box-cover $B_1, \dots, B_t$ of f with generalized areas $\mathcal{A}_p(B_i) \leqslant \varepsilon$ for all $i = 1, \dots, t$.

All our main results will be expressed in terms of $\mathcal{N}_p(f, \varepsilon)$; the other quantity $\mathcal{N}'_p(f, \varepsilon)$ will however be useful in the proofs. The connections between the two are described in Appendix A.3.

We now reinterpret the condition $\big(\sum_{i=1}^t \mathcal{A}_p(B_i)^p\big)^{1/p} \leqslant \varepsilon$ appearing in (i) in an equivalent way. When $p = 1$, this condition corresponds to the total area of the boxes being bounded by ε (as mentioned earlier in the introduction). For general $p \geqslant 1$, an equivalent and useful formulation is the following. Denote by $B_j = [c_{j-1}, c_j] \times [y_j^-, y_j^+]$, $j = 1, \dots, t$, the boxes of the cover, and by $f^-(x) := \sum_{j=1}^t y_j^- \mathbb{1}_{x \in (c_{j-1}, c_j)}$ and $f^+(x) := \sum_{j=1}^t y_j^+ \mathbb{1}_{x \in (c_{j-1}, c_j)}$ the best known lower and upper bounds on the function f inside the boxes B_j. Then, $\big(\sum_{i=1}^t \mathcal{A}_p(B_i)^p\big)^{1/p} \leqslant \varepsilon$ is equivalent to

$$\Big(\sum_{j=1}^t \big(y_j^+ - y_j^-\big)^p \mu\big((c_{j-1}, c_j)\big) \Big)^{1/p} \leqslant \varepsilon \,, \qquad \text{that is,} \qquad \|f^+ - f^-\|_p \leqslant \varepsilon \,. \tag{2}$$

The last condition $\|f^+ - f^-\|_p \leqslant \varepsilon$ implies that f^- and f^+ are good lower and upper bounds on the function f outside of the points c_j. Note an interesting connection with the definition of bracketing entropy in empirical processes theory (see, e.g., [3, Chapter 19] and references therein).

The following lemma shows that $\mathcal{N}_p(f,\varepsilon)$ is always well defined and at most of the order of $1/\varepsilon$. The proof is postponed to Appendix A, where we collect other useful properties about $\mathcal{N}_p(f,\varepsilon)$.

Lemma 1. *For all non-decreasing functions $f:[0,1]\to[0,1]$ and $\varepsilon>0$, the quantity $\mathcal{N}_p(f,\varepsilon)$ is well defined and upper bounded by*

$$\mathcal{N}_p(f,\varepsilon)\leqslant\lceil 1/\varepsilon\rceil\ . \tag{3}$$

Though the rate of $1/\varepsilon$ is tight in the limit $\varepsilon\to 0$ for many functions such as $f:x\mapsto x$ and probability measures such as $\mu=\mathrm{Leb}$, the asymptotic behavior of $\mathcal{N}_p(f,\varepsilon)$ when $\varepsilon\to 0$ does not necessarily reflect the shape of f for a given $\varepsilon\in(0,1]$. Indeed all functions f which are ε_0-close to some piecewise-constant function have a very small box-covering number $\mathcal{N}_p(f,\varepsilon_0)$, even if $\mathcal{N}_p(f,\varepsilon)$ is large in the limit $\varepsilon\to 0$. Importantly, as we show in the next sections, the estimation problem addressed in this paper is very easy for such functions f and scales ε_0.

1.3 Related works

Quadrature formulas (approximation formulas for the computation of an integral) have long been studied for diverse classes of functions and algorithms along the past decades. We focus on references with strong connections to our problem. Sukharev [4] proved that affine methods are minimax optimal in the set of nonadaptive methods for every convex set of functions. It is in particular the case for non-decreasing functions. Since the work of Bakhvalov [5], it is known that nonadaptive methods are as good as adaptive ones for any class of functions that is both convex and symmetric. Note that the last hypothesis is not verified in our problem. However, Kiefer [6] later showed that the trapezoidal rule is optimal among all deterministic (possibly adaptive) methods for integral approximation in the case of monotone functions. His work was completed by the one of Novak [1], who gave optimal bounds for different possible types of algorithms as summarized in Table 1. In particular it is shown that adaption combined with randomization is key to obtaining an improved rate in expectation. These bounds were later extended by Papageorgiou [7] to the integral approximation of multivariate monotone functions. The books from Davis [2] and Brass [8] give a larger panel of results on numerical integration under various assumptions.

Strategy	**Non-adaptive**	**Adaptive**
Deterministic	ε^{-1}	ε^{-1}
Stochastic	ε^{-1}	$\varepsilon^{-2/3}$

Table 1: Minimax rates proved by Novak [1].

Of course, many other function sets were studied. A non-exhaustive list includes work on the set of unimodal functions [9], on the set of functions with bounded variation

[10], on convex and symmetric classes of functions [11–13], and on various classes of multivariate functions [14–16]. Note that the bounds in the previous papers are not f-dependent.

All of the aforementioned works were on integral approximation, an easier problem than L^p approximation. Numerous studies investigate the L^p approximation of diverse classes of functions, mostly using interpolation. A known example is polynomial interpolation for k-times continuously differentiable functions. Many works employ the modulus of smoothness of the function f to bound its L^p approximation, providing f-dependent bounds. It is for example the case of [17] for the approximation of periodic functions using trigonometric polynomials. The computation of a best linear L^p-approximation, where a basis $\phi_1, \ldots, \phi_k$ of functions is previously given and one looks for the weights $w \in \mathbb{R}^k$ that minimize $\|f - \sum_i w_i \phi_i\|_p$, was studied in depth (e.g. [18, 19]). [20] studied the class of functions with their first k derivatives continuous except at one singularity, and showed that adaptive algorithms are better that non-adaptive in the case of integral estimation. The same remark and work was later carried to approximation in [21]. Some work involves k-monotone functions, that is, functions with monotone k-th derivatives. The papers [22, 23] studied the rate of convergence of interpolation methods on this set of functions, and showed results depending on the modulus of smoothness of the function. To our knowledge, little is known about $L^p(\mu)$ approximation of general non-decreasing functions, without any continuity or smoothness assumptions.

Adaptive methods have a long history in numerical integration and other approximation or learning problems. For numerical integration of monotone functions, as mentioned above, adaption combined with randomization is key to improving worst-case rates (see [1] and Table 1 above, as well as [24]). We show that adaption is also key to obtaining less pessimistic, f-dependent error bounds. This work also shares algorithmic principles with online learning methods for (possibly noisy) black-box optimization, such as bandit algorithms [25] or the EGO algorithm [26]. Indeed such algorithms rely on adaptive sampling strategies to reduce the current uncertainty (via, e.g. UCB or Bayesian approaches), which is reminiscent of the way GreedyBox selects the box to be split at time t. Close to our paper is the work by Bachoc et al. [27], who derive f-dependent error bounds for certified black-box Lipschitz optimization.

Lastly, Bonnet et al. [28] explore a close variant of our main algorithm (Algorithm 1), which itself draws significant inspiration from Novak [1] for numerical integration. Bonnet et al. [28] address the problem of adaptively reconstructing a monotone function from imperfect observations. In contrast to our approach, they do not provide any guarantees regarding sample complexity or L^p approximation. Their focus is on asymptotic convergence guarantees, including point-wise, L^1, or L^∞ norm convergence, as the number of function evaluations tends towards infinity. They do not provide any convergence rate information, nor finite time or f-dependente guarantees. Nevertheless, they highlight an intriguing application in uncertainty quantification that could potentially benefit from our analysis.

2 Lower bound

In this section we prove a lower bound on the number of evaluations of f that any deterministic algorithm must request in order to guarantee an ε-approximation of f in $L^p(\mu)$-norm after stopping, when only given the prior knowledge that $f : [0,1] \to [0,1]$ is non-decreasing. The next theorem states that in such a case, at least $\mathcal{N}_p(f, 2\varepsilon) - 1$ evaluations of f are necessary.

In the sequel we write $\tau(f)$ to make it explicit that the stopping time τ of the algorithm depends on the underlying function f (through the sequentially observed values $f(x_1), f(x_2), \ldots$).

Theorem 1. *Let $\varepsilon > 0$ and $\big(\tau(f), (x_t)_{1 \leqslant t \leqslant \tau(f)}, \widehat{f}_{\tau(f)}\big)$ be the output of any deterministic algorithm such that, for all non-decreasing functions $f : [0,1] \to [0,1]$,*

$$\tau(f) < +\infty \qquad \text{and} \qquad \big\|\widehat{f}_{\tau(f)} - f\big\|_p \leqslant \varepsilon \; .$$

Then, for all non-decreasing functions $f : [0,1] \to [0,1]$,

$$\tau(f) \geqslant \mathcal{N}_p(f, 2\varepsilon) - 1 \; .$$

In words, any algorithm that is guaranteed to output an ε-approximation after finitely-many evaluations whatever the non-decreasing function f *must* evaluate each f at least $\mathcal{N}_p(f, 2\varepsilon) - 1$ times before stopping. In Section 3 we show a matching upper bound up to a multiplicative factor of the order of $p \log(1/\varepsilon)$. This indicates that the box-covering number $\mathcal{N}_p(f, 2\varepsilon)$ introduced in Section 1.2 is a key quantity to describe the inherent difficulty of the estimation problem.

Note that the lower bound holds for every f simultaneously. It thus has a similar flavor to distribution-dependent lower bounds that have been proved for the stochastic multi-armed bandit problem in online learning theory (see, e.g., Chapter 16 by Lattimore and Szepesvári [25]). Recently an f-dependent lower bound (also based on a notion of cover) was proved by Bachoc et al. [27] for certified zeroth-order Lipschitz optimization, where algorithms are required to output error certificates (i.e., observable upper bounds on the optimization error).

Proof. Assume for a moment that there exists a non-decreasing function $g : [0,1] \to [0,1]$ such that $\tau(g) < \mathcal{N}_p(g, 2\varepsilon) - 1$. When run on g, the algorithm only uses the $\tau(g)$ query points $x_1, \ldots, x_{\tau(g)}$ before stopping. Let $0 < \widetilde{x}_1 < \ldots < \widetilde{x}_n < 1$ denote the ordered values after removing redundancies and the values 0 and 1 (if applicable), with $0 \leqslant n \leqslant \tau(g)$. Consider the adjacent boxes $B_i = [\widetilde{x}_i, \widetilde{x}_{i+1}] \times [g(\widetilde{x}_i), g(\widetilde{x}_{i+1})]$ for $i \in \{0, \ldots, n\}$ where we set $\widetilde{x}_0 = 0$ and $\widetilde{x}_{n+1} = 1$. The sequence $B_0, \ldots, B_n$ is a box-cover of g (by monotonicity). We construct two functions g_- and g_+ that surround g:

$$g_- : x \mapsto \begin{cases} g(\widetilde{x}_i) & \text{if } \widetilde{x}_i \leqslant x < \widetilde{x}_{i+1} \\ g(1) & \text{if } x = 1 \end{cases} \text{and } g_+ : x \mapsto \begin{cases} g(\widetilde{x}_{i+1}) & \text{if } \widetilde{x}_i < x \leqslant \widetilde{x}_{i+1} \\ g(0) & \text{if } x = 0 \; . \end{cases}$$

Since $g_-(\widetilde{x}_i) = g_+(\widetilde{x}_i) = g(\widetilde{x}_i)$ for all $i \in \{0, \dots, n+1\}$, the algorithm (which is deterministic) would behave the same when run with g_- or g_+ as when run with g, and would construct the same approximation function $\widehat{g}_{\tau(g)}$ after the same number $\tau(g_-) = \tau(g_+) = \tau(g)$ of evaluations. However,

$$\|g_+ - g_-\|_p^p = \sum_{i=0}^{n} (g(\widetilde{x}_{i+1}) - g(\widetilde{x}_i))^p \mu((\widetilde{x}_i, \widetilde{x}_{i+1})) = \sum_{i=0}^{n} \mathcal{A}_p^p(B_i) > (2\varepsilon)^p \ ,$$

where the last inequality follows from $n+1 \leqslant \tau(g)+1 < \mathcal{N}_p(g, 2\varepsilon)$ and the definition of $\mathcal{N}_p(g, 2\varepsilon)$. The triangle inequality then yields

$$\|\widehat{g}_{\tau(g)} - g_-\|_p + \|\widehat{g}_{\tau(g)} - g_+\|_p \geqslant \|g_+ - g_-\|_p > 2\varepsilon \ ,$$

which shows that one of $\|\widehat{g}_{\tau(g)} - g_-\|_p$ or $\|\widehat{g}_{\tau(g)} - g_+\|_p$ is larger than ε. Since (as proved above) $\widehat{g}_{\tau(g)}$ is the approximation function output by the algorithm both with g_- and g_+, which are both non-decreasing, the last conclusion is in contradiction with the assumption that $\|\widehat{f}_{\tau(f)} - f\|_p \leqslant \varepsilon$ for all non-decreasing functions $f : [0,1] \to [0,1]$. This concludes the proof. □

3 Upper bound

In this section we introduce the GreedyBox algorithm and derive an f-dependent sample complexity bound (Theorem 2) that matches the lower bound of Theorem 1 up to a logarithmic factor, for every non-decreasing function f. In Section 4 we will study consequences and derive improved bounds for integral estimation (in expectation) and worst-case approximation of piecewise-C^2 functions.

3.1 Algorithm and main result

We consider Algorithm 1 below, which draws heavily on an algorithm proposed by Novak [1, Section 3.2] for numerical integration, and which we call GreedyBox thereafter. A variant for handling imperfect observations was also considered by Bonnet et al. [28].

It is remarkably simple: at every round, it selects the largest box in the current box-cover of f and replaces it with two smaller boxes by evaluating f at the middle or, more generally, at a conditional median for a general probability measure μ. At any $t \geqslant 1$, we approximate f with the trapezoidal estimator $\widehat{f}_t$ defined as the piecewise-affine function that joins the points $\big(b_k^t, f(b_k^t)\big)_{0 \leqslant k \leqslant t}$ visited up to time t. Note that this estimator uses $t+1$ evaluations of f. We stop Algorithm 1 at time τ_ε, which is the first $t \geqslant 1$ when the certificate $\xi_t = \sum_{k=1}^{t} (a_k^t)^p$ falls below ε^p. (This is because ξ_t is a valid upper bound on $\big\|\widehat{f}_t - f\big\|_p^p$, by Lemma 2 below.)

Algorithm 1: GreedyBox (inspired from Novak [1, Section 3.2]).

Input: $\varepsilon \in (0,1], p \geqslant 1$, probability measure μ on $[0,1]$.
Init: Set $t=1$, $x_0 = b_0^1 = 0$, $x_1 = b_1^1 = 1$, evaluate $f(0)$ and $f(1)$, and set $\xi_1 = (a_1^1)^p = (f(1)-f(0))^p \, \mu\big((0,1)\big)$.

while $\xi_t > \varepsilon^p$ **do**

1. Select a box with largest generalized area: pick $k_*^t \in \arg\max_{k \in \{1,\dots,t\}} a_k^t$.
2. Let x_{t+1} be a median of the conditional distribution $\mu(\cdot | (b_{k_*^t-1}^t, b_{k_*^t}^t))$, and evaluate f at x_{t+1}.
3. Sort the points $x_0, x_1, \dots, x_{t+1}$ in increasing order:
$$b_0^{t+1} = 0 < b_1^{t+1} < \cdots < b_{t+1}^{t+1} = 1 \, .$$
4. Define the generalized areas for all $k \in \{1, \dots, t+1\}$ by
$$a_k^{t+1} := \mu\big((b_{k-1}^{t+1}, b_k^{t+1})\big)^{1/p} \, (f(b_k^{t+1}) - f(b_{k-1}^{t+1})) \, .$$
5. Update the certificate
$$\xi_{t+1} = \sum_{k=1}^{t+1} (a_k^{t+1})^p \, . \tag{4}$$
6. Let $t \leftarrow t+1$.

end

Set $\tau_\varepsilon = t$ and approximate f with the piecewise-affine function $\widehat{f}_{\tau_\varepsilon}$ defined by:
$$\forall x \in [0,1] \qquad \widehat{f}_{\tau_\varepsilon}(x) = \frac{f(b_k^{\tau_\varepsilon}) - f(b_{k-1}^{\tau_\varepsilon})}{b_k^{\tau_\varepsilon} - b_{k-1}^{\tau_\varepsilon}} (x - b_{k-1}^{\tau_\varepsilon}) + f(b_{k-1}^{\tau_\varepsilon}) \, ,$$
for $k \in \{1, \dots, \tau_\varepsilon\}$ such that $b_{k-1}^{\tau_\varepsilon} \leqslant x \leqslant b_k^{\tau_\varepsilon}$.

Output: $(\tau_\varepsilon, (x_t)_{1 \leqslant t \leqslant \tau_\varepsilon}, \widehat{f}_{\tau_\varepsilon})$.

Algorithmic complexity.

We assume that a median of the conditional distribution $\mu(\cdot | (b_{k_*^t-1}^t, b_{k_*^t}^t))$ can be computed exactly at every round t. When μ is the Lebesgue measure, it can indeed be computed in closed form: it is the midpoint $(b_{k_*^t-1}^t + b_{k_*^t}^t)/2$.

For the sake of simplicity, in Algorithm 1 we perform a sort (Step 3) and an argmax operation (Step 1) at each round t, to get the points in increasing order and to choose the box with the largest generalized area. However, one can get rid of the sort operation at each round and do it only once at the end, because GreedyBox does not need the order of the boxes before the last iteration, where it uses sorted points to build $\widehat{f}_{\tau_\varepsilon}$. Furthermore, for the argmax operation, naive methods yield a computational complexity of $\mathcal{O}(t)$ at each time t, resulting in a quadratic complexity for GreedyBox, far worse than the linear complexity of traditional algorithms such as the trapezoidal rule. To speed up GreedyBox, we use a classical algorithmic trick: a max-heap, which is a binary tree that takes logarithmic time to both remove the maximum value and add an element. This provides GreedyBox with a computational complexity of $\mathcal{O}(t \log(t))$ after t rounds, which is closer to the complexity of the trapezoidal rule.

Upper bound on the sample complexity.

Let $\varepsilon \in (0,1]$ be some target accuracy level. The next theorem provides a bound on the sample complexity τ_ε of GreedyBox. The proof appears in Section 3.2.

Theorem 2. *Let $f : [0,1] \to [0,1]$ be non-decreasing,*[5] *$p \geqslant 1$, and $\varepsilon \in (0,1]$. Then,* GreedyBox *defined above (Algorithm 1) satisfies*

$$\left\| \widehat{f}_{\tau_\varepsilon} - f \right\|_p := \left(\int_0^1 \left(\widehat{f}_{\tau_\varepsilon}(x) - f(x) \right)^p dx \right)^{1/p} \leqslant \varepsilon \,,$$

at the stopping time τ_ε. Furthermore, its sample complexity is bounded as follows:

$$\tau_\varepsilon \leqslant 32p^2 \left(\log_2(2/\varepsilon^2) + 2 \right)^2 \mathcal{N}_p(f, \varepsilon) \,.$$

We make three comments before proving the theorem.

A new f-dependent bound. Since $\mathcal{N}_p(f,\varepsilon) \leqslant \lceil 1/\varepsilon \rceil$ for all non-decreasing functions f (by Lemma 1), the above sample complexity bound $\tau_\varepsilon = \mathcal{O}\left(\mathcal{N}_p(f,\varepsilon) \log^2(1/\varepsilon)\right)$ implies the well-known upper bound of $\mathcal{O}(1/\varepsilon)$ up to logarithmic factors in the worst case. Importantly, though the rate of $1/\varepsilon$ is worst-case optimal (see, e.g., [6, Section 5.A]), Theorem 2 yields a much better bound for functions f that are easier to approximate, such as functions close to piecewise-constant functions, because $\mathcal{N}_p(f,\varepsilon)$ is small in that case. Since the GreedyBox algorithm uses no prior knowledge on f (beyond monotonicity) to stop at τ_ε, it is adaptive to the unknown complexity $\mathcal{N}_p(f,\varepsilon)$.

A nearly optimal bound. Note that the lower bound of Theorem 1 is in terms of $\mathcal{N}_p(f, 2\varepsilon)$, while the upper bound of Theorem 2 is proportional to $\mathcal{N}_p(f,\varepsilon)$. By a simple argument (dividing boxes p times to reduce their generalized widths by a factor of 2^p, similarly to the proof of Lemma 4), we can prove that $\mathcal{N}_p(f,\varepsilon) \leqslant 2^p \mathcal{N}_p(f, 2\varepsilon)$. Therefore, the lower and upper bounds of Theorems 1 and 2 match up to a logarithmic factor. For each non-decreasing function $f : [0,1] \to [0,1]$, GreedyBox is thus nearly optimal among all algorithms with guaranteed $L^p(\mu)$ error after stopping.

A possible minor improvement. When μ is the Lebesgue measure, the bound on τ_ε could be slightly improved (in the constants) by replacing the certificate in Equation (4) with $\xi_{t+1} = \frac{1}{1+p} \sum_{k=1}^{t+1} (a_k^t)^p$ (see Lemma 9 in Appendix B.1). While a similar minor improvement is likely to hold for general μ with a slightly different interpolation $\widehat{f}_t$ (non-necessarily piecewise-affine), we decided to focus on piecewise-affine interpolations for the sake of presentation.

3.2 Proof of Theorem 2

Before proving Theorem 2, we first state three lemmas, whose proofs are all postponed to Appendix B.

The first one below shows that, at any round t before stopping, the error of GreedyBox is at most the sum of the generalized areas to the power p of the current-box cover of f. We recall that $a_k^t := (\mu(b_{k-1}^t, b_k^t))^{1/p} (f(b_k^t) - f(b_{k-1}^t))$ denotes the generalized

[5] Recall that the input and output sets of f can be rescaled to any non-empty compact intervals $\mathcal{X}$ and $\mathcal{Y}$ of $\mathbb{R}$, changing the results only by a multiplicative constant.

area of the k-th box at round t, and we define the trapezoidal estimato $\widehat{f}_t$ to be the piecewise-affine function that joins the points $\big(b_k^t, f(b_k^t)\big)_{0 \leqslant k \leqslant t}$ visited up to time t.

Lemma 2. *Let $f : [0,1] \to [0,1]$ be non-decreasing, $p \geqslant 1$ and $\varepsilon \in (0,1]$. For any $t \in \{1, \dots, \tau_\varepsilon\}$,*

$$\|\widehat{f}_t - f\|_p^p := \int_0^1 \left|\widehat{f}_t(x) - f(x)\right|^p d\mu(x) \leqslant \sum_{k=1}^{t} (a_k^t)^p =: \xi_t \,.$$

The next two lemmas are used to control τ_ε. We first show (by a dichotomy argument) that the algorithm can quickly make all boxes equally small. Recall from Section 1.2 that $\mathcal{N}'_p(f,\varepsilon)$ denotes the minimum cardinality of a box-cover of f for which each box has a generalized area below ε.

Lemma 3. *Let $f : [0,1] \to [0,1]$ be non-decreasing, $p \geqslant 1$ and $\varepsilon \in (0,1]$. Define $\tau'_\varepsilon := 2\big(1 + \lceil p \log_2(1/\varepsilon)\rceil\big)\mathcal{N}'_p\big(f,\varepsilon\big)$, and assume that* GreedyBox *is such that $\tau_\varepsilon > \tau'_\varepsilon$. Then, at time τ'_ε, all the boxes maintained by* GreedyBox *have a generalized area bounded from above by ε, i.e., $a_k^{\tau'_\varepsilon} \leqslant \varepsilon$ for all $k \in \{1, \dots, \tau'_\varepsilon\}$.*

The next lemma shows that the certificate $\xi_t = \sum_{k=1}^t (a_k^t)^p$ at round t decreases at least linearly in t.

Lemma 4. *Let $f : [0,1] \to [0,1]$ be non-decreasing, $p \geqslant 1$ and $\varepsilon \in (0,1]$. For any $t \in \{1, \dots, \lfloor \tau_\varepsilon/2 \rfloor\}$, we have $\xi_{2t} \leqslant \xi_t/2$. Therefore, for all $t \leqslant s$ in $\{1, \dots, \tau_\varepsilon\}$,*

$$\xi_s \leqslant \frac{\xi_t}{2^{\lfloor \log_2(s/t) \rfloor}} \leqslant \left(\frac{2t}{s}\right) \xi_t \,. \tag{5}$$

Proof of Theorem 2. We are now ready to prove the theorem. The first inequality follows immediately from Lemma 2 and from the fact that $\xi_{\tau_\varepsilon} \leqslant \varepsilon^p$ by definition of τ_ε.

We now show by contradiction that $\tau_\varepsilon \leqslant 32\big(1+\lceil p \log_2(2/\varepsilon^2)\rceil\big)^2 \mathcal{N}_p(f,\varepsilon)$. Assume thus for a moment that

$$\tau_\varepsilon > 32\big(1 + \lceil p \log_2(2/\varepsilon^2)\rceil\big)^2 \mathcal{N}_p(f,\varepsilon) \,. \tag{6}$$

This assumption will be used implicitly when calling Lemmas 3 and 4 below, since it will imply that $\tau'_{\varepsilon'} \leqslant \tau''_\varepsilon < \tau_\varepsilon \leqslant \tau_{\varepsilon'}$ (so that the algorithm has not stopped before any round considered below). We will see in the end that it raises a contradiction.

Let $n_\varepsilon := \mathcal{N}_p(f,\varepsilon)$. By Lemma 3 applied with $\varepsilon' = \varepsilon/n_\varepsilon^{1/p}$, at time $\tau'_{\varepsilon'} := 2\big(1 + \lceil p \log_2(1/\varepsilon')\rceil\big)\mathcal{N}'_p(f,\varepsilon')$, the $\tau'_{\varepsilon'}$ boxes maintained by GreedyBox all have generalized areas at most of ε' each, so that the certificate $\xi_{\tau'_{\varepsilon'}}$ satisfies

$$\begin{aligned} \xi_{\tau'_{\varepsilon'}} \leqslant \tau'_{\varepsilon'} \cdot (\varepsilon')^p &\leqslant 2\big(1 + \lceil p \log_2(1/\varepsilon')\rceil\big)\mathcal{N}'_p(f,\varepsilon') \cdot \frac{\varepsilon^p}{n_\varepsilon} \\ &\leqslant 4\big(1 + \lceil p \log_2(2/\varepsilon^2)\rceil\big)\, \varepsilon^p \,, \end{aligned} \tag{7}$$

where we used the fact that $\mathcal{N}'_p(f,\varepsilon') \leqslant 2\mathcal{N}_p(f,\varepsilon)$ (by Lemma 8 in Appendix A.3) and that $1/\varepsilon' = \mathcal{N}_p(f,\varepsilon)^{1/p}/\varepsilon \leqslant 2^{1/p}/\varepsilon^{(p+1)/p} \leqslant 2/\varepsilon^2$ (since $\mathcal{N}_p(f,\varepsilon) \leqslant \lceil 1/\varepsilon \rceil \leqslant 2/\varepsilon$). Now, we apply Lemma 4 with $t = \tau'_{\varepsilon'}$ and $s = \tau''_\varepsilon := 8\big(1+\lceil p\log_2(2/\varepsilon^2)\rceil\big)\tau'_{\varepsilon'}$. It yields:

$$\xi_{\tau''_\varepsilon} \leqslant \left(\frac{2\tau'_{\varepsilon'}}{8\big(1+\lceil p\log_2(2/\varepsilon^2)\rceil\big)\tau'_{\varepsilon'}}\right)\xi_{\tau'_{\varepsilon'}} \overset{\text{by (7)}}{\leqslant} \varepsilon^p\,.$$

This raises a contradiction with (6), since (using again $\mathcal{N}'_p(f,\varepsilon') \leqslant 2\mathcal{N}_p(f,\varepsilon)$)

$$\tau''_\varepsilon = 8\big(1+\lceil p\log_2(2/\varepsilon^2)\rceil\big)\tau'_{\varepsilon'} \leqslant 32\big(1+\lceil p\log_2(2/\varepsilon^2)\rceil\big)^2\mathcal{N}_p(f,\varepsilon)$$

and τ_ε is by definition the first time t such that $\xi_t \leqslant \varepsilon^p$. Therefore, (6) must be false, so that $\tau_\varepsilon \leqslant 32\big(1+\lceil p\log_2(2/\varepsilon^2)\rceil\big)^2\mathcal{N}_p(f,\varepsilon)$. Elementary calculations conclude the proof of Theorem 2. □

4 Improvement for special cases

In this section, we derive consequences (with rates faster than the worst-case ε^{-1}) in two specific cases: integral estimation and piecewise-smooth functions.

In the sequel, we adopt a slightly different yet equivalent viewpoint than in Section 3. Though Algorithms 2 and 3, defined in this section, formally stop at round τ_ε, in the proofs, we extend their definitions to all rounds $t \geqslant 1$, by replacing the while condition with $\xi_t > 0$, defining τ_0 as the first round t (if any) where the certificate ξ_t reaches 0, and setting $\widehat{f}_s := \widehat{f}_{\tau_0}$ for all subsequent rounds $s \geqslant \tau_0 + 1$. Note that their approximation error equals zero for all $s \geqslant \tau_0$.

4.1 Side problem: integral estimation

Throughout this subsection, we focus on the case of integral estimation rather than approximation in $L^p(\mu)$-norm. The goal is to approximate the integral $I(f) = \int_0^1 f(x)d\mu(x)$ of a non-decreasing function f on $[0,1]$. This problem is simpler than the $L^1(\mu)$-approximation problem studied previously, and thus GreedyBox can be easily extended to integral estimation while maintaining the same bound.

A deterministic algorithm for the integral estimation problem is defined as a procedure that, given $\varepsilon > 0$, produces a tuple $(\tau_\varepsilon, (x_t)_{1\leqslant\tau_\varepsilon}, \widehat{I}_{\tau_\varepsilon}(f)) \in \mathbb{N}_+ \times \mathcal{X}^{\tau_\varepsilon} \times \mathbb{R}$, where $(x_t)_{t\geqslant 1}$ and τ_ε are defined sequentially, similar to the approximation in $L^p(\mu)$-norm. That is: for all $t \geqslant 2$, x_t is a measurable function of the history h_{t-1} and τ_ε is a stopping time after which the process of the algorithm ends. The algorithm finally outputs an approximation $\widehat{I}_{\tau_\varepsilon}(f)$ of the integral. We now study the convergence speed of GreedyBox in this setting (we replace the definition of $\widehat{f}_{\tau_\varepsilon}$ in the last line of GreedyBox by the computation of $\widehat{I}_{\tau_\varepsilon}(f) = \int_0^1 \widehat{f}_{\tau_\varepsilon}(x)d\mu(x)$) and check whether it achieves optimal convergence speed.

4.1.1 Nearly optimal performance for integral estimation

The following upper bound on the sample complexity of GreedyBox is a direct consequence of Theorem 2 with $p = 1$.

Corollary 1. *Let $f : [0,1] \to [0,1]$ be non-decreasing, and $\varepsilon \in (0,1]$. Then,* GreedyBox *with $p = 1$ satisfies, at the stopping time τ_ε,*

$$\left| \int_0^1 \widehat{f}_{\tau_\varepsilon}(x) d\mu(x) - \int_0^1 f(x) d\mu(x) \right| \leqslant \varepsilon \,.$$

Besides, its sample complexity is bounded from above as follows:

$$\tau_\varepsilon \leqslant 32 p^2 \big(\log_2(2/\varepsilon^2) + 2\big)^2 \mathcal{N}_p(f, \varepsilon) \,.$$

The question is now to check if the lower bound for this weaker problem is still the same one. The following theorem asserts that this is the case.

Theorem 3 (Matching lower bound)**.** *Let $\varepsilon > 0$ and $\mathcal{A}$ be any deterministic algorithm such that, for all non-decreasing functions $f : [0,1] \to [0,1]$:*

$$\tau_\varepsilon(f) < +\infty \text{ and } \left| \widehat{I}_{\tau_\varepsilon(f)}(f) - \int_0^1 f(x) d\mu(x) \right| \leqslant \varepsilon \,.$$

Then, for all non-decreasing functions $f : [0,1] \to [0,1]$

$$\tau_\varepsilon(f) \geqslant \mathcal{N}_p(f, 2\varepsilon) - 1 \,.$$

The proof closely follows that of Theorem 1 in the case of $p = 1$ and is left to the reader. Note that it is inspired from that of the well-known minimax lower bound of $1/(2n+2)$. This lower bound actually implies the lower bound of Theorem 1 for $p = 1$.

4.1.2 Improvement in expectation with randomization

We now provide a stochastic version of our algorithm; see Algorithm 2 below, which we call StochasticGreedyBox. With this randomized variant we prove better guarantees (in expectation only)[6] for the integral estimation problem than with the deterministic version. Note that the improvement is not true for estimating f in $L^p(\mu)$-norm. The idea to use randomization to improve the rates is due to Novak [1, Section 2.2]; we adapt this idea to a fully sequential algorithm, whose bound is now adaptive to the complexity of f. It's worth mentioning that, for the sake of simplicity, the random points X_k in Algorithm 2 are currently sampled at the conclusion of the algorithm. However, an alternative approach could involve sequential sampling when the intervals (b_{k-1}^t, b_k^t) are created, in order to get a sequential estimator $\widehat{I}_t(f)$ for all $t \geqslant 1$.

[6] We could easily derive high probability bound using Hoeffding's lemma.

Algorithm 2: StochasticGreedyBox.

Input: $\varepsilon > 0$, probability measure μ on $[0,1]$

Init: Set $t = 1$, $x_0 = b_0^1 = 0$ and $x_1 = b_1^1 = 1$, evaluate $f(0)$ and $f(1)$ and set $\xi_1 = a_1^1/2 = \mu((0,1))(f(1) - f(0))/2$;

while $\xi_t > \varepsilon$ **do**

1. Select the box with the largest area: find $k_*^t \in \arg\max_{1 \leqslant k \leqslant t} a_k^t$.
2. Let x_{t+1} be a median of the conditional distribution $\mu(\cdot|(b_{k_*^t-1}^t, b_{k_*^t}^t))$ and evaluate f at x_{t+1}.
3. Sort the points $x_0, x_1, \ldots, x_{t+1}$ in increasing order:
$$b_0^{t+1} = 0 < b_1^{t+1} < \cdots < b_{t+1}^{t+1} = 1\,.$$
4. Define the boxes areas for $k \in \{1, \ldots, t+1\}$ by
$$a_k^{t+1} := \mu((b_{k-1}^{t+1}, b_k^{t+1}))(f(b_k^{t+1}) - f(b_{k-1}^{t+1}))\,.$$
5. Update the certificate
$$\xi_{t+1} = \frac{1}{2}\sqrt{\sum_{k=1}^{t+1}(a_k^{t+1})^2}\,.$$
6. Let $t \leftarrow t+1$.

end

Set $\tau_\varepsilon = t$ and let $S_{\tau_\varepsilon} = \left\{k \in \{1, \ldots, \tau_\varepsilon\} \text{ s.t. } \mu((b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon})) > 0\right\}$;

for $k \in S_{\tau_\varepsilon}$ **do**

Sample X_k according to μ conditionally to the interval $(b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon})$ and evaluate f at X_k.

end

Approximate $I(f) = \int_0^1 f(x)d\mu(x)$ with the estimator:

$$\widehat{I}_{\tau_\varepsilon}(f) := \sum_{k \in S_{\tau_\varepsilon}} \mu((b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon}))f(X_k) + \sum_{k=0}^{\tau_\varepsilon} \mu(\{b_k^{\tau_\varepsilon}\})f(b_k^{\tau_\varepsilon})\,.$$

Output: $\left(\tau_\varepsilon, (x_t)_{1 \leqslant t \leqslant \tau_\varepsilon}, \widehat{I}_{\tau_\varepsilon}(f)\right)$.

The following lemma shows that the error of StochasticGreedyBox (Algorithm 2) at stopping time is indeed at most ε.

Lemma 5. *Let $\varepsilon > 0$ and let $f : [0,1] \to [0,1]$ be non-decreasing. Then, the output of Algorithm 2 satisfies*

$$\mathbb{E}\left[\left|\widehat{I}_{\tau_\varepsilon}(f) - I(f)\right|\right] \leqslant \xi_{\tau_\varepsilon} := \frac{1}{2}\sqrt{\sum_{k=1}^{\tau_\varepsilon}(a_k^{\tau_\varepsilon})^2} \leqslant \varepsilon\,.$$

Proof. First, we remark that the stopping time τ_ε and the ordered sequence $b_0^{\tau_\varepsilon}, \ldots, b_{\tau_\varepsilon}^{\tau_\varepsilon}$ are deterministic and do not depend on the randomness of the algorithm. These deterministic points being set, the random variables X_k for $k \in S_{\tau_\varepsilon}$ are distributed according to μ conditionally to each of the deterministic intervals defined by the $b_k^{\tau_\varepsilon}$'s

and satisfy for all $k \in S_{\tau_\varepsilon}$

$$X_k \sim \mu(\cdot|(b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon})) \quad \text{and} \quad \mathbb{E}[f(X_k)] = \frac{1}{\mu((b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon}))} \int_{b_{k-1}^{\tau_\varepsilon}}^{b_k^{\tau_\varepsilon}} f(x) d\mu(x) \,.$$

Therefore,

$$\begin{aligned}
\mathbb{E}\big[\widehat{I}_{\tau_\varepsilon}(f)\big] &= \sum_{k \in S_{\tau_\varepsilon}} \mu((b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon}))\mathbb{E}\big[f(X_k)\big] + \sum_{k=0}^{\tau_\varepsilon} \mu(\{b_k^{\tau_\varepsilon}\}) f(b_k^t) \\
&= \sum_{k=1}^{\tau_\varepsilon} \int_{(b_{k-1}^t, b_k^t)} f(x) d\mu(x) + \sum_{k=0}^{\tau_\varepsilon} \mu(\{b_k^t\}) f(b_k^t) \\
&= \int_0^1 f(x) d\mu(x) = I(f)
\end{aligned}$$

and

$$\begin{aligned}
\mathbb{E}\Big[\big|\widehat{I}_{\tau_\varepsilon}(f) - I(f)\big|\Big]^2 &\overset{\text{Jensen}}{\leqslant} \mathbb{E}\Big[\big(\widehat{I}_{\tau_\varepsilon}(f) - I(f)\big)^2\Big] = \operatorname{Var}\big(\widehat{I}_{\tau_\varepsilon}(f)\big) \\
&\overset{\text{Independence}}{=} \sum_{k \in S_{\tau_\varepsilon}} \mu((b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon}))^2 \operatorname{Var}\big(f(X_k)\big) \,. \qquad (8)
\end{aligned}$$

But since $X_k \in (b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon})$, by monotonicity of f, $f(X_k)$ takes values into the interval $[f(b_{k-1}^{\tau_\varepsilon}), f(b_k^{\tau_\varepsilon})]$. Thus,

$$\operatorname{Var}\big(f(X_k)\big) \leqslant \frac{1}{4}\big(f(b_k^{\tau_\varepsilon}) - f(b_{k-1}^{\tau_\varepsilon})\big)^2 = \frac{(a_k^{\tau_\varepsilon})^2}{4\mu((b_{k-1}^{\tau_\varepsilon}, b_k^{\tau_\varepsilon}))^2}$$

by definition of $a_k^{\tau_\varepsilon}$ (see step 3 of Algorithm 2). Therefore, substituting into Inequality (8) and taking the square root, we get

$$\mathbb{E}\Big[\big|\widehat{I}_{\tau_\varepsilon}(f) - I(f)\big|\Big] \leqslant \frac{1}{2}\sqrt{\sum_{k=1}^{\tau_\varepsilon} (a_k^{\tau_\varepsilon})^2} \,,$$

which is smaller than ε by the stopping criterion. □

Remark that if all boxes at time τ_ε have similar areas $a_k^{\tau_\varepsilon} \approx a$ (which the algorithm aims at getting by splitting only largest boxes), Lemma 5 is asking a to be at most of order $\mathcal{O}(\varepsilon/\sqrt{\tau_\varepsilon})$ in Algorithm 2, while Algorithm 1 required $\mathcal{O}(\varepsilon/\tau_\varepsilon)$. Therefore, this leads to an earlier stopping criterion and smaller sample complexity. Typically, after t rounds, the approximation error of Algorithm 2 is better than the one of Algorithm 1 by a factor $\sqrt{t}$. This is however not so simple to formulate in terms of sample complexity. We will thus only formulate the analog of Theorem 2 under the following assumption.

Assumption 1. *Let $\varepsilon > 0$. There exist $C > 0$ and $0 < \alpha < 1$ such that $\mathcal{N}(f, \varepsilon_1) \leqslant C\varepsilon_1^{-\alpha}$ for all $\varepsilon_1 \geqslant \varepsilon$.*

It is worth to notice that Assumption 1 is mild since $\mathcal{N}(f, \varepsilon) \leqslant \varepsilon^{-1}$ for any non-decreasing f and $\varepsilon > 0$. Furthermore, the assumption is non-asymptotic in ε since the requirement is only for $\varepsilon_1 \geqslant \varepsilon$.

Theorem 4. *Let $f : [0,1] \to [0,1]$ be non-decreasing which satisfies Assumption 1 for some $C, \alpha > 0$. Let $\varepsilon > 0$, then Algorithm 2 satisfies $\mathbb{E}\big[|\widehat{I}_{\tau_\varepsilon}(f) - I(f)|\big] \leqslant \varepsilon$. Besides the number of function evaluations is bounded from above by*

$$\tau_\varepsilon = \mathcal{O}\big(\log(1/\varepsilon)^{3/2}\varepsilon^{-\frac{1}{1/\alpha+1/2}}\big)\,.$$

The proof is postponed to Appendix B.4. The benefit of the stochastic algorithm is thus to replace the rate $\varepsilon^{-\alpha}$ obtained with the deterministic version with $\varepsilon^{-\frac{1}{1/\alpha+1/2}}$. This result generalizes the one obtained by Novak [1] in the case of $\alpha = 1$ (which corresponds our worst-case scenario) for a similar algorithm.

4.2 An intriguing result for piecewise-regular functions

For the sake of simplicity, in this section, we restrict ourselves to the *Lebesgue measure*.

As seen before, GreedyBox has a worst-case sample-complexity of order ε^{-1} up to logarithmic factors. An interesting question is how the error rate improves with regularity for non-decreasing functions, as well as how to adapt GreedyBox to achieve optimal rates for piecewise-smooth functions. More precisely, unlike the rest of the paper, in this section, we focus on analyzing the *effective* L^p-error rate of algorithms when run on piecewise-smooth functions. We control the number of evaluations of f until $\big\|\widehat{f}_t - f\big\|_p$ falls below ε, rather than the number τ_ε of evaluations until the certificate ξ_t falls below ε^p. The first (classical) complexity quantity can be much smaller than τ_ε (since the algorithm is only aware that f is non-decreasing, and lacks any prior regularity knowledge). This does not contradict the lower bound of Theorem 1 and reveals the effective performance of algorithms when run on simpler functions. For instance, on C^2 functions, the trapezoidal rule has an effective rate of order $\varepsilon^{-1/2}$ (a classical result recalled in Appendix B.8), but cannot guarantee ε-accuracy before order ε^{-1} evaluations if only given the knowledge that the underlying function is non-decreasing.

Upper bound on GreedyBox effective sample complexity.

We aim to explore an intriguing question: can we establish these guarantees for GreedyBox without making any modifications? Moreover, the trapezoidal rule fails to adapt to piecewise-C^2 functions, even for simple ones such as $f(x) = \mathbb{1}_{x \geqslant 1/3}$. On the contrary, GreedyBox adapts very well to the discontinuities: it converges exponentially fast for any piecewise-constant function. With this in head, one could ask if GreedyBox learns the jumps quickly enough to ensure an ε-accurate L^p error within

$\widetilde{\mathcal{O}}(\varepsilon^{-1/2})$ sample-complexity for piecewise-C^2 functions.[7] The next theorem shows that the rate can indeed be improved as long as the number of C^1-singularities[8] is at most of order $\mathcal{O}(\varepsilon^{-1})$. It should be noted that the number of C^1-singularities may explode to infinity as ε approaches zero and that the number of C^2-singularities does not affect the upper bound.

Theorem 5. *Let $\alpha > 0$ and $\varepsilon \in (0,1]$. Let $f : [0,1] \to [0,1]$ be a non-decreasing and piecewise-C^2 function with a number of C^1-singularities bounded by $\varepsilon^{-\alpha}$ and such that $|f''(x)| \leqslant 1$ whenever it is defined. Then, there exists*

$$t_\varepsilon = \begin{cases} \widetilde{\mathcal{O}}\left(\varepsilon^{-1+\frac{1}{2p+2}}\right) & \text{if } \alpha \leqslant \frac{1}{2} \\ \widetilde{\mathcal{O}}\left(\varepsilon^{-1+\left(\frac{1-\alpha}{1+p}\right)_+}\right) & \text{if } \alpha \geqslant \frac{1}{2} \end{cases}$$

such that $\left\|\widehat{f}_t - f\right\|_p \leqslant \varepsilon$ for all $t \geqslant t_\varepsilon$, where $\widehat{f}_t$ is the approximation of f returned by GreedyBox *after t rounds.*

Theorem 5 shows that, for piecewise-C^2 functions with $\alpha < 1$, GreedyBox achieves ε-accuracy in $o(\varepsilon^{-1})$ function evaluations which improves the worst-case guarantee of Theorem 2. In particular, when the number of singularities is finite, then $\alpha \to 0$ when $\varepsilon \to 0$, and the L^1-error is asymptotically of order $\mathcal{O}(\varepsilon^{-3/4})$. Note that this result contrasts with what happens for piecewise-C^2 functions without the non-decreasing assumption considered by [21], who showed that as soon as there are strictly more than one discontinuity, any algorithm has a worst-case L^p-error of order $\Omega(\varepsilon^{-p})$.

Minimax lower bound for approximating non-decreasing piecewise-smooth functions.

Interestingly, we now show that this result is optimal (up to logarithmic factors) among deterministic algorithms in the regime $\alpha \geqslant 1/2$ (Proposition 6), that is when the number of C^1-singularities is at least of order $\Omega(\varepsilon^{-1/2})$. In the other regime, which corresponds to more regular functions, we also show that our upper bound on GreedyBox cannot be improved (Proposition 7).

Proposition 6. *Let $p \geqslant 1$, $\varepsilon \in (0,1)$ and $\alpha > 0$. Then, for any deterministic adaptive algorithm $\mathcal{A}$ and for any*

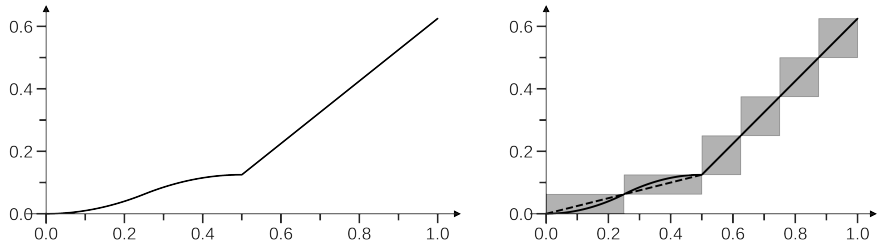
$$t < (2\varepsilon)^{-1+\left(\frac{1-\alpha}{1+p}\right)_+} - 1\,,$$

there exists a non-decreasing piecewise-affine function $f : [0,1] \to [0,1]$ with at most $\max\{2, \lceil \varepsilon^{-\alpha} \rceil\}$ discontinuities, such that $\|f - \widehat{f}_t\|_p > \varepsilon$.

Remarkably, the above lower bound demonstrates a clear difference between regular functions and piecewise-regular functions, even when the number of pieces is finite. Specifically, when considering the case of $p = 1$, the above lower bound shows that, for piecewise-C^∞ function with a constant number of discontinuities ($\alpha = 0$), surpassing

[7] The notation $\widetilde{\mathcal{O}}$ hides logarithmic factors.

[8] We call C^k-singularity of f a point $x \in [0,1]$ such that f is not C^k on any neighborhood of x. A discontinuity is a C^0-singularity.



(a) Worst-case function f_ε. (b) Box-cover produced by GreedyBox.

Fig. 2: Plot of the worst-case function f_ε for $\varepsilon = 0.05$ and box-cover produced by GreedyBox after $t = 6$ iterations.

the bound of $\Omega(\varepsilon^{-1/2})$ is not achievable. This demonstrates the influence of singularities, as C^k functions with no singularities can be approximated at a rate $\varepsilon^{-1/k}$. It is also worth pointing out that the above lower bound maybe easily extended to non-adaptive algorithms (considering Heaviside step adversarial functions) which would require $\Omega(\varepsilon^{-1})$ function evaluations. This underscores, in a new scenario, the need of adaptive algorithms to approximate functions with singularities [21].

Negative result for GreedyBox.

The previous result shows that GreedyBox is (up to logs) optimal for highly non-regular functions ($\alpha \geqslant 1/2$). We now consider the other regime ($\alpha < 1/2$) and prove an almost (up to logs) matching lower bound in the case $p = 1$, $\alpha = 0$. This, shows that the rate $\varepsilon^{-3/4}$ cannot be improved for GreedyBox for such classes of functions.

Proposition 7. *Let $\varepsilon \in (0, 1/12)$. Then, there exists a piecewise-C^2 function $f_\varepsilon : [0,1] \to [0,1]$ with $|f''(x)| \leqslant 1$ whenever it is defined and one C^1-singularity, such that there exists $t \geqslant 2^{-7}\varepsilon^{-3/4}$ with $\left\|\widehat{f}_t - f_\varepsilon\right\|_1 > \varepsilon$, where $\widehat{f}_t$ is the* GreedyBox *approximation after t rounds.*

An example of the worst-case function f_ε built in the proof of Proposition 7 for $\varepsilon = 0.005$, that exhibits poor performance for GreedyBox at $t = 6$, is depicted in Figure 2. The function is formally defined in the proof and consists of two parts: one that oscillates with $|f''(t)| = 1$ for $x \leqslant 1/2$, and the other that is linear for $x > 1/2$. The function is constructed in such a way that at a certain t ($t = 6$ in Fig. 2), GreedyBox selects its points precisely between the oscillations, and focuses too much on the linear part, resulting in a maximum possible L^1 error. It is worth noting that for each value of ε, it is possible to construct an adversarial function f_ε. However, an interesting question arises: can a single function f be devised to work uniformly for all values of ε? We believe that this question is challenging and connected to the 10^{th} open problem raised in [8, Chapter 10].

GreedyWidthBox: an optimal modification of GreedyBox.

GreedyBox can actually be adapted to achieve the optimal rate of $\widetilde{\mathcal{O}}\big(\varepsilon^{-1+(\frac{1-\alpha}{1+p})_+}\big)$ simultaneously for all $\alpha \geqslant 0$, while maintaining our adaptive guarantee in terms of

Algorithm 3: GreedyWidthBox

Input: $\varepsilon > 0$, $p \geqslant 1$

Init: Set $t = 1$, $x_0 = b_0^1 = 0$ and $x_1 = b_1^1 = 1$, evaluate $f(0)$ and $f(1)$, and define $\xi_1 = (a_1^1)^p = (f(1) - f(0))^p$;

while $\xi_t > \varepsilon^p$ **do**

1. **if** *t is even* **then**
   Select the box with the largest width: find k_*^t that maximizes $(b_k^t - b_{k-1}^t)$.
   else
   Select the box with the largest area: find k_*^t that maximizes a_k^t.
   end
2. Evaluate f at the midpoint $x_{t+1} := \big(b_{k_*^t-1}^t + b_{k_*^t}^t\big)/2$.
3. Sort the points $x_0, x_1, \ldots, x_{t+1}$ in increasing order:
   $$b_0^{t+1} = 0 \leqslant b_1^{t+1} < \cdots < b_{t+1}^{t+1} = 1.$$
4. Define the generalized areas for all $k \in \{1, \ldots, t+1\}$ by
   $$a_k^{t+1} = (b_k^{t+1} - b_{k-1}^{t+1})^{1/p}(f(b_k^{t+1}) - f(b_{k-1}^{t+1})).$$
5. Update the certificate
   $$\xi_{t+1} = \sum_{k=1}^{t+1} (a_k^{t+1})^p \,.$$
6. Let $t \leftarrow t + 1$.

end

Set $\tau_\varepsilon = t$ and approximate f with the piecewise-affine function $\widehat{f}_{\tau_\varepsilon}$ defined by:

$$x \in [0,1] \qquad \widehat{f}_{\tau_\varepsilon}(x) = \frac{f(b_k^{\tau_\varepsilon}) - f(b_{k-1}^{\tau_\varepsilon})}{b_k^{\tau_\varepsilon} - b_{k-1}^{\tau_\varepsilon}}(x - b_{k-1}^{\tau_\varepsilon}) + f(b_{k-1}^{\tau_\varepsilon}) \,,$$

for $k \in \{1, \ldots, \tau_\varepsilon\}$ such that $b_{k-1}^{\tau_\varepsilon} \leqslant x \leqslant b_k^{\tau_\varepsilon}$;

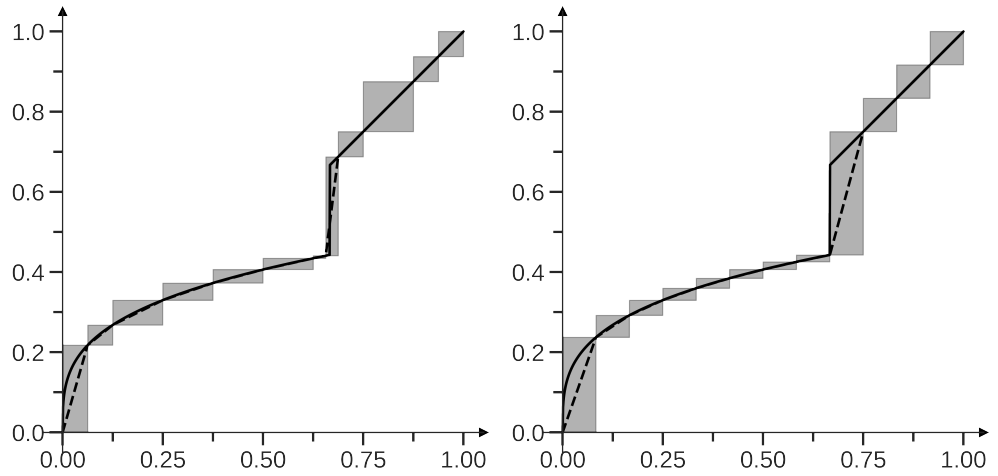
Output: $(\tau_\varepsilon, (x_t)_{1 \leqslant t \leqslant \tau_\varepsilon}, \widehat{f}_{\tau_\varepsilon})$.

$\mathcal{N}_p(f, \varepsilon)$. This can be accomplished by employing the GreedyBox approach for half of the iterations and the trapezoidal rule for the remaining half (see Algorithm 3). The proof, left to the reader, follows closely the one of Theorem 5, in which the upper bound (B9) can be simplified by utilizing the fact that, after conducting t function evaluations, the trapezoidal rule ensures that all widths w_i are at most of the order t^{-1}. In particular, for $p = 1$ and $\alpha = 0$, this provides an algorithm that achieves ε-acccuracy in L^1 norm in $\widetilde{\mathcal{O}}(\varepsilon^{-1/2})$ function evaluations for non-decreasing piecewise-C^2 functions, as soon as the number of C^1-singularities remains constant as $\varepsilon \to 0$.

5 Numerical Experiments

In this section, we study empirically the performance of GreedyBox as compared to the trapezoidal rule. All the experiments are run using $p = 1$ and the Lebesgue measure

for μ. Figure 3 displays the output given by both GreedyBox and the trapezoidal rule on the function f defined by $f(x) = \frac{1}{2}x^{3/10}$ if $x \leqslant \frac{2}{3}$ and $f(x) = x$ otherwise. One can see that on this example the L^1 error of GreedyBox is twice as small as that of the trapezoidal rule. In general, GreedyBox copes far better with discontinuities than the trapezoidal rule.



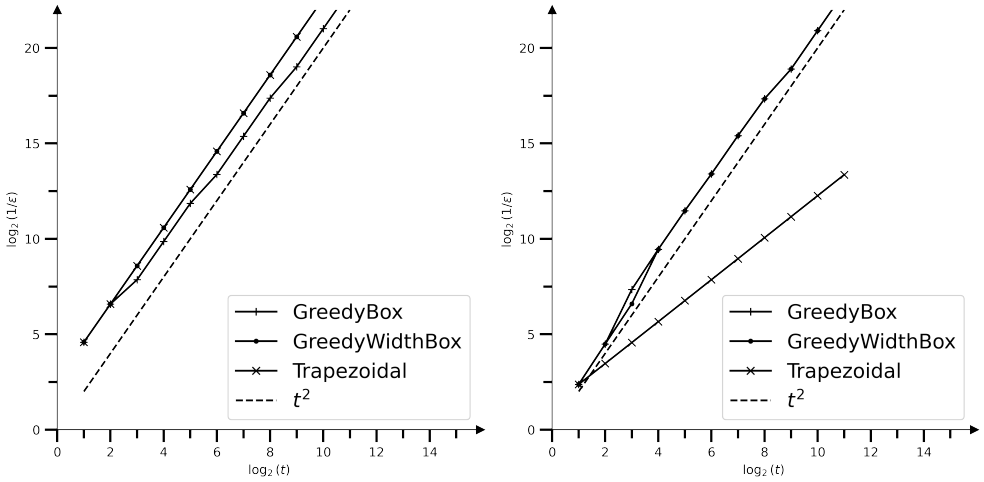
(a) Error of GreedyBox after 12 iterations: 0.005

(b) Error of the trapezoidal rule after 12 iterations: 0.015

Fig. 3: Comparison of GreedyBox and the trapezoidal rule on a piecewise-C^2 function after 12 iterations

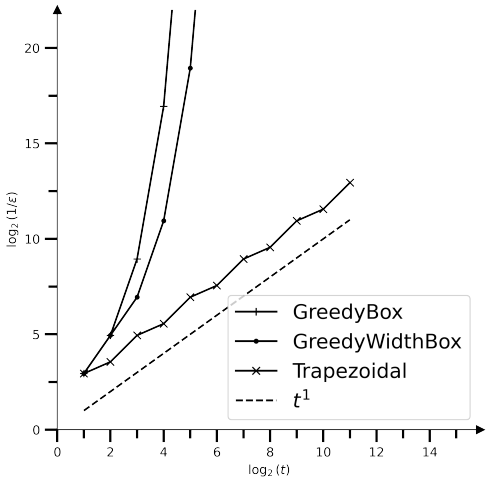
Remark that the trapezoidal rule is usually an offline algorithm that needs the total number t of iterations from the beginning. Fortunately, it can easily be adapted to an online version built on the same model as GreedyBox. Instead of choosing the box with the largest area on Step 1 of GreedyBox, it picks the box with the largest width. This online version matches exactly the offline trapezoidal rule whenever t is a power of 2 and allows for a better comparison with GreedyBox. It is this online version that we use in the next experiments.

The trapezoidal rule is known to have an L^1 error that decreases linearly with the number t of iterations for any non-decreasing functions. This is the same speed of convergence that we proved for GreedyBox in Theorem 2. However, for C^2 functions, the L^1 error of the trapezoidal rule decreases quadratically with the number of iterations, which corresponds to a sample complexity $\varepsilon^{-1/2}$. This is better than the upper bound in $\varepsilon^{-3/4}$ proven in Theorem 5 for GreedyBox. Remember however than no lower bound of order $\varepsilon^{-3/4}$ was proved so far for GreedyBox on C^2 functions with no singularities. Also note that the trapezoidal rule achieves a rate of $\varepsilon^{-1/2}$ only for C^2 functions, but can have an error of order ε^{-1} as soon as the function is discontinuous (see Figure 4c).

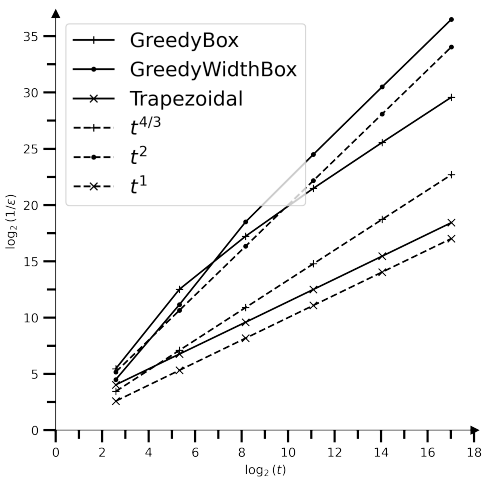


(a) Error rate on $f\colon x \mapsto x^2$

(b) Error rate on $f\colon x \mapsto x^{1/10}$



(c) Error rate on $f(x) = \mathbb{1}_{\{x \geqslant 0.3\}}$



(d) Error rate on g^t

Fig. 4: Comparison of GreedyBox and GreedyWidthBox with the trapezoidal rule. Logarithmic scale of the inverse of the error w.r.t. the number of evaluations.

In order to have best-of-both-worlds theoretical results, we introduced a new algorithm called GreedyWidthBox (Algorithm 3), which achieves the asymptotic rate $\mathcal{O}(\varepsilon^{-1/2})$ for piecewise-C^2 functions, as soon as the number of pieces is finite. In Figure 4, we run the three algorithms for a fixed number t of epochs, and observe the L^1 distance between the approximated function $\widehat{f_t}$ returned by the algorithm and the true function. Figures 4a-4c consider three different functions f with various regularities. In Figure 4d, we examine time-dependent piecewise-C^2 functions

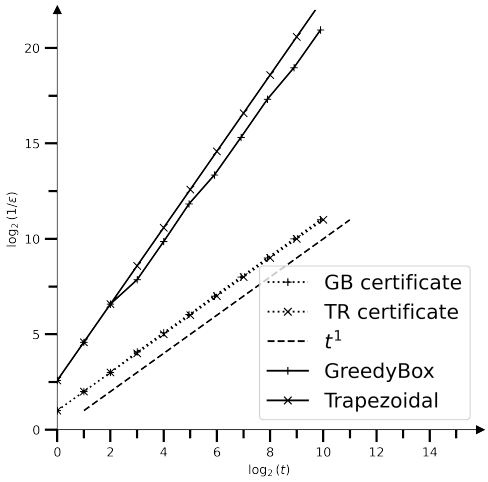
$g^t : x \mapsto \frac{1}{2} f^t(2x) \mathbb{1}_{x \leqslant 1/2} + \mathbb{1}_{x > 1/2}$. Here, f^t corresponds to the worst-case function as defined in (B17) in the proof of Proposition 7, with an additional introduced discontinuity. For a better understanding of the results, we plot the inverse of this error, and plot everything with logarithmic scale. This means that a straight line with slope 1 represents a linear speed of convergence: an error that decreases inversely proportionally with the number t of epochs. Figure 4d precisely confirms the anticipated worst-case rates as determined by the analysis for monotone piecewise-C^2 functions.

Remember however that given a desired precision ε, GreedyBox does not stop when its L^1 error is smaller than ε, but when its certificate (the best upper bound it can get without knowing *a priori* the function) is smaller than ε. Thus, for computation comparison, what really matters is to plot the convergence of the certificate with regard to some target error ε. The certificate of GreedyBox appears to be smaller than both GreedyWidthBox and the trapezoidal rule, regardless of the smoothness of the function. Yet, for most of the existing functions, the certificate of GreedyBox is of the order of ε^{-1}. However, remark that the same bound apply both for the trapezoidal rule and for GreedyWidthBox. This behavior can be seen on Figure 5 for GreedyBox and the trapezoidal rule. In Figure 5d, we can see that GreedyBox provides certification of ε-accuracy approximately 30 times as fast as the trapezoidal rule. For C^2 functions, this certificate will never be a good upper bound of the real error, that often is of order $\varepsilon^{-1/2}$ for all three algorithms.

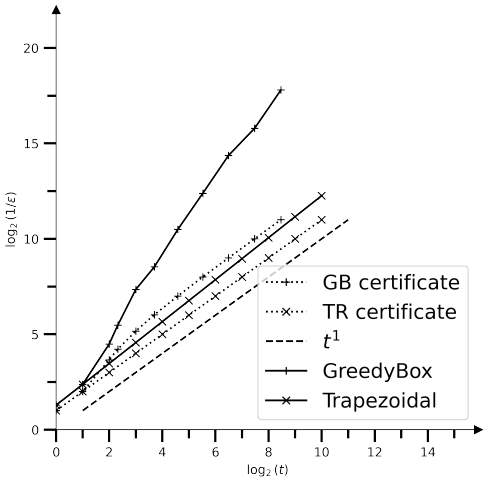
6 Conclusion and future works

In this paper, we studied the problem of approximating a non-decreasing function f in $L^p(\mu)$ norm, with sequential (adaptive) access to its values. We first proved an f-dependent lower bound on the stopping time that holds for all algorithms with guaranteed $L^p(\mu)$ error after stopping. We then presented the GreedyBox algorithm (inspired from Novak [1]) and showed that up to logarithmic factors, it is optimal among all such algorithms, for each non-decreasing function f. As a direct consequence, we showed for the integral estimation problem that GreedyBox can be combined with additional randomization to get an improved rate in expectation. For the L^p-approximation problem, we also investigated to what extent the $L^p(\mu)$ error of GreedyBox can decrease faster (than guaranteed by the algorithm) for piecewise-C^2 functions. Put briefly, up to logarithmic factors, GreedyBox automatically achieves the improved (and optimal) rate of $\varepsilon^{-1+(\frac{1-\alpha}{1+p})_+}$ for a large number $\varepsilon^{-\alpha}$ of singularities, $\alpha \geqslant 1/2$, and a simple algorithmic variant (GreedyWidthBox) achieves this rate for any value of α. In particular, our results highlight multiple performance gaps between adaptive and non-adaptive algorithms, C^k and piecewise-C^k functions, as well as monotone or non-monotone functions. We also provided numerical experiments to illustrate our theoretical results.

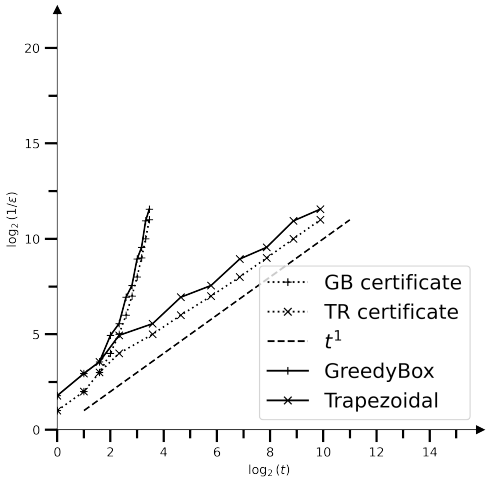
Several interesting questions about GreedyBox are left open. First, similarly to the faster rates proved for piecewise-C^2 functions, it would be interesting to investigate optimal rates for the average $L^p(\mu)$ error, when f is drawn at random from a probability distribution over the set of monotone functions (as, e.g., in Novak [1]). Second,



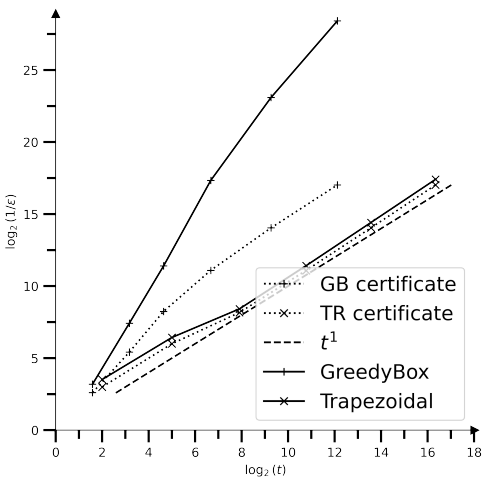
(a) Certificate and error on $f\colon x \mapsto x^2$



(b) Certificate and error on $f\colon x \mapsto x^{1/10}$



(c) Certificate and error on $f(x) = \mathbb{1}_{\{x \geqslant 0.3\}}$



(d) Certificate and error on g^t

Fig. 5: Comparison of GreedyBox with the trapezoidal rule. Logarithmic scale of the inverse of the error w.r.t the number of evaluations.

for any fixed non-decreasing function f, it would be useful to derive the limit (if any) of the empirical distribution of the points queried by GreedyBox. When $p = 1$ and μ is the Lebesgue measure on $[0, 1]$, we conjecture that this limit exists and has a density roughly proportional to the square-root of the derivative of f almost everywhere. Solving this problem would help complete our understanding of the behavior of GreedyBox, and of how it precisely adapts to the complexity of any function f.

Several generalizations would also be worth investigating in the future. A seemingly straightforward direction is to work with Lipschitz functions on compact intervals; an algorithm defined similarly to GreedyBox but with parallelograms instead of rectangular boxes seems perfectly fit for this problem. Other natural directions consist in extending our f-dependent bounds to multivariate monotone functions (in the spirit of, e.g., Papageorgiou [7]), to functions of known bounded variation, or variants of these function classes (e.g., entirely monotone functions, functions of bounded Hardy-Krause variation, see [29]).

Another natural and interesting research avenue is to address the case of noisy evaluations of f. We believe that, under some known assumptions on the noise distribution, similar sample complexity guarantees (with slower rates) can be achieved by using a mini-batch variant of GreedyBox, which reduces the impact of noise by sampling multiple points and computing the average. Solving this problem would be a way to efficiently estimate cumulative distribution functions under a special censored feedback in the same spirit as in Abernethy et al. [30]. A variant of GreedyBox has been designed and analysed by [28] for addressing imperfect observations. However, the authors have solely established asymptotic convergence garantees (both point-wise and in terms of L^1 or L^∞ norms) when the number of function evaluations approaches infinity, and they do not offer any guarantees regarding sample complexity.

7 Acknowledgements

The authors would like to thank Wouter Koolen for insightful discussions at the beginning of the work, as well as Peter Bartlett for fruitful feedback about improved rates for piecewise-regular functions. This work has benefited 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 and Étienne de Montbrun gratefully acknowledge the support of IRT Saint Exupéry and the DEEL project.[9]

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

Appendix A Elementary properties of box-covers

A.1 A general upper bound (proof of Lemma 1)

Lemma 1. *For all non-decreasing functions $f : [0,1] \to [0,1]$ and $\varepsilon > 0$, the quantity $\mathcal{N}_p(f,\varepsilon)$ is well defined and upper bounded by*

$$\mathcal{N}_p(f,\varepsilon) \leqslant \lceil 1/\varepsilon \rceil \ .$$

Proof. Let $n = \lceil 1/\varepsilon \rceil \geqslant 1$. In order to prove the lemma, we exhibit a sequence $\mathcal{B}$ of at most n boxes, show that it is a box-cover of f, and that the generalized areas of the boxes B are such that $\left(\sum_B \mathcal{A}_p(B)^p\right)^{1/p} \leqslant \varepsilon$.

For $i \in \{1,\dots,n\}$ let $x_i = \sup\{0 \leqslant x \leqslant 1 \colon f(x) \leqslant \frac{i}{n}\}$. We also set $x_0 = 0$. Note that $x_n = 1$. For $i \in \{1,\dots,n\}$, we define $B_i = [x_{i-1}, x_i] \times \left[\frac{i-1}{n}, \frac{i}{n}\right]$. Without loss of generality, we assume that $x_{i-1} < x_i$ for all $i \in \{1,\dots,n\}$. (Otherwise, we just remove all B_i's such that $x_{i-1} = x_i$, and remark below that the B_i's still cover the graph of f except maybe at the x_i's.)

We start by showing that $\mathcal{B} = (B_i)_{1 \leqslant i \leqslant n}$ is a box-cover of f. First, $B_1, \dots, B_n$ are adjacent boxes by construction. Note also that the graph of f is included in the union of the B_i's except maybe at the x_i's. Indeed, for all $i \in \{1,\dots,n\}$ and all $x \in (x_{i-1}, x_i)$, we have $\frac{i-1}{n} \leqslant f(x) \leqslant \frac{i}{n}$ by definition of x_{i-1} and x_i, and by monotonicity of f. Therefore, $(x, f(x)) \in B_i$. This proves that $\mathcal{B}$ is a box-cover of f.

We are left to show that $\left(\sum_{i=1}^n \mathcal{A}_p(B_i)^p\right)^{1/p} \leqslant \varepsilon$. We have

$$\sum_{i=1}^n \mathcal{A}_p(B_i)^p = \sum_{i=1}^n \left(\frac{i}{n} - \frac{i-1}{n}\right)^p \mu((x_{i-1}, x_i)) = \frac{1}{n^p} \sum_{i=1}^n \mu((x_{i-1}, x_i)) \leqslant \frac{1}{n^p} \ .$$

This entails that $\left(\sum_{i=1}^n \mathcal{A}_p(B_i)^p\right)^{1/p} \leqslant \frac{1}{n} \leqslant \varepsilon$, which concludes the proof. □

A.2 Technical Lemmas on box-covers

We state here two elementary lemmas about box-covers. The first one below indicates that box-covers of f with desired properties exist not only for $n = \mathcal{N}_p(f,\varepsilon)$ (or $n = \mathcal{N}'_p(f,\varepsilon)$) but also for all larger values of n.

Lemma 6. *Let $\varepsilon > 0$ and $f : [0,1] \to [0,1]$ be any non-decreasing function. Then,*

- *for all $n \geqslant \mathcal{N}_p(f,\varepsilon)$, there exists a box-cover $B_1, \dots, B_n$ of f such that $\left(\sum_{i=1}^n \mathcal{A}_p(B_i)^p\right)^{1/p} \leqslant \varepsilon$;*
- *for all $n \geqslant \mathcal{N}'_p(f,\varepsilon)$, there exists a box-cover $B_1, \dots, B_n$ of f such that $\mathcal{A}_p(B_i) \leqslant \varepsilon$ for all $i = 1, \dots, n$.*

Proof. The proof of the two items is straightforward. For the first one, consider any box-cover $B_1, \dots, B_N$ of f, with $N = \mathcal{N}_p(f,\varepsilon)$, such that $\left(\sum_{i=1}^N \mathcal{A}_p(B_i)^p\right)^{1/p} \leqslant \varepsilon$.

Then, split the last box B_N vertically into $n - N + 1$ sub-boxes $B'_N, \ldots, B'_n$. We can immediately see that

$$\left(\sum_{i=1}^{N-1} \mathcal{A}_p(B_i)^p + \sum_{i=N}^{n} \mathcal{A}_p(B'_i)^p \right)^{1/p} \leqslant \left(\sum_{i=1}^{n} \mathcal{A}_p(B_i)^p \right)^{1/p} \leqslant \varepsilon \,,$$

so that the sequence of boxes $B_1, \ldots, B_{N-1}, B'_N, \ldots, B'_n$ is a box-cover of f with cardinality n that satisfies the desired generalized area property. The second item can be proved similarly. □

The next simple lemma indicates that the values y_j^- and y_j^+ in any box-cover of f are necessary lower and upper bounds on the values of f inside the boxes.

Lemma 7. *Let $f : [0,1] \to [0,1]$ be non-decreasing, and $B_1, \ldots, B_n$ be any box-cover of f, with $B_j = [c_{j-1}, c_j] \times [y_j^-, y_j^+]$ for all $j = 1, \ldots, n$, for some sequence $0 = c_0 < \ldots < c_n = 1$. Then, for all $j = 1, \ldots, n$,*

$$y_j^- \leqslant \inf_{x > c_{j-1}} f(x) \qquad \text{and} \qquad y_j^+ \geqslant \sup_{x < c_j} f(x) \,.$$

Proof. Let $x \in (c_{j-1}, c_j)$. Since $(x, f(x)) \in \cup_{j'=1}^{n} B_{j'}$ and $x \notin [c_{k-1}, c_k]$ for any $k \neq j$, we have $(x, f(x)) \in B_j$ and therefore

$$y_j^- \leqslant f(x) \qquad \text{and} \qquad y_j^+ \geqslant f(x) \,.$$

We conclude by taking the infimum or supremum over all $x \in (c_{j-1}, c_j)$ and by using the fact that f is non-decreasing. □

A.3 Relationship between $\mathcal{N}_p(f, \varepsilon)$ and $\mathcal{N}'_p(f, \varepsilon)$

In this section, we compare the two complexity quantities $\mathcal{N}_p(f, \varepsilon)$ and $\mathcal{N}'_p(f, \varepsilon)$ defined in Section 1.2.

The quantity $\mathcal{N}'_p(f, \varepsilon)$ always satisfies $\mathcal{N}'_p(f, \varepsilon) \leqslant \mathcal{N}_p(f, \varepsilon)$. The next lemma relates the two complexity notions in a tighter way. Intuitively, if all the boxes B of a minimal box-cover of f satisfying $\sum_B \mathcal{A}_p(B)^p \leqslant \varepsilon^p$ had similar generalized areas $\mathcal{A}_p(B)$, these areas would be close to $\varepsilon / \mathcal{N}_p(f, \varepsilon)^{1/p}$, so that

$$\mathcal{N}'_p\big(f, \varepsilon / \mathcal{N}_p(f, \varepsilon)^{1/p}\big) \lesssim \mathcal{N}_p(f, \varepsilon).$$

The following lemma implies that indeed

$$\mathcal{N}'_p\big(f, \varepsilon / \mathcal{N}_p(f, \varepsilon)^{1/p}\big) \approx \mathcal{N}_p(f, \varepsilon)$$

up to a factor of 2.

Lemma 8. *Let $\varepsilon > 0$ and $n \geqslant \mathcal{N}_p(f,\varepsilon)$. Then,*

$$\mathcal{N}'_p\Big(f, \frac{\varepsilon}{n^{1/p}}\Big) \leqslant 2n \qquad \text{and} \qquad \mathcal{N}_p(f,\varepsilon) \leqslant \mathcal{N}'_p\Big(f, \frac{\varepsilon}{\mathcal{N}_p(f,\varepsilon)^{1/p}}\Big)\,.$$

In the proof of Theorem 2 we only use the first inequality $\mathcal{N}'_p\big(f, \frac{\varepsilon}{n^{1/p}}\big) \leqslant 2n$. The second inequality shows that this step is tight up to a constant of 2.

Proof. Let $f : [0,1] \to [0,1]$ be a non-decreasing function and $\varepsilon > 0$.

(a) We prove that $\mathcal{N}'_p(f, \varepsilon/n^{1/p}) \leqslant 2n$. For the sake of readability, we assume that μ is the Lebesgue measure on $[0,1]$, and will explain at the end how to adapt the proof for an arbitrary probability measure μ. By $n \geqslant \mathcal{N}_p(f,\varepsilon)$ and Lemma 6, there exists a box-cover $B_1, \dots, B_n$ of f satisfying $\sum_{i=1}^n \mathcal{A}_p(B_i)^p \leqslant \varepsilon^p$. Following a technique from Novak [1, Proof of Theorem 3], we now divide the B_i's into as many sub-boxes as necessary so that each of their generalized areas is below $\varepsilon/n^{1/p}$. We will show that the overall number of resulting boxes is at most $2n$.

Let $i \in \{1, \dots, n\}$. We split the box $B_i = [x^-, x^+] \times [y^-, y^+]$ in a vertical fashion by splitting $[x^-, x^+]$ into $k_i = \big\lfloor n\mathcal{A}_p(B_i)^p/\varepsilon^p \big\rfloor + 1$ intervals of equal sizes $[x_0, x_1], \dots, [x_{k-1}, x_{k_i}]$, where $x_j = x^- + j(x^+ - x^+)/k_i$. The choice of k_i ensures that $\mathcal{A}_p(B_i)^p \leqslant k_i\varepsilon^p/n$.

Then, we define the smaller boxes $B_i^{(j)} := [x_{j-1}, x_j] \times [y^-, y^+]$ for $j = 1, \dots, k_i$. Their generalized areas are given by $\mathcal{A}_p(B_i^{(j)}) = ((x^+ - x^-)/k_i)^{1/p}(y^+ - y^-) = \mathcal{A}_p(B_i)/k_i^{1/p} \leqslant \varepsilon/n^{1/p}$ as required. Furthermore, the new total number of boxes is

$$\sum_{i=1}^n k_i = \sum_{i=1}^n \left(\left\lfloor \frac{n\mathcal{A}_p(B_i)^p}{\varepsilon^p} \right\rfloor + 1 \right) \leqslant n + \frac{n}{\varepsilon^p} \sum_{i=1}^n \mathcal{A}_p(B_i)^p \leqslant 2n\,.$$

Since $\cup_{i=1}^n \cup_{j=1}^{k_i} B_i^{(j)} = \cup_{i=1}^n B_i$ contains the graph of f except maybe at the B_i's endpoints (and thus everywhere outside the $B_i^{(j)}$'s endpoints), we have shown that $\mathcal{N}'_p\big(f, \varepsilon/n^{1/p}\big) \leqslant 2n$, as desired.

When μ is not the Lebesgue measure on $[0,1]$, the proof can be slightly adapted as follows. For each box $B_i = [x^-, x^+] \times [y^-, y^+]$ as above, we distinguish two cases.
Case 1: if $\mu\big((x^-, x^+)\big) = 0$, we set $k_i = 1$, $x_0 = x^-$, and $x_1 = x^+$ as above.
Case 2: if $\mu\big((x^-, x^+)\big) > 0$, we set $k_i = \big\lfloor n\mathcal{A}_p(B_i)^p/\varepsilon^p \big\rfloor + 1$, $x_0 = x^-$, and $x_{k_i} = x^+$ as above. However, for any $1 \leqslant j < k_i$ (if any), we take x_j as a quantile of order j/k_i of the conditional distribution $\mu\big(\cdot|(x^-, x^+)\big)$. Then, among all sub-boxes $B_i^{(j)} := [x_{j-1}, x_j] \times [y^-, y^+]$, $j = 1, \dots, k_i$, we only keep the non-degenerate ones, i.e., those such that $x_{j-1} < x_j$.

The rest of the proof remains unchanged. In particular, each remaining sub-box satisfies $\mathcal{A}_p(B_i^{(j)}) \leqslant \varepsilon/n^{1/p}$, and the total number of sub-boxes is at most of $\sum_{i=1}^n k_i \leqslant 2n$.

(b) Before proving the second inequality, we prove an intermediate result: for all $n_1 \geqslant \mathcal{N}'_p(f,\varepsilon)$, we have

$$\mathcal{N}_p(f, n_1^{1/p}\varepsilon) \leqslant n_1 \,. \tag{A1}$$

Since $n_1 \geqslant \mathcal{N}'_p(f,\varepsilon)$, we can consider a box-cover $B_1,\dots,B_{n_1}$ of f with generalized areas at most of ε each. The result immediately follows from

$$\Big(\sum_{i=1}^{n_1} \mathcal{A}_p(B_i)^p\Big)^{1/p} \leqslant n_1^{1/p}\varepsilon\,.$$

(c) We write $n_\varepsilon = \mathcal{N}_p(f,\varepsilon)$ for simplicity. We prove that $n_\varepsilon \leqslant \mathcal{N}'_p(f,\varepsilon/n_\varepsilon^{1/p})$ by contradiction. Let us assume that $\mathcal{N}'_p(f,\varepsilon/n_\varepsilon^{1/p}) < n_\varepsilon$ and define

$$\varepsilon' := \varepsilon \cdot \left(\frac{\mathcal{N}'_p\big(f,\varepsilon/n_\varepsilon^{1/p}\big)}{n_\varepsilon}\right)^{1/p} < \varepsilon\,.$$

Then using Inequality (A1), with ε substituted with $\varepsilon/n_\varepsilon^{1/p}$, we get

$$\mathcal{N}_p(f,\varepsilon) \overset{\varepsilon'<\varepsilon}{\leqslant} \mathcal{N}_p(f,\varepsilon') = \mathcal{N}_p\left(f, \mathcal{N}'_p\Big(f, \frac{\varepsilon}{n_\varepsilon^{1/p}}\Big)^{1/p} \frac{\varepsilon}{n_\varepsilon^{1/p}}\right) \overset{\text{(A1)}}{\leqslant} \mathcal{N}'_p\Big(f, \frac{\varepsilon}{n_\varepsilon^{1/p}}\Big)\,,$$

which contradicts the assumption and thus proves the desired inequality.

□

Appendix B Omitted proofs

B.1 Proof of Lemma 2

Lemma 2. *Let $f : [0,1] \to [0,1]$ be non-decreasing, $p \geqslant 1$ and $\varepsilon \in (0,1]$. For any $t \in \{1,\dots,\tau_\varepsilon\}$,*

$$\big\|\widehat{f}_t - f\big\|_p^p := \int_0^1 \Big|\widehat{f}_t(x) - f(x)\Big|^p \, d\mu(x) \leqslant \sum_{k=1}^{t} (a_k^t)^p =: \xi_t \,.$$

Proof. Remark that the approximation $\widehat{f}_t$ is the continuous piecewise-affine function such that $\widehat{f}_t(b_k^t) = f(b_k^t)$ for all $k \in \{0,\dots,t\}$. Therefore, it suffices to show that for each $k = 1,\dots,t$,

$$\int_{(b_{k-1}^t, b_k^t)} \Big|\widehat{f}_t(x) - f(x)\Big|^p \, d\mu(x) \leqslant (a_k^t)^p \,,$$

which follows easily from the fact that $\Big|\widehat{f}_t(x) - f(x)\Big| \leqslant f(b_k^t) - f(b_{k-1}^t)$ on the k-th box (since f is non-decreasing), and by definition of $a_k^t := (\mu(b_{k-1}^t, b_k^t))^{1/p}(f(b_k^t) - f(b_{k-1}^t))$. Summing over $k = 1,\dots,t$ concludes the proof. □

As a supplement, we are left with proving the special case of the Lebesgue measure, for which the result holds with a multiplicative factor of $1/(p+1)$.

Lemma 9. *Let $f : [0,1] \to [0,1]$ be non-decreasing. At any round $t \geqslant 1$,*

$$\left\|\widehat{f}_t - f\right\|_p^p := \int_0^1 \left(\widehat{f}_t(x) - f(x)\right)^p dx \leqslant \frac{1}{1+p} \sum_{k=1}^{t} (a_k^t)^p \,.$$

Proof. Summing over $k = 1, \dots, t$ it suffices to show that for each $k = 1, \dots, t$,

$$\int_{b_{k-1}^t}^{b_k^t} \left|\widehat{f}_t(x) - f(x)\right|^p dx \leqslant \frac{(a_k^t)^p}{1+p} \,.$$

Let $k \in \{1, \dots, t\}$. To ease the notation, we make a change of variables and define for all $u \in [0,1]$:

$$g(u) := \frac{f\big((b_k^t - b_{k-1}^t)u + b_{k-1}^t\big) - f(b_{k-1}^t)}{f(b_k^t) - f(b_{k-1}^t)} \,.$$

We have

$$\begin{aligned}\int_{b_{k-1}^t}^{b_k^t} \left|\widehat{f}_t(x) - f(x)\right|^p dx &= \big(f(b_k^t) - f(b_{k-1}^t)\big)^p (b_k^t - b_{k-1}^t) \int_0^1 |u - g(u)|^p \, du \\ &= (a_k^t)^p \int_0^1 |u - g(u)|^p \, du \,.\end{aligned}$$

The function g is non-decreasing over $[0,1]$ with $g(0) = 0$ and $g(1) = 1$. It remains to control $\int_0^1 |u - g(u)|^p \, du$ by $1/(1+p)$. which is done in the following.

First we remark that we can assume $g(1/2) \neq 1/2$. Otherwise, since g is non-decreasing $|g(u) - u| \in [0, 1/2]$ for all $u \in [0,1]$ and $\int_0^1 |u - g(u)|^p \, du \leqslant 2^{-p} \leqslant (1+p)^{-1}$, which concludes.

We consider the two points $u_- \leqslant 1/2 \leqslant u_+$ such that the sign of $g(x) - x$ does not change over (u_-, u_+). More formally, they are defined as:

$$u_- = \inf\Big\{u \in [0, 1/2] : \quad \forall x \in (u, 1/2] \quad \big(g(1/2) - 1/2\big)\big(g(x) - x\big) > 0\Big\}$$

and

$$u_+ = \sup\Big\{u \in [1/2, 1] : \quad \forall x \in [1/2, u) \quad \big(g(1/2) - 1/2\big)\big(g(x) - x\big) > 0\Big\} \,.$$

Note that the sign of $g(x) - x$ is constant over $\{1/2\} \cup (u_-, u_+)$ since $g(1/2) \neq 1/2$. Now, we show the following two facts

$$\forall u < u_-, \quad g(u) \leqslant u_- \qquad \text{and} \qquad \forall u > u_+, \quad g(u) \geqslant u_+ \,. \tag{B2}$$

Indeed, let $u < u_-$, by definition of u_-, it exists $x \in [u, u_-]$ such that

$$\big(g(1/2) - 1/2\big)\big(g(x) - x\big) \leqslant 0\,,$$

and thus using again the definition of u_- for all $x' \in (u_-, 1/2]$, $\big(g(x') - x'\big)\big(g(x) - x\big) \leqslant 0$. If $g(x) \leqslant x$, we are done since $g(u) \leqslant g(x) \leqslant x \leqslant u_-$ because g is non-decreasing and $u \leqslant x \leqslant u_-$. Otherwise using $u \leqslant x'$, $g(u) \leqslant g(x') \leqslant x'$ and making $x' \to u_-$ concludes the first inequality of (B2). The second inequality can be proved similarly.

Therefore from (B2), for all $u \leqslant u_-$, $|g(u) - u| \leqslant u_-$ and all $u \geqslant u_+$, $|g(u) - u| \leqslant 1 - u_+$ which yields

$$\int_0^{u_-} |g(u) - u|^p \, du \leqslant u_-^{1+p} \quad \text{and} \quad \int_{u_+}^1 |g(u) - u|^p \, du \leqslant (1 - u_+)^{1+p}\,. \tag{B3}$$

Furthermore, $g(u) - u$ does not change sign over (u_-, u_+), which entails:

$$\int_{u_-}^{u_+} |g(u) - u|^p \, du \leqslant \max\left\{ \int_{u_-}^{u_+} (u_+ - u)^p du, \int_{u_-}^{u_+} (u - u_-)^p du \right\} = \frac{(u_+ - u_-)^{p+1}}{p+1}\,. \tag{B4}$$

Summing Inequalities (B3) and (B4), we get

$$\begin{aligned} \int_0^1 |g(u) - u|^p \, du &\leqslant u_-^{1+p} + (1 - u_+)^{1+p} + \frac{(u_+ - u_-)^{1+p}}{1+p} \\ &\leqslant \sup_{0 \leqslant u_- \leqslant 1/2 \leqslant u_+ \leqslant 1} \left\{ u_-^{1+p} + (1 - u_+)^{1+p} + \frac{(u_+ - u_-)^{1+p}}{1+p} \right\} = \frac{1}{1+p}\,. \end{aligned}$$

The supremum is reached for $(u_-, u_+) = (0, 1)$. □

B.2 Proof of Lemma 3

Lemma 3. *Let $f : [0,1] \to [0,1]$ be non-decreasing, $p \geqslant 1$ and $\varepsilon \in (0,1]$. Define $\tau'_\varepsilon := 2\big(1 + \lceil p \log_2(1/\varepsilon) \rceil\big)\mathcal{N}'_p\big(f, \varepsilon\big)$, and assume that* GreedyBox *is such that $\tau_\varepsilon > \tau'_\varepsilon$. Then, at time τ'_ε, all the boxes maintained by* GreedyBox *have a generalized area bounded from above by ε, i.e., $a_k^{\tau'_\varepsilon} \leqslant \varepsilon$ for all $k \in \{1, \ldots, \tau'_\varepsilon\}$.*

Proof. By definition of $m := \mathcal{N}'_p(f, \varepsilon)$, we can fix a box-cover $B_1, \ldots, B_m$ of f such that $\mathcal{A}_p(B_j) \leqslant \varepsilon$ for all j. Recall that these boxes are adjacent, i.e., they are of the form $B_j = [c_{j-1}, c_j] \times \big[y_j^-, y_j^+\big]$ for some nodes $0 = c_0 < \ldots < c_m = 1$. The inequalities $\mathcal{A}_p(B_j) \leqslant \varepsilon$ thus translate into $(y_j^+ - y_j^-)\mu\big((c_{j-1}, c_j)\big)^{1/p} \leqslant \varepsilon$ for all j.

In this proof we will compare the boxes maintained by GreedyBox with the boxes B_j above. We first need the following definition: at each round $t \geqslant 1$, we say that a box $B = \big[b_{k-1}^t, b_k^t\big] \times \big[f(b_{k-1}^t), f(b_k^t)\big]$, $1 \leqslant k \leqslant t$, maintained by GreedyBox is

- *internal* if $\big[b^t_{k-1}, b^t_k\big] \subset (c_{j-1}, c_j)$ for some $j = 1, \dots, m$;
- *overlapping* if $\big[b^t_{k-1}, b^t_k\big] \ni c_j$ for some $j = 1, \dots, m$.

Note that exactly one of the two cases above must hold true. In the sequel, we denote by $\widehat{B}_t = \big[b^t_{k^t_*-1}, b^t_{k^t_*}\big] \times \big[f\big(b^t_{k^t_*-1}\big), f\big(b^t_{k^t_*}\big)\big]$ the box selected by GreedyBox at time t. Since $\widehat{B}_t$ is necessary of one of the two types above, we can distinguish between the following two cases.

Case 1: there exists $t \in \{1, 2, \dots, \tau'_\varepsilon\}$ such that $\widehat{B}_t$ is internal. In this case, letting $j = 1, \dots, m$ be the corresponding index, we have, by Lemma 7 in Appendix A.2,

$$c_{j-1} < b^t_{k^t_*-1} < b^t_{k^t_*} < c_j \qquad \text{and} \qquad y_j^- \leqslant f\big(b^t_{k^t_*-1}\big) \leqslant f\big(b^t_{k^t_*}\big) \leqslant y_j^+ \,.$$

Therefore, the generalized areas of $\widehat{B}_t$ and B_j satisfy $a^t_{k^t_*} = \mathcal{A}_p\big(\widehat{B}_t\big) \leqslant \mathcal{A}_p(B_j) \leqslant \varepsilon$ by construction of B_j. Now, by definition of k^t_* in Algorithm 1:

$$\max_{1 \leqslant k \leqslant t} a^t_k = a^t_{k^t_*} \leqslant \varepsilon \,.$$

This concludes the proof of the lemma in this case, since the quantity $\max_{1 \leqslant k \leqslant t} a^t_k$ can only decrease between rounds t and τ'_ε.

Case 2: $\widehat{B}_t$ is an overlapping box at every round $t \in \{1, 2, \dots, \tau'_\varepsilon\}$. In this case, we say that all nodes c_j lying in $\big[b^t_{k^t_*-1}, b^t_{k^t_*}\big]$ are *activated* at time t. More precisely, we say that a node c_j is *left-activated* when $c_j \in \big(b^t_{k^t_*-1}, b^t_{k^t_*}\big]$ (part of the box $\widehat{B}_t$ lies on the left of c_j), and that it is *right-activated* when $c_j \in \big[b^t_{k^t_*-1}, b^t_{k^t_*}\big)$ (part of the box $\widehat{B}_t$ lies on the right of c_j).

Since $b^t_{k^t_*-1} < b^t_{k^t_*}$ by construction, at least one node c_j is either left-activated or right-activated at every round $t \in \{1, 2, \dots, \tau'_\varepsilon\}$, so that

$$\sum_{t=1}^{\tau'_\varepsilon} \underbrace{\sum_{j=1}^{m} \sum_{\sigma \in \{\text{left}, \text{right}\}} \mathbb{1}_{\{c_j \text{ is } \sigma\text{-activated at round } t\}}}_{\geqslant 1} \geqslant \tau'_\varepsilon \,.$$

Inverting sums and recognizing $N_{j,\sigma} = \sum_{t=1}^{\tau'_\varepsilon} \mathbb{1}_{\{c_j \text{ is } \sigma\text{-activated at round } t\}}$ to be the number of rounds when node c_j is σ-activated, we can see that

$$\sum_{j=1}^{m} \sum_{\sigma \in \{\text{left}, \text{right}\}} N_{j,\sigma} \geqslant \tau'_\varepsilon \,,$$

so that

$$\max_{1 \leqslant j \leqslant m} \max_{\sigma \in \{\text{left}, \text{right}\}} N_{j,\sigma} \geqslant \frac{\tau'_\varepsilon}{2\,m} \,.$$

Combining the last inequality with the definition of τ'_ε and $m := \mathcal{N}'_p(f,\varepsilon)$, we obtain the following intermediate result.

Fact 1. *There exists* $j \in \{1,\dots,\mathcal{N}'_p(f,\varepsilon)\}$ *and* $\sigma \in \{\text{left},\text{right}\}$ *such that the node* c_j *is* σ*-activated at least*

$$\frac{\tau'_\varepsilon}{2\,\mathcal{N}'_p(f,\varepsilon)} = 1 + \lceil p \log_2(1/\varepsilon) \rceil$$

times within the set of rounds $\{1,\dots,\tau'_\varepsilon\}$.

We now focus on a single pair (j,σ) provided by Fact 1 above. We follow the evolution of the box $\widetilde{B}_t$ maintained by GreedyBox that lies on the σ-side of c_j. More formally, if we write $\widetilde{B}_t = \big[b^t_{k-1}, b^t_k\big] \times \big[f\big(b^t_{k-1}\big), f\big(b^t_k\big)\big]$, this means that $c_j \in \big(b^t_{k-1}, b^t_k\big]$ if $\sigma = \text{left}$, and that $c_j \in \big[b^t_{k-1}, b^t_k\big)$ if $\sigma = \text{right}$. Note that, at any round t, such a box $\widetilde{B}_t$ indeed exists and is unique.

Note that, at all rounds $t \in \{1,\dots,\tau'_\varepsilon\}$ when c_j is σ-activated, we have $\widetilde{B}_t = \widehat{B}_t$, so that the box $\widetilde{B}_t$ is replaced (see Step 3 in Algorithm 1) by two boxes whose generalized widths are at most half that of $\widetilde{B}_t$. This is because x_{t+1} is a median of the conditional distribution $\mu(\cdot|(b^t_{k^t_*-1}, b^t_{k^t_*}))$. Since $\widetilde{B}_{t+1}$ is among these two boxes, we thus have

$$\text{width}\big(\widetilde{B}_{t+1}\big) \leqslant \text{width}\big(\widetilde{B}_t\big)/2$$

at each round $t \in \{1,\dots,\tau'_\varepsilon\}$ when c_j is σ-activated. Note also that $\widetilde{B}_{t+1} = \widetilde{B}_t$ at all other rounds $t \in \{1,\dots,\tau'_\varepsilon\}$; at such rounds, $\text{width}\big(\widetilde{B}_{t+1}\big) = \text{width}\big(\widetilde{B}_t\big)$. Combining the last two properties, and denoting by τ the round when c_j is σ-activated for the $\lceil p \log_2(1/\varepsilon) \rceil$-th time, we get

$$\text{width}\big(\widetilde{B}_{\tau+1}\big) \leqslant 2^{-\lceil p \log_2(1/\varepsilon) \rceil} \leqslant \varepsilon^p \,.$$

To conclude, denote by τ' the round when c_j is σ-activated for the $\big(\lceil p \log_2(1/\varepsilon) \rceil + 1\big)$-th time. Note that $\widehat{B}_{\tau'} = \widetilde{B}_{\tau'}$ so that

$$\text{width}\big(\widehat{B}_{\tau'}\big) = \text{width}\big(\widetilde{B}_{\tau'}\big) \leqslant \text{width}\big(\widetilde{B}_{\tau+1}\big) \leqslant \varepsilon^p \,,$$

where the first inequality follows from $\tau' \geqslant \tau + 1$ and the fact that $\text{width}\big(\widetilde{B}_t\big)$ is non-increasing over time. Therefore, by definition of $\widehat{B}_{\tau'}$, all boxes maintained by GreedyBox at time τ' have generalized areas bounded by

$$\max_{1 \leqslant k \leqslant \tau'} a_k^{\tau'} \leqslant \mathcal{A}_p\big(\widehat{B}_{\tau'}\big) \leqslant (1-0) \times \text{width}\big(\widehat{B}_{\tau'}\big)^{1/p} \leqslant \varepsilon \,.$$

We conclude the proof by noting that $\tau' \leqslant \tau'_\varepsilon$, so that $\max_{1 \leqslant k \leqslant \tau'_\varepsilon} a_k^{\tau'_\varepsilon} \leqslant \max_{1 \leqslant k \leqslant \tau'} a_k^{\tau'} \leqslant \varepsilon$. □

B.3 Proof of Lemma 4

Lemma 4. *Let $f : [0,1] \to [0,1]$ be non-decreasing, $p \geqslant 1$ and $\varepsilon \in (0,1]$. For any $t \in \{1, \dots, \lfloor \tau_\varepsilon/2 \rfloor\}$, we have $\xi_{2t} \leqslant \xi_t/2$. Therefore, for all $t \leqslant s$ in $\{1, \dots, \tau_\varepsilon\}$,*

$$\xi_s \leqslant \frac{\xi_t}{2^{\lfloor \log_2(s/t) \rfloor}} \leqslant \left(\frac{2t}{s}\right) \xi_t \,.$$

Proof. We first prove that $\xi_{2t} \leqslant \xi_t/2$ for all $t \in \{1, \dots, \lfloor \tau_\varepsilon/2 \rfloor\}$. To do so, recall that at any round $1 \leqslant t' \leqslant \tau_\varepsilon$, GreedyBox maintains t' boxes given by $\left[b^{t'}_{k-1}, b^{t'}_k\right] \times \left[f\left(b^{t'}_{k-1}\right), f\left(b^{t'}_k\right)\right]$, $k = 1, \dots, t'$. We write $a^{t'}_{(1)} \geqslant a^{t'}_{(2)} \geqslant \dots \geqslant a^{t'}_{(t')}$ for their generalized areas $a^{t'}_k$ sorted in decreasing order.

Part 1: We first show by induction on $k = 1, \dots, t$ that

(i) the box selected by GreedyBox at round $t - 1 + k$ (Step 1 in Algorithm 1) has a generalized area $a^{t-1+k}_{k^{t-1+k}_*}$ larger than or equal to $a^t_{(k)}$;
(ii) at least $t - k$ boxes of round $t + k$ are identical to boxes of round t.

Both (i) and (ii) are straightforward for $k = 1$. Assume they are true for some $k \in \{1, \dots, t-1\}$. Next we prove (i) and (ii) with the index value $k + 1$. At round $t + k$, the generalized area of the box selected by GreedyBox must be at least as large as the maximum generalized area of the $t - k$ boxes that are identical to boxes of round t (by (ii) with k). Therefore, this generalized area is larger than or equal to $a^t_{(k+1)}$, which proves (i). Property (ii) is immediate since only one box is selected at every round. This completes the induction.

Part 2: Note that, at each round $t-1+k \in \{t, t+1, \dots, 2t-1\}$, the box $\widehat{B}_{t-1+k}$ selected by GreedyBox is replaced with two smaller boxes B' and B'' whose generalized areas satisfy

$$\mathcal{A}_p(B')^p + \mathcal{A}_p(B'')^p \leqslant \mathcal{A}_p(\widehat{B}_{t-1+k})^p/2 \,,$$

since x_{t+1} is a median of $\mu(\cdot|(b^t_{k^t_*-1}, b^t_{k^t_*}))$ and $(\delta')^p + (\delta'')^p \leqslant (\delta' + \delta'')^p$ for any $\delta', \delta'' \geqslant 0$. Therefore, and by Property (i) above, at least $\mathcal{A}_p(\widehat{B}_{t-1+k})^p/2 \geqslant \left(a^t_{(k)}\right)^p/2$ is lost when summing the generalized areas to the power p at round $t + k$, compared to round $t + k - 1$. Therefore, the certificate ξ_{t+k} of the box-cover at the next round satisfies $\xi_{t+k} \leqslant \xi_{t-1+k} - (a^t_{(k)})^p/2$. Summing over $k = 1, \dots, t$ we get:

$$\xi_{2t} \leqslant \xi_t - \sum_{k=1}^{t} \frac{(a^t_{(k)})^p}{2} = \frac{\xi_t}{2} \,.$$

To see why this implies (5), it suffices to note that $s \geqslant \widetilde{s} := 2^{\lfloor \log_2(s/t) \rfloor} \cdot t$, so that

$$\xi_s \leqslant \xi_{\widetilde{s}} \leqslant \frac{\xi_t}{2^{\lfloor \log_2(s/t) \rfloor}} \,,$$

where the first inequality is because the certificate ξ_s can only decrease over time, and where the last inequality follows from the property $\xi_{2t} \leqslant \xi_t/2$ shown above. This concludes the proof. □

B.4 Proof of Theorem 4

Theorem 4. *Let $f : [0,1] \to [0,1]$ be non-decreasing which satisfies Assumption 1 for some $C, \alpha > 0$. Let $\varepsilon > 0$, then Algorithm 2 satisfies*

$$\mathbb{E}\Big[\big|\widehat{I}_{\tau_\varepsilon}(f) - I(f)\big|\Big] \leqslant \varepsilon.$$

Besides the number of function evaluations is bounded from above by

$$\tau_\varepsilon = \mathcal{O}\big(\log(1/\varepsilon)^{3/2}\varepsilon^{-\frac{1}{1/\alpha+1/2}}\big)\,.$$

Proof. Let $\varepsilon > 0$. First, we remark that the points $x_0, \dots, x_t$ defined by the deterministic version GreedyBox (defined in Algorithm 1) and the stochastic version (Algorithm 2) are identical. Only the stopping criterion and definition of $\widehat{I}_{\tau_\varepsilon}(f)$ differs from Algorithm 1. Therefore, we can apply Lemma 3.

Let $\varepsilon_1 \geqslant \varepsilon$ that will be fixed later as a function of ε. Denote $n_{\varepsilon_1} = C\varepsilon_1^{-\alpha} \geqslant \mathcal{N}(f, \varepsilon_1)$ by Assumption 1. Applying Lemma 3 to $\varepsilon_2 = \varepsilon_1/n_{\varepsilon_1} = \varepsilon_1^{1+\alpha}/C$, we get the following result. At time

$$\tau'_{\varepsilon_2} := \big(1 + \lceil \log_2(1/\varepsilon_2)\rceil\big)\mathcal{N}'(f, \varepsilon_2)$$

all the boxes maintained by Algorithm 2 have an area bounded from above by ε_2, i.e.,

$$a_k^{\tau'_{\varepsilon_2}} \leqslant \varepsilon_2, \qquad \forall k = 1, \dots, \tau'_{\varepsilon_2}\,.$$

From Lemma 8 since $n_{\varepsilon_1} \geqslant \mathcal{N}(f, \varepsilon_1)$ by Assumption 1,

$$\mathcal{N}'(f, \varepsilon_2) = \mathcal{N}'\Big(f, \frac{\varepsilon_1}{n_{\varepsilon_1}}\Big) \leqslant 2n_{\varepsilon_1} = 2C\varepsilon_1^{-\alpha}\,,$$

which substituted into the definition of τ'_{ε_2} yields

$$\tau'_{\varepsilon_2} \leqslant 2C\big(1 + \lceil \log_2(1/\varepsilon_2)\rceil\big)\varepsilon_1^{-\alpha}\,.$$

Thus, the certificate up to time τ'_{ε_2} is bounded from above as

$$\xi_{\tau_{\varepsilon_2}} = \frac{1}{2}\sum_{k=1}^{\tau'_{\varepsilon_2}}(a_k^{\tau'_{\varepsilon_2}})^2 \leqslant \frac{\tau'_{\varepsilon_2}(\varepsilon_2)^2}{2} \leqslant \frac{1}{C}\big(1 + \lceil \log_2(1/\varepsilon_2)\rceil\big)\varepsilon_1^{2+\alpha}\,.$$

Now, to get rid of the multiplicative term, similarly to Theorem 2, we can apply Lemma 4 (which also works for StochasticGreedyBox) to replace $\xi_{\tau'_{\varepsilon_2}}$ with ξ_s for $s \geqslant \tau'_{\varepsilon_2}$. The choice $s = \tau'_{\varepsilon_2}\big(\frac{1}{C}\big(1 + \lceil \log_2(1/\varepsilon_2)\rceil\big)\big)^{1/2}$ yields

$$\xi_s \leqslant \varepsilon_1^{2+\alpha}\,.$$

Then, choosing $\varepsilon_1 = \varepsilon^{\frac{2}{2+\alpha}}$ implies $\xi_s \leqslant \varepsilon^2$. Therefore by definition of the stopping criterion

$$\tau_\varepsilon \leqslant s = \Big(\frac{1}{C}\big(1 + \lceil \log_2(1/\varepsilon_2)\rceil\big)\Big)^{3/2} \varepsilon_1^{-\alpha} = \widetilde{\mathcal{O}}\big(\varepsilon^{-\frac{1}{1/\alpha+1/2}}\big)\,.$$

This concludes the proof. □

B.5 Proof of Theorem 5

Theorem 5. *Let $\alpha > 0$ and $\varepsilon \in (0,1]$. Let $f : [0,1] \to [0,1]$ be a non-decreasing and piecewise-C^2 function with a number of C^1-singularities bounded by $\varepsilon^{-\alpha}$ and such that $|f''(x)| \leqslant 1$ whenever it is defined. Then, there exists*

$$t_\varepsilon = \begin{cases} \widetilde{\mathcal{O}}\Big(\varepsilon^{-1+\frac{1}{2p+2}}\Big) & \text{if } \alpha \leqslant \frac{1}{2} \\ \widetilde{\mathcal{O}}\Big(\varepsilon^{-1+\left(\frac{1-\alpha}{1+p}\right)_+}\Big) & \text{if } \alpha \geqslant \frac{1}{2} \end{cases}$$

such that $\big\|\widehat{f}_t - f\big\|_p \leqslant \varepsilon$ for all $t \geqslant t_\varepsilon$, where $\widehat{f}_t$ is the approximation of f returned by GreedyBox *after t rounds.*

Proof. Let $c \in (0,1]$, and $\gamma > 0$ be two constants to be fixed later by the analysis and set $\varepsilon' = c\varepsilon^\gamma$.

Step 1. We will now establish an upper bound on the number of evaluations, denoted by τ_ε, required by GreedyBox to ensure that all individual areas are smaller than ε'. From Lemma 3, we know that

$$\tau_\varepsilon \leqslant 2\big(1 + \lceil p \log_2(1/\varepsilon')\rceil\big)\mathcal{N}'_1(f, \varepsilon')\,, \tag{B5}$$

which can be further bounded from above by the use of Lemma 8 in the appendix with the choice $n := c^{-p}\big\lceil \varepsilon^{-\frac{\gamma p}{p+1}}\big\rceil$. Indeed, since $c \leqslant 1$ and by Lemma 1, $n \geqslant \big\lceil \varepsilon^{-\frac{\gamma p}{p+1}}\big\rceil \geqslant \mathcal{N}_p(f, \varepsilon^{\frac{\gamma p}{p+1}})$. Thus, Lemma 8 entails

$$\mathcal{N}'_1(f, \varepsilon') = \mathcal{N}'_1(f, c\varepsilon^\gamma) = \mathcal{N}'_1\Big(f, \frac{c\varepsilon^{\frac{\gamma p}{p+1}}}{\varepsilon^{-\frac{\gamma}{p+1}}}\Big) \leqslant \mathcal{N}'_1\Big(f, \frac{\varepsilon^{\frac{\gamma p}{p+1}}}{n^{1/p}}\Big) \leqslant 2n = 2c^{-p}\big\lceil \varepsilon^{-\frac{\gamma p}{p+1}}\big\rceil\,,$$

Therefore, plugging back into Inequality (B5), the number of required evaluations is bounded as

$$\tau_\varepsilon \leqslant 4\big(1 + \lceil p \log_2(1/\varepsilon')\rceil\big)c^{-p}\big\lceil \varepsilon^{-\frac{\gamma p}{p+1}}\big\rceil = \widetilde{\mathcal{O}}(c^{-p}\varepsilon^{-\frac{\gamma p}{p+1}})\,. \tag{B6}$$

Step 2. We will now fix the values of $c \in (0,1]$ and $\gamma > 0$ in a way that ensures an approximation error in the L^p-norm smaller than ε.

For $1 \leqslant i \leqslant t$, let us denote by $B_i = [x_i, x_{i+1}] \times [f(x_i), f(x_{i+1})]$ the i^{th} box maintained by GreedyBox, define $w_i := x_{i+1} - x_i$ to be the width of the ith box and let $\widehat{f}_t$ be the piecewise-linear function returned by GreedyBox after t epochs. By the first step of this proof, for all $1 \leqslant i \leqslant t$,

$$\mathcal{A}_p(B_i) := \Big((f(x_{i+1}) - f(x_i))^p (x_{i+1} - x_i) \Big)^{1/p} \leqslant \varepsilon' \,,$$

which implies by construction of $\widehat{f}_t$ (see Proof of Lemma 2), for all $1 \leqslant i \leqslant t$

$$\int_{x_i}^{x_{i+1}} \left| \widehat{f}_t(x) - f(x) \right|^p dx \leqslant (\varepsilon')^p \,. \tag{B7}$$

The remaining part of the proof revolves around using the above upper bound for the non-smooth pieces and a more refined upper bound for the C^1 pieces. Denote by $\mathcal{J} \subseteq \{1, \dots, t\}$ the indices of all boxes such that f is C^1 over $[x_i, x_{i+1}]$. Because f is piecewise-C^1 with at most K pieces, $\mathrm{Card}(\mathcal{J}^c) \leqslant K$. Therefore, the L^p error after t evaluations may be decomposed as

$$\begin{aligned} \left\| \widehat{f}_t - f \right\|_p^p = \int_0^1 \left| \widehat{f}_t(x) - f(x) \right|^p dx &= \sum_{i=1}^{t} \int_{x_i}^{x_{i+1}} \left| \widehat{f}_t(x) - f(x) \right|^p dx \\ &\overset{(B7)}{\leqslant} \min\{K, t\} (\varepsilon')^p + \sum_{i \in \mathcal{J}} \int_{x_i}^{x_{i+1}} \left| \widehat{f}_t(x) - f(x) \right|^p dx \end{aligned} \tag{B8}$$

We are now left with bounding from above the L^p errors on the right-hand-side for all $i \in \mathcal{J}$. On one side, the bound (B7) is valid for all $i \in \mathcal{J}$, which we bound further by $(\varepsilon')^p$. On the other side, we can use Lemma 11 that bounds the L^p approximation error of $\widehat{f}_t$ for any C^1 and piecewise-C^2 function to obtain

$$\int_{x_i}^{x_{i+1}} \left| \widehat{f}_t(x) - f(x) \right|^p dx \leqslant M^p w_i^{2p+1} \,, \tag{B9}$$

where $M \geqslant \frac{3}{2} \sup_{x \notin \mathcal{X}_2} |f''(x)|$, where $\mathcal{X}_2$ denotes the set of C^2-singularities. This prompt us to introduce the following function ϕ, that depends on the width of the intervals of $\mathcal{J}$:

$$\phi\left((w_i)_{i \in \mathcal{J}}\right) = \sum_{i \in \mathcal{J}} \min\left\{ (\varepsilon')^p, M^p w_i^{2p+1} \right\} \,.$$

From (B8), the total error is thus bounded from above by

$$\left\| \widehat{f}_t - f \right\|_p^p \leqslant \min\{K, t\} (\varepsilon')^p + \phi\left((w_i)_{i \in \mathcal{J}}\right) \,. \tag{B10}$$

It now remains to bound the function ϕ for any set $(w_i)_{i\in\mathcal{J}}$ of possible widths that could arise from GreedyBox. We thus need to solve the maximization problem

$$\max_{(w_i)_{i\in\mathcal{J}}} \phi((w_i)_{i\in\mathcal{J}}) \text{ such that } \sum_{i\in\mathcal{J}} w_i \leqslant 1 \,.$$

ϕ may be re-written in two terms:

$$\phi((w_i)_{i\in\mathcal{J}}) = \sum_{i\in\mathcal{J},\varepsilon'\leqslant M w_i^{2+\frac{1}{p}}} (\varepsilon')^p + \sum_{i\in\mathcal{J},\varepsilon'>M w_i^{2+\frac{1}{p}}} M^p w_i^{2p+1} \,. \tag{B11}$$

Let us first handle the first term. Since $\sum_{i\in\mathcal{J}} w_i \leqslant 1$, we have in particular

$$\left|\left\{i\in\mathcal{J}\colon \varepsilon' \leqslant M w_i^{2+\frac{1}{p}}\right\}\right| \cdot \left(\frac{\varepsilon'}{M}\right)^{\frac{p}{2p+1}} \leqslant \sum_{i\in\mathcal{J},\varepsilon'\leqslant M w_i^{2+\frac{1}{p}}} w_i \leqslant 1 \,.$$

This shows that the number of intervals in the first term verifies

$$\left|\left\{i\in\mathcal{J}\colon \varepsilon' \leqslant M w_i^{2+\frac{1}{p}}\right\}\right| \leqslant M^{\frac{p}{2p+1}} (\varepsilon')^{-\frac{p}{2p+1}} \,, \tag{B12}$$

which implies that the first term of (B11) is bounded as

$$\sum_{i\in\mathcal{J},\varepsilon'\leqslant M w_i^{2+\frac{1}{p}}} (\varepsilon')^p \leqslant M^{\frac{p}{2p+1}} (\varepsilon')^{\frac{2p^2}{2p+1}} \,. \tag{B13}$$

Now, let us bound from above the second term of the sum. Since $x \mapsto x^{2p+1}$ is strictly convex on $[0,1]$, Lemma 10 (in the appendix) applied with $\alpha = (\varepsilon'/M)^{p/(2p+1)}$ and $\beta \leqslant \sum_{i\in\mathcal{J}} w_i \leqslant 1$ states that the second term of (B11) is bounded in the following way:

$$M^p \sum_{i\in\mathcal{J},\varepsilon'>M w_i^{2+\frac{1}{p}}} w_i^{2p+1} \leqslant 2M^p\alpha^{2p} = 2M^p \left(\frac{\varepsilon'}{M}\right)^{\frac{2p^2}{2p+1}} = 2M^{\frac{p}{2p+1}} (\varepsilon')^{\frac{2p^2}{2p+1}} \,. \tag{B14}$$

Substituting the two upper bounds (B13) and (B14) into (B11), we have

$$\phi((w_i)_{i\in\mathcal{J}}) \leqslant 3M^{\frac{p}{2p+1}} (\varepsilon')^{\frac{2p^2}{2p+1}} \,,$$

which substituted into (B10) yields

$$\left\|\widehat{f}_t - f\right\|_p^p \leqslant \min\{K,t\}(\varepsilon')^p + 3M^{\frac{p}{2p+1}} (\varepsilon')^{\frac{2p^2}{2p+1}}$$

$$\leqslant \min\{\varepsilon^{-\alpha}, t\} c^p \varepsilon^{\gamma p} + 3M^{\frac{p}{2p+1}} c^{\frac{2p^2}{2p+1}} \varepsilon^{\frac{2\gamma p^2}{2p+1}} . \tag{B15}$$

Now, we finalize the proof by considering three cases depending on the value of α and by optimizing γ and c for each situation.

- *Case 1:* $0 \leqslant \alpha \leqslant 1/2$. Then, $\varepsilon^{-\alpha} \leqslant \varepsilon^{-1/2}$, which substituted in (B15), implies

$$\|\widehat{f}_t - f\|_p^p \leqslant c^p \varepsilon^{\gamma p - 1/2} + 3M^{\frac{p}{2p+1}} c^{\frac{2p^2}{2p+1}} \varepsilon^{\frac{2\gamma p^2}{2p+1}} .$$

Choosing $\gamma = 1 + \frac{1}{2p}$ yields

$$\|\widehat{f}_t - f\|_p^p \leqslant \Big(c^p + 3M^{\frac{p}{2p+1}} c^{\frac{2p^2}{2p+1}}\Big)\varepsilon^p .$$

The choice $c = \min\{2^{-1/p}, (6M)^{-\frac{1}{2p}}\}$ implies for $t \geqslant \tau_\varepsilon$

$$\|\widehat{f}_t - f\|_p \leqslant \varepsilon ,$$

and by (B6), the number of required evaluations is of order

$$\tau_\varepsilon = \widetilde{\mathcal{O}}\Big(c^{-p} \varepsilon^{-\frac{\gamma p}{p+1}}\Big) = \widetilde{\mathcal{O}}\Big(\varepsilon^{-1+\frac{1}{2p+2}}\Big) ,$$

which concludes the first statement of the theorem.

- *Case 2:* $\alpha \geqslant 1$. Then, one may then use Theorem 2, which does note use the piecewise-regularity assumption of f, and implies that $\|\widehat{f}_t - f\|_p \leqslant \varepsilon$ for $t \geqslant \tau_\varepsilon$ with

$$\tau_\varepsilon = \mathcal{O}\big(\log(1/\varepsilon)^2 \mathcal{N}_p(f, \varepsilon)\big) = \mathcal{O}\big(\log(1/\varepsilon)^2 \varepsilon^{-1}\big) = \widetilde{\mathcal{O}}\Big(\varepsilon^{-1+\left(\frac{1-\alpha}{1+p}\right)_+}\Big) .$$

- *Case 3:* $1/2 \leqslant \alpha \leqslant 1$. Then, (B15) yields

$$\|\widehat{f}_t - f\|_p^p \leqslant c^p \varepsilon^{\gamma p - \alpha} + 3M^{\frac{p}{2p+1}} c^{\frac{2p^2}{2p+1}} \varepsilon^{\frac{2\gamma p^2}{2p+1}} .$$

Substituting the choice $\gamma = 1 + \frac{\alpha}{p}$ into (B15) further entails

$$\|\widehat{f}_t - f\|_p^p \leqslant c^p \varepsilon^p + 3M^{\frac{p}{2p+1}} c^{\frac{2p^2}{2p+1}} \varepsilon^{\left(\frac{2p+2\alpha}{2p+1}\right)p} \leqslant \Big(c^p + 3M^{\frac{p}{2p+1}} c^{\frac{2p^2}{2p+1}}\Big)\varepsilon^p .$$

Similary to the first case, choosing $c = \min\{2^{-1/p}, (6M)^{-\frac{1}{2p}}\}$ implies

$$\|\widehat{f}_t - f\|_p \leqslant \varepsilon ,$$

and for any $t \geqslant \tau_\varepsilon$ of order

$$\tau_\varepsilon = \widetilde{\mathcal{O}}\Big(c^{-p}\varepsilon^{-\frac{\gamma p}{p+1}}\Big) = \widetilde{\mathcal{O}}\Big(\varepsilon^{-1+\frac{1-\alpha}{1+p}}\Big)\,.$$

This concludes the proof. □

B.6 Proof of Proposition 6

Proposition 6. *Let $p \geqslant 1$, $\varepsilon \in (0,1)$ and $\alpha > 0$. Then, for any deterministic adaptive algorithm $\mathcal{A}$ and for any*

$$t < (2\varepsilon)^{-1+\left(\frac{1-\alpha}{1+p}\right)_+} - 1\,,$$

there exists a non-decreasing piecewise-affine function $f : [0,1] \to [0,1]$ with at most $\max\{2, \lceil \varepsilon^{-\alpha} \rceil\}$ discontinuities, such that $\|f - \widehat{f}_t\|_p > \varepsilon$.

Proof. Let $p \geqslant 1$, $\varepsilon > 0$ and $\alpha > 0$. Let $\mathcal{A}$ be an adaptive algorithm and fix $t \geqslant 1$ a number of evaluations. We aim to design a function f with at most $K = \lceil \varepsilon^{-\alpha} \rceil$ discontinuities such that $\|f - \widehat{f}_t\|_p > \varepsilon$ if t is sufficiently small.

Let $g : x \to x$ be the identity function on $[0,1]$. Let $x_0 = 0$ and $x_{t+1} = 1$ and denote by $0 \leqslant x_1 \leqslant \ldots \leqslant x_t \leqslant 1$ the points that the algorithm would have chosen after t evaluations, if it was applied on g and by $\widehat{f}_t$ its estimation. Then, we define two functions g_- and g_+ such that

$$\|g_- - g_+\|_p^p \geqslant \min\{Kt^{-1+p}, t^{-p}\}.$$

Let $K_- = \min\{K, t+1\}$. For $i \in \{1, \ldots, t+1\}$, we define $w_i = x_i - x_{i-1}$ the width of the i-th interval. Let $\mathcal{J}$ denotes the set of indexes that correspond to the K_- largest intervals (i.e., such that $w_i \geqslant w_j$ for all $i \in \mathcal{J}$ and $j \notin \mathcal{J}$ and $|\mathcal{J}| = K_-$). Then, we define for all $x \in [0,1]$

$$g_-(x) = \begin{cases} x & \text{if } \exists i \notin \mathcal{J}, \text{ such that } x \in [x_{i-1}, x_i] \\ x_{i-1} & \text{if } x \in [x_{i-1}, x_i) \text{ for } i \in \mathcal{J} \end{cases}$$

and

$$g_+(x) = \begin{cases} x & \text{if } \exists i \notin \mathcal{J}, \text{ such that } x \in [x_{i-1}, x_i] \\ x_i & \text{if } x \in (x_{i-1}, x_i] \text{ for } i \in \mathcal{J} \end{cases}.$$

Then, g_- and g_+ have at most K discontinuities and are such that $g_-(x_i) = g_+(x_i) = g(x_i) = x_i$ for all $i \in \{1, \ldots, t\}$. Thus, since $\mathcal{A}$ is deterministic, the function estimation and the points chosen by $\mathcal{A}$ on g_- and g_+ after t evaluations would also respectively be $\widehat{f}_t$ and $x_1, \ldots, x_t$.

Furthermore, we have

$$\big\|g_- - g_+\big\|_p^p = \int_0^1 \big|g_-(x) - g_+(x)\big|^p dx$$

$$
\begin{aligned}
&= \sum_{i\in\mathcal{J}} \int_{x_{i-1}}^{x_i} \left|g_-(x) - g_+(x)\right|^p dx \\
&= \sum_{i\in\mathcal{J}} w_i^{p+1} \\
&\geqslant K_-\Big(\frac{1}{K_-}\sum_{i\in\mathcal{J}} w_i\Big)^{p+1} && \leftarrow \quad \text{by Jensen's Inequality} \\
&\geqslant K_-\Big(\frac{1}{t+1}\sum_{i=1}^{t+1} w_i\Big)^{p+1} && \leftarrow \quad \text{by Definition of } \mathcal{J} \\
&= K_-(t+1)^{-(p+1)} && \leftarrow \quad \text{because } \sum_{i=1}^{t+1} w_i = 1\,.
\end{aligned}
$$

By triangular inequality, this yields

$$
\begin{aligned}
\max_{f\in\{g_-,g_+\}} \left\|\widehat{f}_t - f\right\|_p &\geqslant \frac{1}{2}\Big(\left\|\widehat{f}_t - g_-\right\|_p + \left\|\widehat{f}_t - g_+\right\|_p\Big) \\
&\geqslant \frac{1}{2}\left\|g_- - g_+\right\|_p \\
&\geqslant \frac{1}{2}K_-^{\frac{1}{p}}(t+1)^{-\frac{p+1}{p}}\,.
\end{aligned}
$$

Therefore, $\max_{f\in\{g_-,g_+\}} \left\|\widehat{f}_t - f\right\|_p > \varepsilon$ if

$$
\frac{1}{2}K_-^{\frac{1}{p}}(t+1)^{-\frac{p+1}{p}} > \varepsilon
$$

which, using $K_- \geqslant \min\{t+1, \varepsilon^{-\alpha}\}$, is satisfied for

$$
t < \min\left\{(2\varepsilon)^{-1+(\frac{1-\alpha}{1+p})}, (2\varepsilon)^{-1}\right\} - 1\,.
$$

Noting that g_- and g_+ have at most $K \leqslant \lceil\varepsilon^{-\alpha}\rceil$ discontinuities and that $g''(x) = 0$ elsewhere concludes the proof. □

B.7 Proof of Proposition 7

Proposition 7. *Let $\varepsilon \in (0, 1/12)$. Then, there exists piecewise-C^2 function f_ε with one C^1-singularity, such that there exists $t \geqslant 2^{-7}\varepsilon^{-3/4}$ with $\left\|\widehat{f}_t - f_\varepsilon\right\|_1 > \varepsilon$ where $\widehat{f}_t$ is the* GreedyBox *approximation at t evaluations.*

Proof. Let $k \in \mathbb{N}$ that will be chosen later by the analysis. Set $s = 2^{3k}$, $s' = 2^{2k}$ and $t = \frac{s+s'}{2}$.

Step 1. Design of a worst-case function We will design a worst-case function f^t that will cause GreedyBox to incur a large L^1-error after t iterations. The function will

consist of two components: one for $x \leqslant 1/2$ that oscillates with a second derivative $|(f^t)''(x)| = 1$, and another that is linear for $x \geqslant 1/2$. f^t has a continuous derivative and its second derivative is piecewise-continuous wit $K = \frac{s'}{2}$ singularities. An illustration is given in Figure 2 for $k = 2$ ($t = 40$).

Let us first design the first oscillating part, that we call $g_{s'}$, and is defined recursively for all $x \in [0, 1]$ by

$$g_{s'}(x) = \begin{cases} x^2 & \text{if } x \in \left[0, \frac{1}{s'}\right] \\ -\left(x - \frac{1}{s'}\right)^2 + \frac{2}{s'} \times \left(x - \frac{1}{s'}\right) + \left(\frac{1}{s'}\right)^2 & \text{if } x \in \left[\frac{1}{s'}, \frac{2}{s'}\right] \\ g_{s'}\left(x - \frac{2i}{s'}\right) + 2i\left(\frac{1}{s'}\right)^2 & \text{if } x \in \left[\frac{2i}{s'}, \frac{2(i+1)}{s'}\right], i \in \left\{1, \dots, \frac{s'}{2} - 1\right\} \end{cases}. \tag{B16}$$

Then, we define f^t by: for all $x \in [0, 1]$

$$f^t(x) = \mathbb{1}_{\{x \leqslant 1/2\}} g_{s'}(x) + \mathbb{1}_{\{x > 1/2\}} \left(x - \frac{1}{2} + \frac{1}{2s'}\right). \tag{B17}$$

Step 2. Expression of the approximation $\widehat{f}_t$ provided by GreedyBox after t iterations on f^t. Let us understand how GreedyBox behaves during the first t iterations when given this function. Once done, we will be able to retrieve the expression of the approximation $\widehat{f}_t$ of f^t made by GreedyBox.

Remark that for $i \in \{0, \dots, s'/2 - 1\}$, (during the first oscillating part), the area of the box $B_i = \left[\frac{i}{s'}, \frac{i+1}{s'}\right] \times \left[f^t\left(\frac{i}{s'}\right), f^t\left(\frac{i+1}{s'}\right)\right]$ is

$$\mathcal{A}_p(B_i) = \frac{1}{s'} \times \frac{1}{s'^2} = \frac{1}{s'^3} = \frac{1}{s^2}.$$

Similarly, for $j \in \left\{\frac{s}{2}, \dots, s - 1\right\}$, (during the linear part), the area of the box $B_j = \left[\frac{j}{s}, \frac{j+1}{s}\right] \times \left[f^t\left(\frac{j}{s}\right), f^t\left(\frac{j+1}{s}\right)\right]$ is

$$\mathcal{A}_p(B_j) = \frac{1}{s^2}.$$

Thus, because GreedyBox tends to equalize the areas of the different boxes, and because it maintains areas of the form $[i/2^j, (i+1)/2^j]$ for some j in $\mathbb{N}^*$ and $i \in \{0, \dots, 2^j - 1\}$ (it can only splits intervals in 2), the box cover we just described is a potential output of GreedyBox after $t = (s' + s)/2$ epochs.

Let us formally prove that it is precisely the case. For any $0 \leqslant j' \leqslant 3k$ and $i \in \{1, \dots, 2^{j'-1} - 1\}$, the area of the box $[i2^{-j'}, (i+1)2^{-j'}] \times [\widehat{f}_t(i2^{-j'}) + \widehat{f}_t((i+1)2^{-j'})]$ is $1/8^{j'}$. Likewise, for any $0 \leqslant j \leqslant 2k$ and $i \in \{2^{j-2}, \dots 2^{j-1} - 1\}$, the area of the box $[i2^{-j}, (i+1)2^{-j}] \times [\widehat{f}_t(i2^{-j}) + \widehat{f}_t((i+1)2^{-j})]$ is $1/4^j$. Thus, if for some epoch $t' \leqslant t$, one box on the left of $1/2$ has a width greater than $3k$, its area will be greater than or equal to $8/s^2$, and if one box on the right of $1/2$ has a width greater than $2k$, its area will be greater than or equal to $4/s^2$. In both cases, the area is strictly greater than $1/s^2$, and GreedyBox would choose at Line 1 to split this box rather than any other box that already has area $1/s^2$.

Splitting one by one boxes with area greater than $1/s^2$, GreedyBox obtains at time t the t boxes with exact area $1/s^2$ described previously. $\widehat{f}_t$ is the linear interpolation between all the points $(x_i, f^t(x_i))$ where x_i is the extremity of one of the box.

Step 3. Error of GreedyBox on f^t after t iterations. Now that we exhibited the exact value of $\widehat{f}_t$, we are left with computing the total error made by GreedyBox on f^t after t steps. Since f^t is linear on the interval $[1/2, 1]$, $\widehat{f}_t$ equals f^t on this segment. Then,

$$\begin{aligned}\left\|f^t - \widehat{f}_t\right\|_1 &= \int_0^{\frac{1}{2}} \left|f^t(x) - \widehat{f}_t(x)\right| dx = \sum_{i=1}^{s'/2-1} \int_{\frac{i}{s'}}^{\frac{i+1}{s'}} \left|f^t(x) - \widehat{f}_t(x)\right| dx \\ &= \frac{s'}{2} \int_0^{\frac{1}{s'}} \left|f^t(x) - \widehat{f}_t(x)\right| dx = \frac{s'}{2} \int_0^{\frac{1}{s'}} \left|x^2 - \frac{x}{s'}\right| dx \\ &= \frac{s'}{2}\left(\frac{1}{2{s'}^3} - \frac{1}{3{s'}^3}\right) = \frac{1}{12(s')^2}\,.\end{aligned}$$

Step 4. Choice of k. We are left with choosing k as large as possible such that the above error is at least ε. That is,

$$\left\|f^t - \widehat{f}_t\right\|_1 > \varepsilon$$

which can be rewritten as

$$\frac{2^{-4k}}{12} = \frac{1}{12{s'}^2} > \varepsilon \qquad \Leftrightarrow \qquad k < \frac{1}{4}\log_2\left(\frac{1}{12\varepsilon}\right).$$

Thus choosing $k = \left\lfloor \frac{1}{4}\log_2\left(\frac{1}{12\varepsilon}\right)\right\rfloor$, yields

$$t = \frac{2^{2k} + 2^{3k}}{2} \geqslant 2^{3k-1} \geqslant \left(12\varepsilon\right)^{-\frac{3}{4}} 2^{-4} > 2^{-7}\varepsilon^{-\frac{3}{4}}\,.$$

Note that the designed function has a number of C^2-singularities

$$K = \frac{s'}{2} = 2^{2k-1} \leqslant \frac{1}{2}(12\varepsilon)^{-1/2} \leqslant \varepsilon^{-1/2}\,,$$

but only has one C^1-singularity at $x = 1/2$. This concludes the proof. □

B.8 Technical lemmas for regular functions

In this section, we establish two technical lemmas that are used in our analysis of GreedyBox for piecewise-C^2 functions. The first lemma, presented below, is a property of strictly convex functions.

Lemma 10. *Let $n \in \mathbb{N}^*$, $\beta \in \mathbb{R}^+$, $\alpha \in [\beta/n, \beta]$ and f be a strictly convex function on $[0, \beta]$. Then*

$$\max\left\{\sum_{i=1}^{n} f(x_i) \colon \forall i, x_i \leqslant \alpha, \sum_{i=1}^{n} x_i = \beta\right\} = mf(\alpha) + f(\beta - m\alpha) + (n - m + 1)f(0)\ ,$$

and the maximum is reached for $x_1^\star = \ldots = x_m^\star = \alpha$, $x_{m+1}^\star = \beta - m\alpha$, $x_{m+2}^\star = \ldots = x_n^\star = 0$, where $m = \lfloor \beta/\alpha \rfloor$.

Proof. Let $x^\star \in \mathbb{R}^n$ be as in the statement of the theorem, and let us show that it is optimal. Let x be a real-number. Since only the sum of the $f(x_i)$ matters, we can assume without loss of generality that the x_i's are sorted in decreasing order. Now, assume that $x_m < \alpha = x_m^\star$, and let us show that $\sum_{i=1}^n f(x_i)$ is not a maximum.

Because the sum of the x_i's still needs to be equal to β, either $x_{m+1} > x_{m+1}^\star = \beta - m\alpha$, or $x_{m+2} > x_{m+2}^\star = 0$. Assume first that $x_{m+2} > x_{m+2}^\star$. Furthermore, assume that $\min\{x_m^\star - x_m, x_{m+2} - x_{m+2}^\star\} = x_m^\star - x_m$. This means that x_m is closer to $x_m^\star$ than x_{m+2} is from $x_{m+2}^\star = 0$. Let $x = x_m^\star$, $\lambda = \frac{x_m^\star - x_{m+2}}{2x_m^\star - x_m - x_{m+2}} \in (0, 1)$ and $y = x_{m+2} + x_m - x_m^\star \in [0, x_{m+2})$. Then,

$$\begin{aligned} f(x_m) + f(x_{m+2}) &= f(\lambda x + (1-\lambda)y) + f((1-\lambda)x + \lambda y) \\ &< \lambda f(x) + (1-\lambda)f(y) + (1-\lambda)f(x) + \lambda f(y) \\ &\leqslant f(x_m^\star) + f(y)\ . \end{aligned}$$

Then for $x_m' = x$, $x_{m+2}' = y$ and $x_i' = x_i$ for $i \neq \{m, m+2\}$, $\sum_{i=1}^n f(x_i') < \sum_{i=1}^n f(x_i)$, which shows that $(x_i)_{1 \leqslant i \leqslant n}$ is suboptimal. We can do the same kind of construction when $x_{m+1} > x_{m+1}^\star$ or when $\min\{x_m^\star - x_m, x_{m+2} - x_{m+2}^\star\} = x_{m+2} - x_{m+2}^\star$. All these different cases show that $\sum_{i=1}^n f(x_i)$ is maximized only when the largest possible amount of the x_i's are at the extremity of the constraint set, that is when $x_i = 0$ or $x_i = \alpha$. This concludes the proof. □

The second technical Lemma below recalls a classical result and provides an upper bound on the L^p-error achieved by an affine approximation of a C^2 function. In particular, this lemma implies a sample complexity of order $\mathcal{O}(\varepsilon^{-\frac{1}{2}})$ for the trapezoidal rule to provide an ε-approximation in L^p norm for C^2 functions.

Lemma 11. *Let $a < b$. Assume that $f : [a, b] \to [0, 1]$ is C^1 and piecewise-C^2 and such that $|f^{(2)}(x)| \leqslant M$ for all x where it is defined. Then,*

$$\int_a^b \left|\widehat{f}(x) - f(x)\right|^p dx \leqslant \left(\frac{3M}{2}\right)^p (b-a)^{2p+1}\ ,$$

where $\widehat{f}$ is the affine approximation defined for all $x \in [a, b]$ by:

$$\widehat{f}(x) = f(a) + \big(f(b) - f(a)\big)\frac{x-a}{b-a}\ .$$

Proof. Since f is piecewise-C^2 with bounded second derivative, f' is absolutely continuous, we can thus apply the Taylor-Lagrange Theorem with the integral form of the remainder with $k=1$. We have for all $x, y \in [a,b]$

$$\left|f(y) - f(x) - f'(x)(y-x)\right| = \left|\int_x^y f''(u)(y-u)du\right| \leqslant \frac{M}{2}(y-x)^2 .$$

We now control the L^p-error of $\widehat{f}$. The above inequality yields that for all $x \in [a,b]$ there exists $r(x)$ such that $|r(x)| \leqslant \frac{M}{2}(x-a)^2$ and

$$f(x) = f(a) + f'(a)(x-a) + r(x) .$$

Thus, applying it with $x = b$ entails

$$\frac{f(b)-f(a)}{b-a} = f'(a) + \frac{r(b)}{b-a} ,$$

which in turns implies that

$$\begin{aligned}
\left|\widehat{f}(x) - f(x)\right| &= \left|f(a) + (x-a)\frac{\big(f(b)-f(a)\big)}{b-a} - f(x)\right| \\
&= \left|(x-a)\frac{r(b)}{b-a} + r(b) - r(x)\right| \\
&\leqslant \frac{M}{2}\big[(b-a)(x-a) + (b-a)^2 + (x-a)^2\big] .
\end{aligned}$$

Therefore, with the change of variable $x = a + (b-a)u$, we get

$$\begin{aligned}
\int_a^b \left|\widehat{f}(x) - f(x)\right|^p dx &\leqslant \Big(\frac{M}{2}\Big)^p \int_a^b \big[(b-a)(x-a) + (b-a)^2 + (x-a)^2\big]^p dx \\
&= \Big(\frac{M}{2}\Big)^p (b-a)^{2p+1} \int_0^1 (1+u+u^2)^p du \\
&\leqslant \Big(\frac{3M}{2}\Big)^p (b-a)^{2p+1} ,
\end{aligned}$$

which concludes the proof. □

Appendix C Numerical experiments: the L^2-norm case

In the core of the paper, we gave some plots on the error of GreedyBox, and GreedyWidthBox as compared to the trapezoidal rule in Section 5. In this section, we complete the comparison by displaying some plots for the L^2-norm. In this case, the area of a box $B = [c^-, c^+] \times [y^-, y^+]$ is worth $(y^+ - y^-)(c^+ - c^-)^2$.

We can notice several interesting facts on the plots of Figure C1. First, the trapezoidal rule, and GreedyWidthBox behaves better than GreedyBox for the square function. This is the first case we could find where the trapezoidal rule has a better speed of convergence than GreedyBox. Another point is that on Figure C1b, the trapezoidal rule has a bound of order $t^{11/20}$. However, in Theorem 2, we proved that GreedyBox has an error at worst of the order of $\mathcal{N}(f, \varepsilon)$, which we proved to be smaller than $\lceil \varepsilon^{-1} \rceil$. Yet, for this example, the trapezoidal does not converge in $\mathcal{O}(t)$, which proves that GreedyBox satisfies better worst case property than the trapezoidal rule for L^p-norm with $p > 1$. Figure C1c shows results similar to the L^1-norm. In, Figure C1d, we consider the L^2 for approximating $g^t : x \mapsto \frac{1}{2} f^{t^{9/10}}(2x) \mathbb{1}_{x \leqslant 1/2} + \mathbb{1}_{x > 1/2}$, where f^s is the function defined in equation (B17) at $s = t^{9/10}$. The empirical rates seem to confirm once again the anticipated worst-case rates as determined by our analysis.

References

[1] Novak, E.: Quadrature formulas for monotone functions. Proceedings of the American Mathematical Society **115**(1), 59–68 (1992)

[2] Davis, P.J., Rabinowitz, P.: Methods of Numerical Integration. Elsevier Science, Netherlands (1984)

[3] Vaart, A.W.: Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (1998)

[4] Sukharev, A.G.: On the existence of optimal affine methods for approximating linear functionals. Journal of Complexity **2**(4), 317–322 (1986)

[5] Bakhvalov, N.S.: On the optimality of linear methods for operator approximation in convex classes of functions. USSR Computational Mathematics and Mathematical Physics **11**(4), 244–249 (1971)

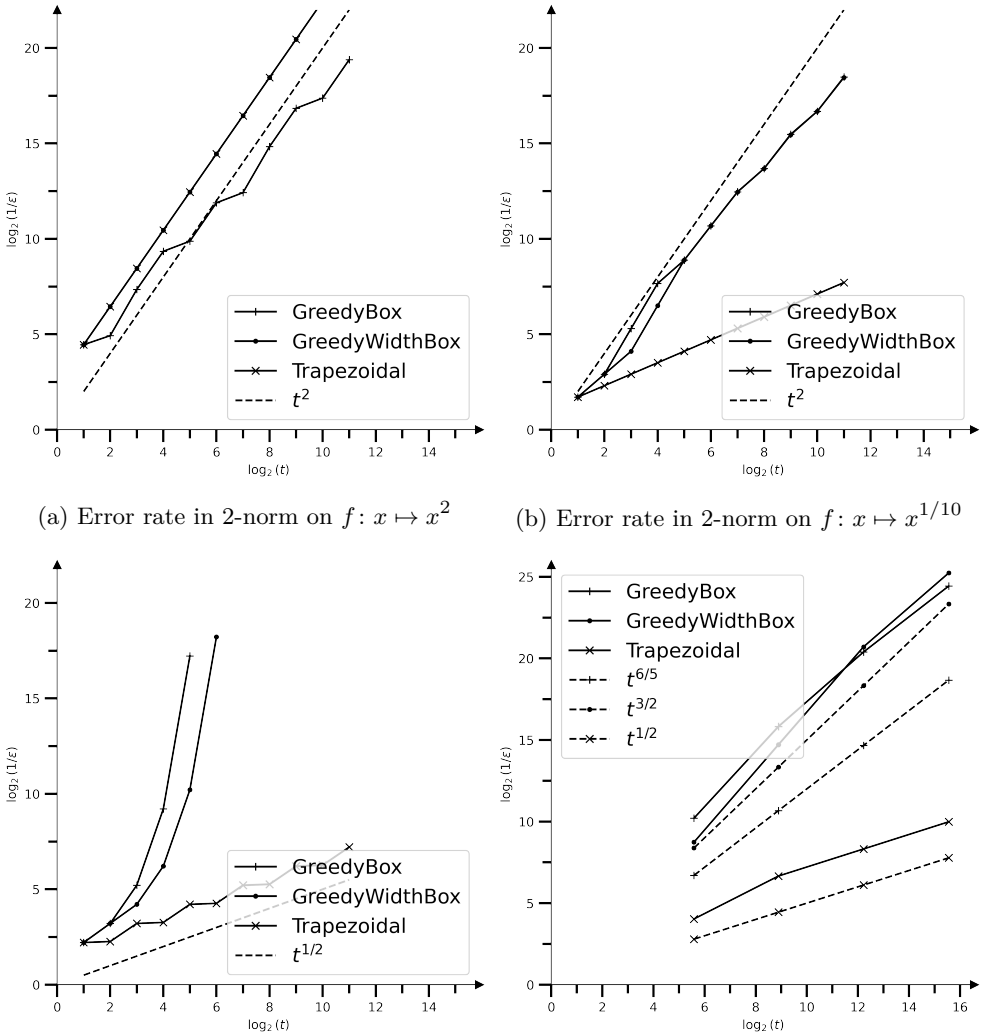
[6] Kiefer, J.: Optimum sequential search and approximation methods under minimum regularity assumptions. Journal of the Society for Industrial and Applied Mathematics **5**(3), 105–136 (1957)

[7] Papageorgiou, A.: Integration of monotone functions of several variables. Journal of Complexity **9**(2), 252–268 (1993)

[8] Brass, H., Petras, K.: Quadrature Theory: the Theory of Numerical Integration on a Compact Interval. Mathematical Surveys and Monographs, vol. 178. American Mathematical Soc., Rhode Island, USA (2011)

[9] Novak, E., Roschmann, I.: Numerical integration of peak functions. Journal of Complexity **12**(4), 358–379 (1996)

[10] Graf, S., Novak, E.: The average error of quadrature formulas for functions of



(a) Error rate in 2-norm on $f\colon x \mapsto x^2$

(b) Error rate in 2-norm on $f\colon x \mapsto x^{1/10}$

(c) Error rate in 2-norm on $f(x) = \mathbb{1}_{\{x \geqslant 0.3\}}$

(d) Error rate in 2-norm on g^t

Fig. C1: Comparison of GreedyBox and GreedyWidthBox with the trapezoidal rule for the 2-norm.

bounded variation. The Rocky Mountain journal of mathematics **20**(3), 707–716 (1990)

[11] Novak, E.: Quadrature formulas for convex classes of functions. In: Numerical Integration IV: Proceedings of the Conference at the Mathematical Research Institute, Oberwolfach, November 8–14, 1992, pp. 283–296 (1993). Springer

[12] Novak, E.: Optimal recovery and n-widths for convex classes of functions. Journal of Approximation Theory **80**(3), 390–408 (1995)

[13] Hinrichs, A., Novak, E., Woźniakowski, H.: The curse of dimensionality for the class of monotone functions and for the class of convex functions. Journal of Approximation Theory **163**(8), 955–965 (2011)

[14] Ritter, K., Wasilkowski, G.W., Woźniakowski, H.: On multivariate integration for stochastic processes. In: Numerical Integration IV: Proceedings of the Conference at the Mathematical Research Institute, Oberwolfach, November 8–14, 1992, pp. 331–347 (1993). Springer

[15] Katscher, C., Novak, E., Petras, K.: Quadrature formulas for multivariate convex functions. Journal of Complexity **12**(1), 5–16 (1996)

[16] Krieg, D., Novak, E.: A universal algorithm for multivariate integration. Foundations of Computational Mathematics **17**, 895–916 (2017)

[17] Chandra, P.: Trigonometric approximation of functions in Lp-norm. Journal of Mathematical Analysis and Applications **275**(1), 13–26 (2002)

[18] Fletcher, R., Grant, J., Hebden, M.: The calculation of linear best Lp approximations. The Computer Journal **14**(3), 276–279 (1971)

[19] Fletcher, R., Grant, J., Hebden, M.: Linear minimax approximation as the limit of best l_p-approximation. SIAM Journal on Numerical Analysis **11**(1), 123–136 (1974)

[20] Plaskota, L., Wasilkowski, G.W.: Adaption allows efficient integration of functions with unknown singularities. Numerische Mathematik **102**, 123–144 (2005)

[21] Plaskota, L., Wasilkowski, G., Zhao, Y.: The power of adaption for approximating functions with singularities. Mathematics of computation **77**(264), 2309–2338 (2008)

[22] Kopotun, K., Shadrin, A.: On k-monotone approximation by free knot splines. SIAM journal on mathematical analysis **34**(4), 901–924 (2003)

[23] Kopotun, K.A., Leviatan, D., Prymak, A.: Nearly monotone and nearly convex approximation by smooth splines in Lp, p striclty greater than 0. Journal of Approximation Theory **160**(1-2), 103–112 (2009)

[24] Novak, E.: On the power of adaption. Journal of Complexity **12**(3), 199–237 (1996)

[25] Lattimore, T., Szepesvári, C.: Bandit Algorithms. Cambridge University Press, Cambridge (2020)

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

[27] Bachoc, F., Cesari, T., Gerchinovitz, S.: Instance-dependent bounds for zeroth-order lipschitz optimization with error certificates. Advances in Neural Information Processing Systems **34**, 24180–24192 (2021)

[28] Bonnet, L., Akian, J.-L., Savin, É., Sullivan, T.J.: Adaptive reconstruction of imperfectly observed monotone functions, with applications to uncertainty quantification. Algorithms **13**(8), 196 (2020)

[29] Fang, B., Guntuboyina, A., Sen, B.: Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and Hardy-Krause variation. The Annals of Statistics **49**(2), 769–792 (2021)

[30] Abernethy, J.D., Amin, K., Zhu, R.: Threshold bandits, with and without censored feedback. Advances in Neural Information Processing Systems **29** (2016)