Active Learning for Convex Regression

Max Simchowitz*, Kevin Jamieson*, Jordan W. Suchow†, Thomas L. Griffiths†
Department of Electrical Engineering and Computer Sciences*,
and Department of Psychology†
University of California, Berkeley, Berkeley, CA 94720
{msimchow,kjamieson,suchow,tom_griffiths}@berkeley.edu

June 29, 2017

Abstract

In this paper, we introduce the first principled adaptive-sampling procedure for learning a convex function in the L_∞ norm, a problem that arises often in economics, psychology, and the social sciences. We present a function-specific measure of complexity and use it to prove that our algorithm is information-theoretically near-optimal in a strong, function-specific sense. We also corroborate our theoretical contributions with extensive numerical experiments, finding that our method substantially outperforms passive, uniform sampling for favorable synthetic and data-derived functions in low-noise settings with large sampling budgets. Our results also suggest an idealized "oracle strategy", which we use to gauge the potential for deploying the adaptive-sampling strategy on any function in any particular setting.

1 Introduction

In convex regression, one seeks to learn a function from noisy observations and exploit the knowledge that it is convex[1]. Many functions describing individual utility in economics, the output of manufacturing processes, and natural phenomena in the social sciences are either convex or concave (i.e., $-f$ is convex). For example, a classic effect in behavioral economics is *temporal discounting*, where people value delayed rewards less than immediate rewards [8, 9]. Discounting rates are measured through a task in which the participant chooses between a smaller, sooner reward and a larger, later reward. The indicator denoting that the smaller sooner reward is chosen over a larger reward at time x is modeled as a Bernoulli random variable $y_i \in \{0, 1\}$, where $\mathbb{E}[y_i|x_i] = \mathbb{P}(y_i = 1|x_i) = f(x_i)$ and f is convex. Convexity is natural because the effect of delay on discounting decreases over time: the prospect of waiting 5 more days is more painful if the initial wait is just 1 day versus 10 days.

Standard experiment designs in the behavioral sciences manipulate a variable (e.g., delay) over a fixed, uniformly spaced grid (e.g., $x_i = \frac{i-1}{n-1}$) of ≈ 5 points, placing trials of an experimental task at each value on this grid [7]. Having collected the data, the experimenter then fits a parametric function of assumed form (e.g., exponential or hyperbolic) using maximum likelihood estimation. This approach has many shortcomings, including the problem of model mismatch, where the true function f lies outside the assumed class of functions, and the problem of inefficient confidence-interval construction, where error bars become increasingly loose as the number of design points n grows.

Non-parametric convex regression (c.f. [6]) apparently corrects for these shortcomings by combining a fixed, uniformly spaced grid of $n \gg 5$ points with methods that construct error bars at any $x \in [0, 1]$ (see [2]). However, like most non-parametric estimators, non-parametric convex regression can require a vast amount of data to recover the underlying function, making this method impractical.

This paper proposes methods that achieve the best of both worlds — non-parametric convex regression that does not require many samples — by collecting the data *adaptively*. Specifically, we consider:
- **Passive Sampling**: At each time t, x_t depends only on $\{x_s\}_{s=1}^{t-1}$.
- **Active Sampling**: At each time t, x_t may depend on $\{(x_s, y_s)\}_{s=1}^{t-1}$.

[1] A differentiable function f is convex if and only if $f'(y) \geq f'(x)$ for all $x < y$.

Examples of passive sampling include i) random sequences where x_s is drawn uniformly at random on $[0, 1]$, ii) quasi-random or other low-discrepancy sequences (e.g. the Sobol sequence), and iii) fixed horizon uniform spacing where some $n \in \mathbb{N}$ is fixed and $x_t = \frac{t-1}{n-1}$ for $t = 1, \ldots, n$.

In general, active sampling has limited benefits when the function class f lives in a relatively unconstrained space (e.g., Holder classes [3]), but as we show in this work, if f is known to be convex and one wishes to find an estimator $\widehat{f}$ that minimizes $||f - \widehat{f}||_\infty := \max_{x \in [0,1]} |f(x) - \widehat{f}(x)|$, adaptive sampling can yield substantial improvements over passive designs.

1.1 Problem Statement and Contributions

Adaptive sampling can be viewed as a game where, at each time t, the player chooses a point $x_t \in [0, 1]$ and nature reveals $f(x_t) + w_t$, where w_t is an independent, zero-mean, σ^2-subgaussian random variable (e.g., $w_t \sim \mathcal{N}(0, \sigma^2)$). For a class of functions $\mathcal{F}$, an algorithm Alg is said to be (ϵ, δ)-*correct* over $\mathcal{F}$ if, for a given ground-truth $f \in \mathcal{F}$, Alg returns an estimator $\widehat{f}$ such that $\mathbb{P}[\|f - \widehat{f}\|_\infty > \epsilon] \leq 1 - \delta$ at some stopping time T (probability is taken over the randomness of the noise and in the algorithm). We assume all functions in $\mathcal{F}$ are convex, which captures the guarantee that the true f is indeed convex. This paper studies how many adaptively chosen samples are required to learn a *particular* ground truth function f, subject to the (ϵ, δ)-correctness guarantee.

Building on [2], we define a local complexity parameter, the "modulus of continuity" (Equation 1), which we leverage to provide a lower bound on the sample complexity of adaptively learning a convex function f in the L_∞-norm. The packing argument for constructing our lower bound immediately suggests a clairvoyant sampling allocation that we call the "oracle strategy." This corresponds to the optimal allocation of samples one would choose to use in hindsight, given knowledge of the true function f and desired accuracy ϵ. We observe that both the oracle allocations and resulting sampling complexities vary considerably for different convex functions f. Moreover, the oracle strategy can be used to gauge the "best-case" performance of any sampling strategy, allowing us to characterize the potential advantages of adaptive sampling in any given application domain.

We propose an active-sampling algorithm for any noise level that efficiently converges to an approximation of the oracle strategy and matches our information theoretic-lower bounds for recovering f in L_∞ up to logarithmic factors. The algorithm relies on a remarkable fact about convex functions: that approximation error of a secant over an interval can be estimated using only a single point. In particular, our results show that to achieve ϵ accuracy, the difference between active and passive sampling can be *polynomial* in $1/\epsilon$ (see Section 2).

Finally, we validate our theoretical claims with an empirical study using both synthetic functions and those derived from real data. We observe that in low noise settings or when the sampling budget is large, active sampling can substantially outperform passive uniform sampling. Moreover, our algorithm constitutes the first theoretically justified algorithm (passive or active) that guarantees uniform accuracy, even at the boundaries of the interval [2, 6]. However, when the noise is high or the sampling budget is modest, the active sampling algorithm does not clearly outperform passive uniform sampling. We are encouraged that comparing the performance of our active algorithm to the oracle sampling strategy suggests room for modest but non-negligible improvements.

1.2 Related work

Examples of learning convex functions can be found across the behavioral sciences. For example, temporal discounting functions as described earlier are often modeled as exponential, $f(x) = Ae^{-Bx}$, or hyperbolic, $f(x) = A(1 + Bx)^{-p}$, for parameters A, B, p [14, 9, 8]. On the theoretical side, there is a rich literature on convex and monotonic function regression which studies estimating the ground truth f in the L_p norm $\|f - \widehat{f}\|_p^p = \int_0^1 |f(x) - \widehat{f}(x)|^p dx$ using on-adaptive, fixed design setting, $x_i = \frac{1}{n-1}$ for $i = 0, \ldots, n-1$ [10]. If $\mathcal{F}$ is the set of Lipschitz, convex functions, then the L_∞ error is known to decrease like $(\log(n)/n)^{1/3}$, whereas if the convex function has Lipschitz gradients, the rate improves to $(\log(n)/n)^{2/5}$ [6]. There is also a large body of work which studies recovering in the L_2-norm, and characterizes both minimax rates and sample complexities for natural subclasses of convex functions (e.g. k-piecewise linear) [11, 4].

Our work draws heavily upon Cai et al. [2] (which in turn builds on [5]) which aims to characterize the function-specific sample complexity of estimating a convex f at a given point in the interior of $[0,1]$, from uniform measurements. We extend these tools to characterize the complexity of estimating f with uniform accuracy over the interval $[0,1]$, from measurements which may be chosen in an adaptive, function-dependent manner. We are thus able to obtain exceptionally granular, instance-specific results similar to those in the multi-arm arm bandit literature [12], and in recent work studying the local minimax sample complexity of convex optimization [15].

Active non-parametric regression has been previously studied in the context of Holder classes, where it was shown that active sampling cannot improve over passive sampling in the L_2 metric in general [3], though it can yield improvements for restricted subclasses, such as those well-approximated by piecewise constant functions [3, 13].

2 Efficiently Learning a Convex Function

We begin by establishing preliminary notation. For an interval $I = [a,b] \subseteq [0,1]$, define the right-, middle- and left-endpoints as $x_{l(I)} = a, x_{m(I)} = \frac{a+b}{2}, x_{r(I)} = b$. We define the secant approximation of f on an interval $I \subset [0,1]$ as $\text{Sec}[f,I](x) = \frac{x_{r(I)}-x}{x_{r(I)}-x_{l(I)}} f(x_{l(I)}) + \frac{x-x_{l(I)}}{x_{r(I)}-x_{l(I)}} f(x_{r(I)})$, and note that for a convex function, this approximation always overestimates f ($\text{Sec}[f,I](x) \geq f(x)$).

Definition 1 *For any convex function f and interval $I \subseteq [0,1]$, let $\Delta(f,I) := \text{Sec}[f,I](x_{m(I)}) - f(x_{m(I)}) = \frac{f(x_{l(I)})+f(x_{r(I)})}{2} - f(x_{m(I)})$ denote the error of the secant approximation to f on I at the midpoint $x_{m(I)}$. In addition, we overload notation so that for any x, t such that $x \in [t, 1-t]$ we have $\Delta(f,x,t) := \Delta(f,[x-t,x+t])$.*

We now state a remarkable fact about convex functions that is at the core of our analysis.

Lemma 2.1 *For any convex f, $\Delta(f,I) \leq \max_{x \in I} \{\text{Sec}[f,I](x) - f(x)\} \leq 2\Delta(f,I)$.*

Lemma 2.1 is a special case of a more general lemma stated in Appendix B that upper bounds the supremum of the secant approximation error by a constant using just a single point within the interval. Convexity is critical to the proof of this lemma and such a property does not hold, for instance, on merely monotonic functions. We remark that the first inequality is trivial, whereas the second inequality is tight in the sense that it is achieved by $f(x) = (1-x)^p$ on interval $I = [0,1]$ as $p \to \infty$.

The above observations motivate our strategy of approximating f with secant approximations on disjoint intervals whose union is $[0,1]$. The next definition relates the secant approximation error to the required sampling density.

Definition 2 (Local Modulus) *Let $f : [0,1] \to \mathbb{R}$ be convex. We define the left- and right-approximation points as*

$$t_{\text{left}}(f,\epsilon) := \inf\{t \leq 1/2 : \Delta(f,t,t) \geq \epsilon\} \quad t_{\text{right}}(f,\epsilon) = \inf\{t \leq 1/2 : \Delta(f,1-t,t) \geq \epsilon\}$$

For any $x \in [t_{\text{left}}(f,\epsilon), 1 - t_{\text{right}}(f,\epsilon)]$, we define the ϵ-approximation modulus of f at a point $x \in [0,1]$ as the least t such that the midpoint secant approximation to f on $[x-t, x+t]$ has bias ϵ:

$$\omega(f,x,\epsilon) := \min\{t \in [0, \min\{x, 1-x\}] : \Delta(f,x,t) \geq \epsilon\}. \tag{1}$$

Intuitively, $\omega(f,x,\epsilon)$ describes on what scale f "looks" linear around some x, up to a tolerance ϵ. As is common in the literature for ℓ_∞ risk bounds [6], we restrict the interval of interest in the definition of the modulus to $[t_{\text{left}}(f,\epsilon), 1 - t_{\text{right}}(\epsilon)]$ since this allows us to avoid some of the pecularities that convex functions may exhibit on the endpoints of a compact interval. Nevertheless, the algorithms in this work guarantee accuracy on the whole interval $[0,1]$. Within $[t_{\text{left}}(f,\epsilon), 1 - t_{\text{right}}(\epsilon)]$, we will show that the midpoint errors $\Delta(f,x,t)$ concisely describe how densely one would need to sample f in the neighborhood of $x \in [0,1]$ in order to estimate it up to a desired accuracy ϵ in the ℓ_∞-norm. At a high level, the central message of this paper is:

*Though the sample complexity of any **passive sampling** convex regression algorithm scales with the **worst-case** sampling density* $\sup_{x\in[t_{\text{left}}(f,\epsilon),1-t_{\text{right}}(f,\epsilon)]}\omega^{-1}(f,x,\epsilon)$*, **active sampling** algorithms can have a sample complexity that scale with the **average** sampling density*

$$\Lambda(f,\epsilon) := \int_{t_{\text{left}}(f,\epsilon)}^{1-t_{\text{right}}(f,\epsilon)} \omega(f,x,\epsilon)^{-1}dx\ . \tag{2}$$

Moreover, no algorithm can estimate f using substantially fewer samples than $\Lambda(f,\epsilon)$.

To provide intuition, if $\sup_{x\in[0,1]} f'''(x) < \infty$ and ϵ is small, then $\omega(f,x,\epsilon)^{-1} \approx \sqrt{\frac{f''(x)}{2\epsilon}}$, which means that the difference between the worst-case and average-case is most stark when the function has areas of high localized curvature, e.g. $f(x) = 1-\sqrt{x}$ or $f(x) = \frac{1}{100}\log(1+\exp(-100(x-\frac{1}{2}))$. Piecewise linear functions, e.g. $f(x) = \max\{1-7x, 0\}$, constitute an extreme case, where average modulus scales like $\log(1/\epsilon)$, yet the largest modulus is $\approx 1/\epsilon$. The average and worst-case moduli are equal when the function has constant curvature, e.g. $f(x) = x^2$.

2.1 Sample Complexity of Approximating a Convex Function

We first consider the passive uniform sampling scheme where $x_i = \frac{i}{n-1}$ for $i = 1,\dots,n-1$ for some value of n. If $x_* := \arg\inf\{\omega(f,x,\epsilon) : x \in [t_{\text{left}}(f,\epsilon), 1-t_{\text{right}}(f,\epsilon)]\}$ and no x_i was contained in $I_* := [x_* - \omega(f,x_*,\epsilon), x_* + \omega(f,x_*,\epsilon)]$, then it would be impossible to disambiguate between $f(x)$ and the function $\tilde{f}_{I^*} := f(x) + \mathbb{I}(x\in I^*)\cdot(\text{Sec}[f,I^*](x) - f(x))$, the function obtained by replacing f with its secant approximation on I^*, with an approximation error less than ϵ. This implies that any (ϵ,δ)-correct uniform sampling procedure requires $n \gtrsim \sup_{x\in[t_{\text{left}}(f,\epsilon),1-t_{\text{right}}(f,\epsilon)]}\omega(f,x,\epsilon)^{-1}$ measurements to ensure that $1/(n-1)$ is smaller than $|I_*| = \inf_{x\in[0,1]} 2\omega(f,x,\epsilon)$.

By the same token, if any (even adaptive) algorithm does not measure f on an interval I for which $\Delta(f,I) \geq \epsilon$, then f cannot distinguish between f and the alternative function $\tilde{f}_I := f(x) + \mathbb{I}(x\in I)(\text{Sec}[f,I](x) - f(x))$. Thus, a key step to showing that $\Lambda(f,\epsilon)$ effectively lower bounds the number of evaluations is to show that it roughly lower bounds the number of intervals I for which $\Delta(f,I) \geq \epsilon$. The following packing lemma is proved in Section C.

Lemma 2.2 (Packing) *Let $f:[0,1]\to\mathbb{R}$ be a convex function, $\epsilon > 0$, and define*

$$\underline{N}(f,\epsilon) = \frac{\Lambda(f,\epsilon)}{4(1+\log(\omega_{\max}(f,\epsilon)/\omega_{\min}(f,\epsilon)))} - 3 \tag{3}$$

where $\omega_{\max}(f,\epsilon) = \max_{x\in[t_{\text{left}}(f,\epsilon),1-t_{\text{right}}(f,\epsilon)]}\omega(f,x,\epsilon)$ and $\omega_{\min}$ is defined analogously. Then, for any $n \leq \underline{N}(f,\epsilon)$ there exists $\{z_i\}_{i=1}^n \subset [0,1]$ such that the intervals $I_i^\epsilon := [z_i - \omega(z_i,f,\epsilon), z_i + \omega(z_i,f,\epsilon)]$ have disjoint interiors, are contained in $[2t_{\text{left}}(f,\epsilon), 1-2t_{\text{right}}(f,\epsilon)]$ and satisfy $\sup_{x\in I_i^\epsilon}\text{Sec}[f,I_i^\epsilon](x) - f(x) \geq \Delta(f,I_i^\epsilon) > \epsilon$ for all i.

Thus, if we define the local class of alternatives functions (all of which are also convex)

$$\mathcal{G}_{f,\epsilon} := \Big\{ f(x) + \sum_{i=1}^n \beta_i \mathbb{I}\{x\in I_i^\epsilon\}(\text{Sec}[f,I_i^\epsilon](x) - f(x)) : \beta\in\{0,1\}^n \Big\}, \tag{4}$$

then $f\in\mathcal{G}_{f,\epsilon}$ and for any $\{x_i\}_{i=1}^{n-1}\subset[0,1]$ there exists a convex function $g\in\mathcal{G}_{f,\epsilon}$ such that $f(x_i) = g(x_i)$ for all i and $\sup_{x\in[0,1]} g(x) - f(x) > \epsilon$. Hence, if a sample is not placed in each interval I_i^ϵ, one cannot disambiguate between the true function f and the $g\in\mathcal{G}_{f,\epsilon}$ that differs by f only over the interval I_i^ϵ. Inflating ϵ by a factor of 2 a we see that if Alg cannot distinguish between f and $g\in\mathcal{G}_{f,2\epsilon}$, then f will incur an error $> \epsilon$ on one of the two instances. Thus,

Theorem 2.1 (Active Lower Bound - Noiseless) *Fix some convex function f and $\epsilon > 0$. Let $\underline{N}(f,\epsilon)$ and $\mathcal{G}_{f,2\epsilon}$ be defined as in Lemma 2.2 and Equation (4). If Alg is (ϵ,δ)-correct over the class $\mathcal{G}_{f,2\epsilon}$, then $\mathbb{E}_{f,\mathsf{Alg}}[T] \geq (1-2\delta)\underline{N}(f,2\epsilon)$.*

The above theorem, whose proof is found in Section D, demonstrates a lower bound on the true number of samples collected when approximating f, provided that Alg is uniformly correct over the class of instances $\mathcal{G}_{f,2\epsilon}$. While the above lower bound holds a fortiori for the class of all convex functions, we emphasize that it applies *even* when we consider these local alternatives.

By inspecting the proof of the packing, we see that we can *cover* the interval $[0,1]$ with $\underline{N}(f,\epsilon)+3$ intervals I such that $\Delta(f,I)=\epsilon$. By our main lemma, this implies that obtaining noiseless measurements of f on the endpoints and midpoints of these intervals is sufficient to certify that the estimator $\widehat{f}$ obtained via the secant approximation to f on each interval, using the measurements, satisfies $\|\widehat{f}-f\|_\infty \le 2\epsilon$. We therefore refer to the consturction in our packing as the *oracle strategy*, since it constitutes the best possible allocation of samples for identifying f up to a constant rescaling of ϵ (adressed below), and a small additive constant. Suprisingly, a close approximation to this oracle strategy can be attained by efficient sampling procedure, detailed in Section 2.3. We state its guarantee as follows:

Theorem 2.2 *Let $f:[0,1]\to\mathbb{R}$ be convex, let* Alg *denote the algorithm of Section 2.3, and fix $\epsilon>0$. Then once* Alg *evaluates f at $n=\overline{N}(f,\epsilon):=\lceil 12\Lambda(f,\epsilon/2)+9\rceil$ locations,* Alg *can certify that the convex function $\widehat{f}_n$ defined in Section 2.3 satisfies $\|f-\widehat{f}_n\|_\infty\le\epsilon$.*

In the above theorem, Alg is able to certify $\|f-\widehat{f}_n\|_\infty\le\epsilon$ by measuring f on the endpoints $x_{l(I)}, x_{r(I)}$ and midpoints $x_{m(I)}$ of a family $\mathcal{I}$ of adaptively chosen intervals , such $\bigcup_{I\in\mathcal{I}} I=[0,1]$, and $\Delta(f,I)\le\epsilon/2$ for all $I\in\mathcal{I}$. With this information, Alg can construct a secant approximation to f on each interval which then satisfies $\sup|\mathrm{Sec}[f,x]-f(x)|\le 2\cdot\epsilon/2=\epsilon$ by Lemma 2.1. The key challenge is designing the algorithm to only require $\lesssim\Lambda(f,\epsilon/2)$ of these intervals.

2.2 Comparing the upper and lower bound

The upper bound of Theorem 2.2 and lower bound of Theorem 2.1, nearly match, with two exceptions: The first is that the upper bound is in terms of $\Lambda(f,\epsilon/2)$, whereas the lower bound is in terms of $\Lambda(f,2\epsilon)$. The two quantities can be related by the following proposition, proved in Appendix H.

Proposition 2.3 *For any $0<c\le 1$, $\epsilon>0$ and any convex f, $\omega(f,x,\epsilon)\ge\omega(f,x,c\epsilon)\ge c\omega(f,x,\epsilon)$ for all $x\in[t_{\mathrm{left}}(f,\epsilon),1-t_{\mathrm{right}}(f,\epsilon)]$. Moreover,*

$$\Lambda(f,\epsilon)+\log\frac{t_{\mathrm{left}}(f,\epsilon)t_{\mathrm{right}}(f,\epsilon)}{t_{\mathrm{left}}(f,c\epsilon)t_{\mathrm{right}}(f,c\epsilon)}\le\Lambda(f,c\epsilon)\le\frac{1}{c}\left\{\Lambda(f,\epsilon)+\log\frac{t_{\mathrm{left}}(f,\epsilon)t_{\mathrm{right}}(f,\epsilon)}{t_{\mathrm{left}}(f,c\epsilon)t_{\mathrm{right}}(f,c\epsilon)}\right\}.$$

Hence, ignoring the contributions of the endpoints t_{left} and t_{right}, rescaling ϵ by a multiplicative constant c changes $\Lambda(f,\epsilon)$ by at most c.

The second gap in the upper and lower bounds comes from the fact that $\underline{N}(f,\epsilon)$ requires dividing through by $\log(\omega_{\max}/\omega_{\min})$. This is not simply an artificat of the proof. For example, for any k-piecewise linear function, there exists a set of k intervals $\mathcal{I}=\{I_i\}_{1\le i\le k}$ such that f is linear on each interval; measuring f at $\{x_{l(I)},x_{r(I)},x_{m(I)}\}_{I\in\mathcal{I}}$ would be enough to estimate f with zero-error. Hence, we must have $\underline{N}(f,\epsilon)\lesssim k$. On the other hand, $\Lambda(f,\epsilon)\approx k\log(1/\epsilon)$, and indeed one can show that for a k-piecewise linear function, $\log(\omega_{\max}/\omega_{\min})\approx\log(1/\epsilon)$, yielding the necessary cancelation. As with the term $\log\frac{t_{\mathrm{left}}(f,\epsilon)t_{\mathrm{right}}(f,\epsilon)}{t_{\mathrm{left}}(f,c\epsilon)t_{\mathrm{right}}(f,c\epsilon)}$, $\log(\omega_{\max}/\omega_{\min})$ grows like at most $\log(1/\epsilon)$ for most reasonable functions, and can be bounded by $\log(\max_{x\in[0,1]}f''(x)/\min_{x\in[0,1]}f''(x))$ for twice-differentiable functions f. Overall, we conjecture that the true sample complexity lies closer to $\underline{N}(f,\epsilon)$, because in the noiseless setting, one can approximate left- and right-derivatives of f to arbitrary accuracy using just two-points. This makes it possible to learn a 2-piecewise linear function with a constant number of function evaluations, rather than the $O(\log(1/\epsilon))$ implied by $\Lambda(f,\epsilon)$.

2.3 Recursive Secant Approximation Algorithm

We now describe the algorithm that achieves the claimed sample complexity of Theorem 2.2. Consider a binary tree $\mathcal{T}$ whose nodes represent intervals on $[0,1]$ such that the root is $[0,1]$, and each internal node representing $[a,b]$ has children $[a,(a+b)/2]$ and $[(a+b)/2,b]$. If the leaves of $\mathcal{T}$ are denoted as $\mathcal{L}(\mathcal{T})$ then

we conclude that the intersection of any two leaves $I, I' \in \mathcal{L}(\mathcal{T})$ is at most a point $|I \cap I'| = 0$, and that the leaves cover the full interval $\bigcup_{I \in \mathcal{L}(\mathcal{T})} I = [0,1]$. For a given tree $\mathcal{T}$ and function f define $\mathrm{Sec}[f, \mathcal{T}](x)$ as $\mathrm{Sec}[f, I](x)$ for the $I \in \mathcal{L}(\mathcal{T})$ such that $x \in I$ (breaking ties arbitrarily). Note that $\mathrm{Sec}[f, \mathcal{T}](x)$ evaluates f only at $|\mathcal{L}(\mathcal{T})| + 1$ locations. Moreover, we have that $\sup_{x \in [0,1]} |\mathrm{Sec}[f, \mathcal{T}](x) - f(x)| \leq \max_{I \in \mathcal{L}(\mathcal{T})} 2\Delta(f, I)$ and given a tree $\mathcal{T}$, one can compute $\Delta(f, I)$ for each $I \in \mathcal{L}(\mathcal{T})$ using just $|\mathcal{L}(\mathcal{T})|$ additional samples, which could then be reused if the leaves produced children.

Our algorithm starts with a tree $\mathcal{T}$ of depth 2 such that $\mathcal{L}(\mathcal{T}) = \{[0, 1/2], [1/2, 1]\}$ and at each time step, two children are appended to the leaf $\arg\max_{I \in \mathcal{L}(\mathcal{T})} \Delta(f, I)$. This process continues until the first time in which $\max_{I \in \mathcal{L}(\mathcal{T})} \Delta(f, I) \leq \epsilon/2$. Then it is easy to see that in the resulting tree, for each $I \in \mathcal{L}(\mathcal{T})$ we have $\Delta(\mathsf{parent}(I)) > \epsilon/2$ and $\Delta(f, I) \leq \epsilon/2$. Let $\mathcal{T}'$ be the largest tree such that $\Delta(f, I) > \epsilon/2$ for all $I \in \mathcal{L}(\mathcal{T}')$. If each leaf of $\mathcal{T}'$ has two children then the resulting tree's leaves will each have $\Delta(f, I) \leq \epsilon/2$, and because $\mathcal{T}'$ was the largest tree, we have $|\mathcal{L}(\mathcal{T})| \leq 2|\mathcal{L}(\mathcal{T}')|$. The next lemma, proved in Section 2.4, will be instrumental in relating the cardinality of this $\epsilon/2$-suboptimal "largest tree" to a digestable quantity that depends only on f.

Lemma 2.4 *Let $[a, b] \subset [0, 1]$, and suppose that $\Delta(f, [a, b]) \geq \epsilon$. Then for any $\alpha \in (0, 1)$, $\int_a^b \omega(f, x, \epsilon(1 - \alpha))^{-1} dx \geq \frac{2\alpha}{1+\alpha}$.*

Recall the definition of $\mathcal{T}'$ from above, where each $I \in \mathcal{L}(\mathcal{T}')$ satisfies $\Delta(f, I) > \epsilon/2$. By definition, $t_{\mathrm{left}}(f, \epsilon/2) := \inf\{t \leq 1/2 : \Delta(f, [0, 2t]) \geq \epsilon/2\}$ which implies that there is an $I_{left} \in \mathcal{L}(\mathcal{T}')$ such that $[0, 2t_{\mathrm{left}}(f, \epsilon/2)] \subseteq I_{left}$, define I_{right} analogously. We can bound the number of leaves:

$$\begin{aligned}
|\mathcal{L}(\mathcal{T}')| &= 2 + \sum_{I \in \mathcal{L}(\mathcal{T}') \setminus \{I_{left}, I_{right}\}} 1 \overset{(i)}{\leq} 2 + \sum_{I \in \mathcal{L}(\mathcal{T}') \setminus \{I_{left}, I_{right}\}} \frac{3}{2} \int_{x \in I} \omega(f, x, \epsilon/4)^{-1} dx \\
&\overset{(ii)}{\leq} 2 + \frac{3}{2} \int_{2t_{\mathrm{left}}(f, \epsilon/2)}^{1 - 2t_{\mathrm{right}}(f, \epsilon/2)} \omega(f, x, \epsilon/4)^{-1} dx \overset{(iii)}{\leq} 2 + 3 \int_{t_{\mathrm{left}}(f, \epsilon/2)}^{1 - t_{\mathrm{right}}(f, \epsilon/2)} \omega(f, x, \epsilon/2)^{-1} dx \\
&= 2 + 3\Lambda(f, \epsilon/2)
\end{aligned}$$

where (i) follows from Lemma 2.4 with $\alpha = 1/2$ applied to each I in the sum, (ii) follows from the the fact that $[0, 2t_{\mathrm{left}}(f, \epsilon/2)] \subseteq I_{left}$ and $[1 - 2t_{\mathrm{right}}(f, \epsilon/2), 1] \subseteq I_{right}$, and (iii) follows from $\omega(f, x, \epsilon/4) \geq \frac{1}{2}\omega(f, x, \epsilon/2)$ for $x \in [t_{\mathrm{left}}(f, \epsilon/2), 1 - t_{\mathrm{right}}(f, \epsilon/2)]$ by Proposition 2.3 and that we trivially have $[2t_{\mathrm{left}}(f, \epsilon/2), 1 - 2t_{\mathrm{right}}(f, \epsilon/2)] \subseteq [t_{\mathrm{left}}(f, \epsilon/2), 1 - t_{\mathrm{right}}(f, \epsilon/2)]$.

By a calculation above, we concluded that a tree $\mathcal{T}$ could be constructed (including validating that $\Delta(f, I) \leq \epsilon/2$ on the leaves) with just $2|\mathcal{L}(\mathcal{T})| + 1$ evaluations of f. Since $|\mathcal{L}(\mathcal{T})| \leq 2|\mathcal{L}(\mathcal{T}')|$, we conclude that $\mathcal{T}$ is constructed after evaluating f at most $9 + 12\Lambda(f, \epsilon/2)$ times, obtaining the claimed result of Theorem 2.2.

3 Efficiently Learning a Convex Function with Noise

In this section we return to the motivating problem where we sample at $x_s \in [0, 1]$ and observe $y_s = f(x_s) + w_s$ where w_s is mean-zero, σ^2-subgaussian noise. The previous section proposed an algorithm that was shown to achieve ϵ error using $\approx \Lambda(f, \epsilon)$ noiseless evaluations of f. Roughly speaking, if $\{y_i\}_{i=1}^m$ are noisy estimates of f at x, then $|\frac{1}{m}\sum_{i=1}^m y_i - f(x)| \leq \epsilon$ with constant probability whenever $m \gtrsim \sigma^2\epsilon^{-2}$. Intuitively, these observations suggest the existance of an algorithm with sample complexity that roughly scales like $\Lambda(f, \epsilon) \max\{1, \sigma^2\epsilon^{-2} \log(\Lambda(f, \epsilon)))\}$, the log factor accounting for a union bound over the $\Lambda(f, \epsilon)$ estimates being accurate simultaneousy. This section proposes a simple algorithm and shows formally that this number of samples is essentially necessary and sufficient.

The noiseless lower bound of Theorem 2.1 follows from the packing, requiring any algorithm run on an instance $g \in \mathcal{G}_{f, 2\epsilon}$ to put at least one sample in each I_i to unqiuely identify $g \in \mathcal{G}_{f, 2\epsilon}$. The noisy lower bound is intuitively similar, but requires more care to show that each of the $\underline{N}(f, 2\epsilon)$ intervals has to be sampled about $\sigma^2\epsilon^{-2} \log(\underline{N}(f, 2\epsilon)/\delta)$ times, the main challenge showing that the $\underline{N}(f, 2\epsilon)$ appears in the log to account for the fact that any one of the interval estimations can err.

Theorem 3.1 *Fix convex $f : [0,1] \to \mathbb{R}$, $\epsilon > 0$, and $\delta \in (0, 1/2)$. Let $\underline{N}(f,\epsilon)$ and $\mathcal{G}_{f,\epsilon}$ be as in Lemma 2.2 and Equation (4). If* Alg *is (ϵ,δ)-correct over $\mathcal{G}_{f,2\epsilon}$, then $\mathbb{E}_{f,\mathsf{Alg}}[T] \gtrsim \underline{N}(f,2\epsilon)\max\{1, \frac{\sigma^2}{2\epsilon^2}\log(1/\delta)\}$, and the average sample complexity over $\mathcal{G}_{f,2\epsilon}$ is at least*

$$\frac{1}{|\mathcal{G}_{f,2\epsilon}|}\sum_{g\in\mathcal{G}_{f,2\epsilon}} \mathbb{E}_{g,\mathsf{Alg}}[T] \gtrsim \underline{N}(f,2\epsilon)\max\{1, \frac{\sigma^2}{2\epsilon^2}\log(\underline{N}(f,2\epsilon)/\delta)\}.$$

In Theorem 3.1, the first part characterizes the difficulty of estimating a *particular* fixed f, and follows from considering the difficulty of a composite hypothesis test: is the true function f, or some $g \in \mathcal{G}_{f,2\epsilon}$ with $g \neq f$? The second part characterizes the average difficulty over the whole class $\mathcal{G}_{f,2\epsilon}$ and follows from a multiple hypothesis testing problem.

3.1 Recursive Secant Approximation with Noise

Algorithm 1: Active Learning Algorithm for Convex Regression

1 **Input** Bias-variance tradeoff β, confidence δ
2 **Global** mutable dictionaries/maps $\widetilde{f}, T, \widetilde{\delta}$
3 **Initialize** Tree of intervals $\mathcal{T} = \{[0,1]\}$, $\widetilde{f}(x) = 0, T(x) = 0, \widetilde{\delta}(x) = \delta/6$ for $x \in \{0, 1/2, 1\}$
4 **For** round $t = 1, 2, \ldots$
5 **While** $\exists I : (1+\beta)\overline{B}(I, T, \widetilde{\delta}) \leq \Delta(\widetilde{f}, I)$
6 Bisect I into two even intervals I_1 and I_2
7 **For** $j = 1, 2$
8 **Append** I_j to $\mathcal{T}$ as a child of I
9 $\widetilde{\delta}(x_{m(I_j)}) \leftarrow \delta/2|\mathcal{L}(\mathcal{T})|^2$, $\widetilde{f}(x_{m(I_j)}) \leftarrow 0$, $T(x_{m(I_j)}) \leftarrow 0$
10 $I^* \leftarrow \arg\max_{I\in\mathcal{L}(\mathcal{T})} \max\{\underline{B}(I, T, \widetilde{\delta}), \Delta(\widetilde{f}, I) + \overline{B}(I, T, \widetilde{\delta})\}$ (break ties arbitrarily)
11 **If** $\underline{B}(I^*, T, \widetilde{\delta}) \leq \Delta(\widetilde{f}, I^*) + \overline{B}(I^*, T, \widetilde{\delta})$ and $\underline{B}(I^*, T, \widetilde{\delta}) \leq \phi(T(x_{m(I^*)}), \widetilde{\delta}(x_{m(I^*)}))$
12 $x_t \leftarrow x_{m(I^*)}$
13 **Else** $x_t \leftarrow$ maximizer of $\phi(T(x), \widetilde{\delta}(x))$ for $x \in \{x_{l(I^*)}, x_{r(I^*)}\}$
14 **Sample** at x_t to observe y_t
15 **Update** $T(x_t) \leftarrow T(x_t) + 1$, $\widetilde{f}(x_t) \leftarrow y_t \cdot \frac{1}{T(x_t)} + \widetilde{f}(x_t) \cdot \frac{T(x_t)-1}{T(x_t)}$

We now describe how to generalize the algorithm of Section 2.3 to noisy observations. Fix some time t and let $\{(x_s, y_s)\}_{s=1}^t$ be the collection of noisy function evaluation pairs. In the algorithm, $T(x) = \sum_{s=1}^t \mathbf{1}\{x_s = x\}$ will denote the number of times the point $x \in [0,1]$ has been sampled and $\widetilde{f}(x) = \frac{1}{T(x)\vee 1}\sum_{s=1}^t \mathbf{1}\{x_s = x\}\ y_s$. Let $\phi(t,\delta)$ denote an anytime confidence interval such that $\mathbb{P}\left(\bigcup_{t=1}^\infty \{|\frac{1}{t}\sum_{s=1}^t w_s| \geq \phi(t,\delta)\}\right) \leq \delta$[2]. In addition to T and $\widetilde{f}$, we maintain a function $\widetilde{\delta}$ such that $\mathbb{P}\left[\forall t : |\frac{1}{t}\sum_{s=1}^t \mathbf{1}\{x_s = x\}w_s| \leq \phi(T(x), \widetilde{\delta}(x))\right] \geq 1 - \widetilde{\delta}(x)$. Finally, define confidence bounds

$$\underline{B}(I, T, \widetilde{\delta}) = \tfrac{1}{2}\max_{x\in\{x_{l(I)}, x_{r(I)}\}} \phi(T(x), \widetilde{\delta}(x)) \ , \quad \overline{B}(I, T, \widetilde{\delta}) = \underline{B}(I, T, \widetilde{\delta}) + \phi(T(x_{m(I)}), \widetilde{\delta}(x_{m(I)}))$$

One can show that the error of the empirical secant approximations using $\widetilde{f}$ are controlled via the sandwich relationship $-\underline{B}(I, T, \widetilde{\delta}) \leq \frac{\mathrm{Sec}[\widetilde{f}, I](x) - f(x)}{2} \leq \Delta(\widetilde{f}, I) + \overline{B}(I, T, \widetilde{\delta})$.

For some $\beta > 0$, Algorithm 1 maintains the condition $(1+\beta)\overline{B}(I, T, \widetilde{\delta}) \geq \Delta(\widetilde{f}, I)$ using the while loop of Line 5. This is to ensure that the stochastic variance always dominates the bias of the approximation. The parameter $\beta > 0$ appears to have little effect on performance as long as it is smaller than 1; we recommend setting $\beta = 1/2$. The definition of I^* is motivated by sandwich relationship noted above. And in each case,

[2] For example, $\phi(t,\delta) = \sqrt{16\sigma^2\log(\log_2(2t)/\delta)/t}$ suffices but we recommend using [12, Theorem 8].

x_t is chosen in order to minimize the maximum confidence bound relevant to the interval I^*. The values of $\widetilde{\delta}(x_{m(I_j)})$ are chosen such that $\sum_{x:T(x)>0} \widetilde{\delta}(x) \leq \delta$ since $3 \cdot \frac{1}{6} + \sum_{k=2}^{\infty} \frac{1}{2k^2} \leq 1$. The proof of the following guarantee is found in Appendix G.

Theorem 3.2 *Fix convex $f : [0,1] \to \mathbb{R}$ and some $\delta \in (0, 1/2)$. Let $\overline{N}(f, \epsilon)$ be defined as in Theorem 2.2. With probability at least $1 - \delta$, for all $\epsilon > 0$, once the algorithm has made*

$$\overline{N}(f, \tfrac{\beta}{(2+\beta)}\epsilon) \max\{1, (2+\beta)^2\sigma^2\epsilon^{-2} \log(\overline{N}(f, \tfrac{\beta}{(2+\beta)}\epsilon) \log((2+\beta)^2\epsilon^{-2})/\delta)\}$$

noisy observations of f we have $\sup_{x\in[0,1]} |\mathrm{Sec}[f, \mathcal{T}](x) - f(x)| \leq \epsilon$.

4 Empirical Results

In this section, we validate our theoretical results through empirical comparisons of active and passive sampling using simulated data and data drawn from the behavioral literature. In all experiments, a query at $x \in [0,1]$ results in an observation $y \overset{i.i.d.}{\sim} \mathcal{N}(f(x), \sigma^2)$ where σ depends on the experiments and is known to the algorithm. We construct our confidence intervals $\phi(t, \delta)$ using [12, Theorem 8], scaled by σ. Let $\vec{s}$ be the dyadic sequence $\vec{s} = \{0, 1, 1/2, 1/4, 3/4, 1/8, \dots\}$. Let dyadic sampling on an interval $[a, b]$ be defined as the sequence $a + (b-a)\vec{s}$. We consider three methods: 1) `passive-dyadic`: dyadic sampling on $[0,1]$, 2) `active`: Algorithm 1, and 3) `active-dyadic`: on odd times sample according to Algorithm 1 and on even times pick the interval in $\mathcal{L}(\mathcal{T})$ that has been sampled the fewest times on even times and sample at the next location in the dyadic sequence on that interval. Regardless of the method, to compute an estimator we use all the data collected from the method to compute the least squares estimator subject to the function being convex[3][4].

We begin by confirming theoretical predictions. The empirical results are presented in Figure 1 for $\sigma = .01$. The dotted and solid lines represent the L_∞ error on $[0,1]$ and $[0.1, 0.9]$, respectively. When $f(x) = \max\{1 - 7x, 0\}$, the optimal strategy is to concentrate samples on $\{0, 1/7, 1\}$, and we accordingly observe that `active` and `active-dyadic` both adapt, outperforming `passive-dyadic`. We point out that `active` has a saw-tooth behavior due to splitting times and `active-dyadic` corrects for this. When $f(x) = (1-x)^2$, the curvature is constant and the active and passive methods perform equally well. Finally, we turn to exploring the potential impact of active sampling on real functions. Participants (250) were asked to choose between a hypothetical reward of \$100 given immediately and a reward of \$115 given at a time $x = 0, 1, 2, \dots, 64$ days in the future (times were randomized and rescaled to be in $[0,1]$). We fit a convex function to this data using least squares and sampled from it as above. At $\sigma = .01$, we observe that active slightly outperforms perform passive for high sample sizes on this function, shown in the right panel of Figure 1.

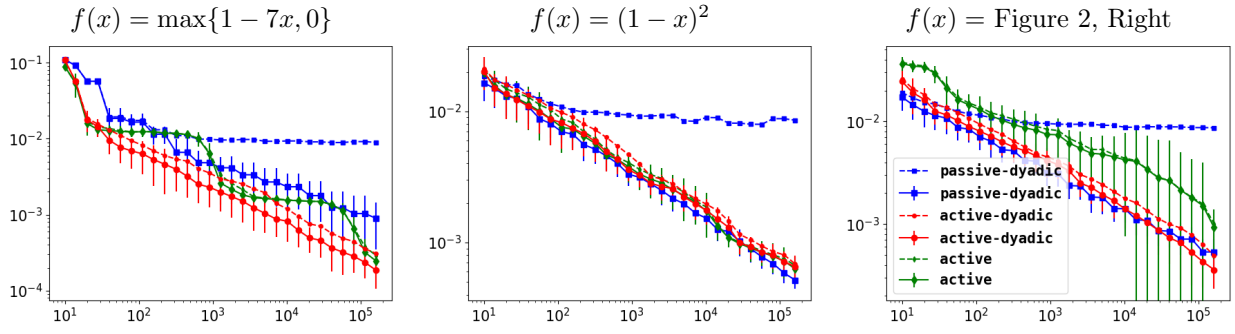

Figure 1: Active versus passive performance on three functions. The x-axis is the number of samples taken, and the y-axis is L_∞ error.

However, when we increase the noise by a factor of 10 to $\sigma = .1$, our active sampling procedure yields little improvements over passive (Figure 2, Left). To deduce if any algorithm can do substantially better, we

[3]For a finite number of observations, convex regression is a quadratic program [1].
[4]That is, the `active` estimator is *not* a collection of secant approximations.

consider sampling at the endpoints of the lower-bound packing at different values of ϵ (see the discussion following Theorem 2.1). We believe the oracle strategy is a proxy for the optimal strategy and can expose how much room there is for improvement. Noting the log-scale, we observe that the oracle strategy outperforms `passive-dyadic` by a non-negligible amount, leaving the door open for improved algorithms.

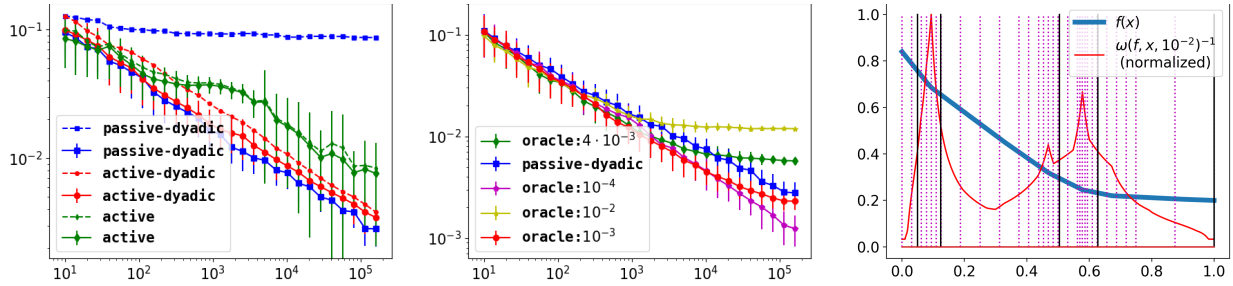

Figure 2: Left: passive and active methods. Center: passive and oracle methods for different values of ϵ. Right: The fitted function (bold), $\omega(f, x, 0.01)^{-1}$ (solid), oracle sampling locations with $\epsilon = .01$ (vertical solid), `active` sampling locations (vertical dotted).

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[2] T Tony Cai, Mark G Low, Yin Xia, et al. Adaptive confidence intervals for regression functions under shape constraints. *The Annals of Statistics*, 41(2):722–750, 2013.

[3] Rui Castro, Rebecca Willett, and Robert Nowak. Faster rates in regression via active learning. In *NIPS*, volume 18, pages 179–186, 2005.

[4] Sabyasachi Chatterjee. An improved global risk bound in concave regression. *Electronic Journal of Statistics*, 10(1):1608–1629, 2016.

[5] Lutz Dümbgen et al. Optimal confidence bands for shape-restricted curves. *Bernoulli*, 9(3):423–449, 2003.

[6] Lutz Dümbgen, Sandra Freitag, and Geurt Jongbloed. Consistency of concave regression with an application to current-status data. *Mathematical methods of statistics*, 13:69–81, 2004.

[7] Ronald Aylmer Fisher. *The design of experiments*. Oliver And Boyd; Edinburgh; London, 1937.

[8] Shane Frederick, George Loewenstein, and Ted O'donoghue. Time discounting and time preference: A critical review. *Journal of economic literature*, 40(2):351–401, 2002.

[9] Leonard Green and Joel Myerson. A discounting framework for choice with delayed and probabilistic rewards. *Psychological bulletin*, 130(5):769, 2004.

[10] Piet Groeneboom and Geurt Jongbloed. *Nonparametric estimation under shape constraints*, volume 38. Cambridge University Press, 2014.

[11] Adityanand Guntuboyina and Bodhisattva Sen. Global risk bounds and adaptation in univariate convex regression. *Probability Theory and Related Fields*, 163(1-2):379–411, 2015.

[12] Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *The Journal of Machine Learning Research*, 2015.

[13] Alexander Korostelev. On minimax rates of convergence in image models under sequential design. *Statistics & Probability Letters*, 43(4):369–375, 1999.

[14] Joel Myerson and Leonard Green. Discounting of delayed rewards: Models of individual choice. *Journal of the experimental analysis of behavior*, 64(3):263–276, 1995.

[15] Yuancheng Zhu, Sabyasachi Chatterjee, John Duchi, and John Lafferty. Local minimax complexity of stochastic convex optimization. *Advances in Neural Information Processing Systems*, 29, 2016.

A Notation

Notation
$[n] = \{1, 2, \ldots, n\}$ for any $n \in \mathbb{N}$
$I = [a, b] = \{x \in \mathbb{R} : a \leq x \leq b\}$
$x_{l(I)} = a$
$x_{r(I)} = b$
$x_{m(I)} = \frac{a+b}{2}$
$\mathrm{Sec}[f, I](x) = \frac{x_{r(I)} - x}{x_{r(I)} - x_{l(I)}} f(x_{l(I)}) + \frac{x - x_{l(I)}}{x_{r(I)} - x_{l(I)}} f(x_{r(I)})$
$\Delta(f, I) := \mathrm{Sec}[f, I](x_{m(I)}) - f(x_{m(I)})$
$\Delta(f, x, t) := \Delta(f, [x - t, x + t])$
$t_{\mathrm{left}}(f, \epsilon) := \inf \{t \leq 1/2 : \Delta(f, t, t) \geq \epsilon\}$
$t_{\mathrm{right}}(f, \epsilon) := \inf \{t \leq 1/2 : \Delta(f, 1 - t, t) \geq \epsilon\}$
$\epsilon^*(f, x) := \Delta(f, x, \min\{x, 1 - x\})$
$\overline{I}(f, \epsilon) := [t_{\mathrm{left}}(f, \epsilon), t_{\mathrm{right}}(f, \epsilon)]$
$\omega(f, x, \epsilon) := \min \{t \in [0, \min \{x, 1 - x\}] : \Delta(f, x, t) \geq \epsilon\}$
$\omega_{\max}(f, \epsilon) = \max_{x \in [t_{\mathrm{left}}(f,\epsilon), 1 - t_{\mathrm{right}}(f,\epsilon)]} \omega(f, x, \epsilon)$
$\omega_{\min}(f, \epsilon) = \min_{x \in [t_{\mathrm{left}}(f,\epsilon), 1 - t_{\mathrm{right}}(f,\epsilon)]} \omega(f, x, \epsilon)$
$\Lambda(f, \epsilon) := \int_{t_{\mathrm{left}}(f,\epsilon)}^{1 - t_{\mathrm{right}}(f,\epsilon)} \omega(f, x, \epsilon)^{-1} dx$
$\mathcal{T}$, Binary tree of intervals on $[0, 1]$ (Section 2.3)
$\mathcal{L}(\mathcal{T})$ Leaves, disjoint intervals, of binary tree. $\bigcup_{I \in \mathcal{L}(\mathcal{T})} I = [0, 1]$

B Bounding ℓ_∞ with error at a single point

Lemma B.1 *Let $f : [0, 1] \to \mathbb{R}$ be convex. For any $x, z \in (t_1, t_2) \subset [0, 1]$, one has that*

$$\mathrm{Sec}[f, [t_1, t_2]](x) - f(x) \geq \{\mathrm{Sec}[f, [t_1, t_2]](z) - f(z)\} \cdot \min \left\{ \frac{t_2 - x}{t_2 - z}, \frac{x - t_1}{z - t_1} \right\} \tag{5}$$

Proof

Note that adding an affine function to f does not change the value of $\mathrm{Sec}[f, [t_1, t_2]](x) - f(x)$. Thus, we may assume that $f(t_1) = f(t_2) = 0$. Without loss of generality, we may also take $t_1 = 0$ and $t_2 = 1$. With these simplifications,

$$\mathrm{Sec}[f, [0, 1]] = \mathrm{Sec}[f, [t_1, t_2]] = 0 \tag{6}$$

and hence our goal is show that

$$-f(x) \geq \begin{cases} \frac{x}{z} \cdot (-f(z)) & 0 \leq x \leq z \leq 1 \\ \frac{1-x}{1-z} \cdot (-f(z)) & 0 \leq z \leq x \leq 1 \end{cases}, \tag{7}$$

or equivalently,

$$f(z) \geq \begin{cases} \frac{z}{x} \cdot f(x) & 0 \leq x \leq z \leq 1 \\ \frac{1-z}{1-x} \cdot f(x) & 0 \leq z \leq x \leq 1 \end{cases}. \tag{8}$$

To this end, fix a subgradient $g \in \partial f(x)$ and some $z \geq x$. By the definition of the subgradient, it holds that

$$f(x) + g(t-x) \leq f(t) \quad \forall t \in [0,1] \ . \tag{9}$$

By choosing $t = 0$ and $t = z$ in the above display, we verify that

$$f(x) + g(0-x) \overset{(a)}{\leq} 0 \quad \text{and} \quad f(x) + g(z-x) \overset{(b)}{\leq} f(z) \ .$$

Combining (a) and (b), and noting that $z \geq x$

$$\begin{aligned} f(z) &\overset{(b)}{\geq} f(x) + g(z-x) \\ &\overset{(a)}{\geq} f(x) + \frac{1}{x} f(x)(x-z) \\ &= f(x)\left(1 + \frac{z-x}{x}\right) \quad = \frac{x}{z} \cdot f(x) \ , \end{aligned} \tag{10}$$

as needed. On the other hand, suppose $x \geq z$. Noting that the function $\widetilde{f}(t) = f(1-t)$ is convex and satisfies $\widetilde{f}(0) = \widetilde{f}(1) = 0$, we have

$$f(z) = \widetilde{f}(1-z) \overset{\text{(Equation (10))}}{\geq} \left(\frac{(1-z)}{(1-x)}\right) \widetilde{f}(1-x) = \left(\frac{(1-z)}{(1-x)}\right) f(x) \tag{11}$$

■

We will also find the following fact useful

Lemma B.2 *Let $\varphi(t)$ be a convex function satisfying $\varphi(0) = 0$. Then for all $c \in [0,1]$, $\varphi(ct) \leq c\varphi(t)$.*

Proof $c\varphi(t) = c\varphi(t) + (1-c)\varphi(0) \leq \varphi(ct + (1-c)0)\varphi(ct)$, where the first equality uses $\varphi(0) = 0$ and the second uses convexity. ■

In particular, this fact implies that if $[a,b] \subset [c,d]$, then the secant approximation obtained by using $[a,b]$ is no worse than the one obtained by using $[c,d]$

Lemma B.3 *For any $x \in [a,b] \subset [c,d]$, $\mathrm{Sec}[f,[a,b]](x) \leq \mathrm{Sec}[f,[c,d]](x)$*

Proof It suffices to prove that this is the case when $a = c$ or $d = b$, since then $\mathrm{Sec}[f,[a,b]](x) \leq \mathrm{Sec}[f,[c,b]](x) \leq \mathrm{Sec}[f,[c,d]](x)$. We assume without los of generality that $a = c$. Then, it suffices to show that, for $t \geq x$, the map $t \mapsto \mathrm{Sec}[f,[a,a+t]](x) =$ is non-decreasing. We have

$$\mathrm{Sec}[f,[a,a+t]](x) = \frac{f(a)(a+t-x) + (x-a)f(a+t)}{t} = \frac{\varphi(t)}{t} \tag{12}$$

where $\varphi(t) = f(a)(a+t-x) + (x-a)f(a+t)$. Since φ is convex (sum of affine function and convex function as $x \geq a$, and $t \mapsto f(a+t)$ is convex), and $\varphi(0) = 0$, we conclude by Lemma B.2 that $\frac{\varphi(t)}{t}$ is non-decreasing, as needed. ■

C Proof of Packing, Lemma 2.2

We prove the lemma by showing (Theorem C.1) that we can construct a packing of intervals I_i of size at least $\underline{N}(f,\epsilon) + 1$ such that $\Delta(f, I_i) = \epsilon$:

Theorem C.1 *Define the lower sample complexity*

$$\underline{N}(f,\epsilon) = \frac{1}{4(1 + \log(\omega_{\max}(f,\epsilon)/\omega_{\min}(f,\epsilon)))} \int_{t_{\text{left}}(f,\epsilon)}^{1-t_{\text{right}}(f,\epsilon)} \omega(f,u,\epsilon)^{-1} du - 3 \ , \tag{13}$$

where $\omega_{\max}(f,\epsilon) := \max\{\omega(f,u,\epsilon) := u \in [t_{\text{left}}(f,\epsilon), 1 - t_{\text{right}}(f,\epsilon)]$ and $\omega_{\min}(f,\epsilon)$ is defined analogously. Then, there exists intervals $\{I_i = [a_i, b_i]\}_{1 \leq i \leq \underline{N}(f,\epsilon)+1}$ with disjoint interiors, such that $\bigcup_{i=1}^{\underline{N}(f,\epsilon)+1} I_i \subset [2t_{\text{left}}(\epsilon,f), 1 - 2t_{\text{right}}(\epsilon,f)]$, and $\Delta(f, I_i) = \epsilon$.

Note that Lemma 2.2 requires a packing of intervals I_i such that $\Delta(f, I_i) > \epsilon$, that is, a strict inequality. To prove this slightly stronger packing, we use the fact that $\underline{N}(f, \epsilon)$ is right-continuous by Proposition H.1. Thus, there exist some $\eta > 0$ such that $\underline{N}(f, \epsilon + \eta) + 1 \geq \underline{N}(f, \epsilon)$. We then use the packing Theorem C.1 with ϵ replaced by $\epsilon + \eta$, which yields a family of $\underline{N}(f, \epsilon + \eta) + 1 \geq \underline{N}(f, \epsilon)$ intervals I_i such that $\Delta(f, I_i) = \epsilon + \eta > \epsilon$, thereby proving Lemma 2.2

C.1 Proof of Theorem C.1

We construct the packing by choosing a sequence of interval midpoints m_i and interval lengths t_i, such that the intervals $I_i := [m_i - t_i, m_i + t_i] = [a_i, b_i]$ only overlap at their endpoints, and such that $\Delta(f, m_i, t_i) = \epsilon$. To do this, we define $t_0 = m_0 = t_{\text{left}}(f, \epsilon)$. By definition of $t_{\text{left}}(f, \epsilon)$, we have the equality $\Delta(f, t_{\text{left}}(f, \epsilon), t_{\text{left}}(f, \epsilon)) = \epsilon$. Now, for each $i \geq 1$, we define

$$t_i = \sup \left\{ t \in [0, \frac{1 - b_{i-1}}{2}] : \Delta(f, b_{i-1} + t, t) \leq \epsilon \right\} \quad \text{and} \quad (a_i, m_i, b_i) = (b_{i-1}, b_{i-1} + t_i, b_{i-1} + 2t_i) \tag{14}$$

One can think of t_i as as the equivalent of t_{left}, but starting at b_{i-1} rather than zero. Note that $\Delta(f, b_{i-1} + t, t)$ is non-decreasing and continuous in t (Lemma H.4), and thus, if there exists a $t \in [0, \frac{1-b_{i-1}}{2}]$ such that $\Delta(f, b_{i-1} + t, t) \geq \epsilon$, then the supremum in the definition of t_i will be attained for a t_i such that $\Delta(f, b_{i-1} + t, t) = \Delta(f, m_i, t_i) = \epsilon$. Thus, we will terminate the construction at $i = n$, where n is the first number satisfying $b_n \geq 1 - 2t_{\text{right}}(f, \epsilon)$, or $\Delta(f, b_n + t_{n+1}, t_{n+1}) < \epsilon$.

Collecting what we have established thus far,

1. $\Delta(f, m_i, t_i) = \Delta(f, b_i + t_i, t_i) = \epsilon$
2. By definition, $a_1 = 2t_{\text{left}}(f, \epsilon)$. And, by the stopping condition, $b_n \geq 1 - 2t_{\text{right}}(f, \epsilon) \geq a_n$
3. Hence, since $b_i = a_{i+1}$, we have that $\bigcup_{I_i} = [a_1, b_n] \supseteq [2t_{\text{left}}(f, \epsilon), 1 - 2t_{\text{right}}(\epsilon, f)]$, and that I_i have disjoint interiors.

To conclude, we want to show that there is some $K(f, \epsilon)$ such that

$$\int_{I_i} \omega(f, u, \epsilon)^{-1} du \leq K(f, \epsilon) \tag{15}$$

for all $i \in [n]$. Since then

$$\int_{2t_{\text{left}}(f,\epsilon)}^{1-2t_{\text{right}}(f,\epsilon)} \omega(f, u, \epsilon)^{-1} du \leq \int_{a_1}^{b_n} \omega(f, u, \epsilon)^{-1} du \leq \sum_{i=1}^{n} \int_{I_i} \omega(f, u, \epsilon)^{-1} du \leq nK(f, \epsilon) \tag{16}$$

This is accomplished by the following lemma, which is a rephrasing of Lemma C.3, proved in the subsection below.

Lemma C.1 *Let $m \in \overline{I}(f, \epsilon)$ and choose $t = \omega(f, m, \epsilon)$, so that $\Delta(f, m, t) = \epsilon$. Then if $\omega_{\min} = \inf_{u \in [m-t, m+t]} \omega(f, u, \epsilon)$, one has*

$$\int_{m-t}^{m+t} \omega(f, u, \epsilon)^{-1} du \leq 4 \left(1 + \log \frac{t}{\omega_{\min}} \right) \tag{17}$$

Similarly, if $0 \leq m - t$ or $m + t \leq 1$, one has

$$\max\{ \int_{m-t}^{m} \omega(f, u, \epsilon)^{-1} du, \int_{m}^{m+t} \omega(f, u, \epsilon)^{-1} du \} \leq 2 \left(1 + \log \frac{t}{\omega_{\min}} \right) \tag{18}$$

Hence, if $\omega_{\min} = \min\{\omega(f, u, \epsilon) : u \in [t_{\text{left}}(f, \epsilon), 1 - t_{\text{right}}(f, \epsilon)]$ and $\omega_{\max} = \min\{\omega(f, u, \epsilon) : u \in [t_{\text{left}}(f, \epsilon), 1 - t_{\text{right}}(f, \epsilon)]$, we have

$$\int_{2t_{\text{left}}(f,\epsilon)}^{1-2t_{\text{right}}(f,\epsilon)} \omega(f, u, \epsilon)^{-1} du \leq 4n(1 + \log \frac{\omega_{\max}}{\omega_{\min}})$$

and that

$$\int_{t_{\rm left}(f,\epsilon)}^{2t_{\rm left}(f,\epsilon)} \omega(f,u,\epsilon)^{-1}du + \int_{1-2t_{\rm right}(f,\epsilon)}^{1-t_{\rm right}(f,\epsilon)} \omega(f,u,\epsilon)^{-1}du \leq 2\cdot 2(1+\log\frac{\omega_{\max}}{\omega_{\min}}) \ .$$

Hence

$$\int_{t_{\rm left}(f,\epsilon)}^{1-t_{\rm right}(f,\epsilon)} \omega(f,u,\epsilon)^{-1}du \quad \leq \quad 4(n+1)(1+\log\frac{\omega_{\max}}{\omega_{\min}}) \tag{19}$$

and the number of intervals satisfies $n \geq \frac{1}{4(1+\log\frac{\omega_{\max}}{\omega_{\min}})}\int_{t_{\rm left}(f,\epsilon)}^{1-t_{\rm right}(f,\epsilon)} \omega(f,u,\epsilon)^{-1}du - 1$. Finally, we remove the last interval I_n. Since the right endpoint a_n of I_n satisfies $a_n \leq 1-2t_{\rm right}(f,\epsilon)$, the intervals $I_1,\ldots,I_{n-1}$ are contained within $[2t_{\rm right}(f,\epsilon), 1-2t_{\rm right}(f,\epsilon)]$, and

$$n-1 \geq \frac{1}{4(1+\log\frac{\omega_{\max}}{\omega_{\min}})}\int_{t_{\rm left}(f,\epsilon)}^{1-t_{\rm right}(f,\epsilon)} \omega(f,u,\epsilon)^{-1}du - 2.$$

C.2 Proof of Lemma C.3

Lemma C.2 *Let $x \in \overline{I}(f,\epsilon)$, and $\tau \in [-1,1]$, such that $u := x+\tau\omega(f,x,\epsilon) \in \overline{I}(f,\epsilon)$. Then,*

$$\omega(f,u,\epsilon) \geq \frac{(1-|\tau|)\omega(f,x,\epsilon)}{2} \tag{20}$$

Proof

We may assume without loss of generality that $\tau \in [0,1]$. For ease of notation, set

$$t = \omega(f,x,\epsilon) \qquad u = x+\tau t \qquad \widetilde{\epsilon} = \Delta(f,u,(1-\tau)t)$$

noting that $\widetilde{\epsilon}$ is the midpoint secant error on the interval $[u-(1-\tau)t, u+(1-\tau)t]$. Since $u+(1-\tau)t = x+t$ and $[u-(1-\tau)t, u+(1-\tau)t] \subseteq [x-t, x+t]$, we have that

$$\begin{aligned}\widetilde{\epsilon} &= \Delta(f,[u-(1-\tau)t, u+(1-\tau)t]) \\ &\overset{Lemma\ B.3}{\leq} \sup_{y\in[x-t,x+t]} \mathrm{Sec}[f,[x-t,x+t]](y) - f(y) \\ &\overset{Lemma\ 2.1}{\leq} 2\Delta(f,[x-t,x+t]) = 2\epsilon.\end{aligned}$$

First, if $\widetilde{\epsilon} \leq \epsilon$ then

$$\omega(f,u,\epsilon) \overset{Lemma\ H.6}{\geq} \omega(f,u,\widetilde{\epsilon}) \overset{Lemma\ H.5}{=} (1-\tau)t. \tag{21}$$

On the other hand, if $\epsilon < \widetilde{\epsilon} \leq 2\epsilon$ then

$$\omega(f,u,\epsilon) \overset{Lemma\ H.6}{\geq} \frac{\epsilon}{\widetilde{\epsilon}}\omega(f,u,\widetilde{\epsilon}) \geq \frac{1}{2}\,\omega(f,u,\widetilde{\epsilon}) \overset{Lemma\ H.5}{=} \frac{(1-\tau)t}{2}.$$

■

Lemma C.3 *Fix an ϵ and f, and let $[a,b] \subset [t_{\rm left}(f,\epsilon), 1-t_{\rm left}(f,\epsilon)]$ such that $\Delta(f,[a,b]) = \epsilon$. If $m = \frac{a+b}{2}$ then*

$$\int_a^m \omega(f,u,\epsilon)^{-1}du \leq 2(1+\log\frac{\omega(f,m,\epsilon)}{\min\limits_{u\in[a,m]}\omega(f,u,\epsilon)})$$

and

$$\int_m^b \omega(f,u,\epsilon)^{-1}du \leq 2(1+\log\frac{\cdot\omega(f,m,\epsilon)}{\min\limits_{u\in[m,b]}\omega(f,u,\epsilon)}).$$

Proof [Proof of Lemma C.3] We will prove the result for the integral for $u \in [m, b]$. The other direction is analogous. We can write $u \in [m, b]$ as $u = m + \tau t$, where $t = \omega(f, m, \epsilon)$ and $\tau \in [0, 1]$. Now, set $\omega_{\min} = \min_{u \in [m,b]}(f, \epsilon, u)$. Using Lemma C.2, we can integrate

$$\begin{aligned}
\int_m^b \omega(f, u, \epsilon)^{-1} du &= t \int_0^1 \omega(f, m + \tau t, \epsilon)^{-1} d\tau \\
&\overset{(a)}{\leq} t \int_{\tau=0}^1 \min\left\{ \frac{2}{(1-\tau)t}, \omega_{\min}^{-1} \right\} d\tau \\
&= \int_{\tau=0}^1 \min\left\{ 2/\tau, t/\omega_{\min} \right\} d\tau \\
&= \int_{\tau=0}^{2\omega_{\min}/t} t/\omega_{\min} d\tau' + \int_{2\omega_{\min}/t}^1 2/\tau' \; d\tau' \\
&= 2 + 2\log(t/(2\omega_{\min}))
\end{aligned}$$

where (a) is precisely Lemma C.2. Lastly, since $m \in [t_{\text{left}}(f, \epsilon), 1 - t_{\text{right}}(f, \epsilon)]$ and $\Delta(f, m, t) = \Delta(f, I) = \epsilon$, we conclude that $t = \omega(f, m, \epsilon)$ by Lemma H.5. ■

D Proof of Noiseless Lower Bound, Theorem 2.1

The following lemma lets us reduce to proving lower bounds on estimating the binary vector β used to to define the alternatives $G_{f,2\epsilon}$ in Equation 4:

Lemma D.1 *Fix a convex function f, an algorithm* Alg *which outputs an estimate $\widehat{f}$, and let $\mathcal{G}_{f,2\epsilon}$ denote the collection of alternatives from Equation 4, parameterized by $\beta \in \{0,1\}^n$. Define the quantity $\widehat{\beta} = \widehat{\beta}(\widehat{f})$ by*

$$\widehat{\beta}_i = \mathbb{I}(|\widehat{f}(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| > \epsilon) \tag{22}$$

for all $i \in [n]$. If Alg *is run on some $g \in \mathcal{G}_{f,2\epsilon}$ corresponding to the parameter $\beta^{(g)} \in \{0,1\}^n$, defined by*

$$g = \Big\{ f(x) + \sum_{i=1}^n \beta_i^{(g)} \mathbf{1}\{x \in I_i^{2\epsilon}\}(\text{Sec}[f, I_i^{2\epsilon}](x) - f(x)) \Big\} \in \mathcal{G}_{f,2\epsilon} \; , \tag{23}$$

then $\widehat{\beta} = \beta^{(g)}$ whenever $\|\widehat{f} - g\|_\infty \leq \epsilon$.

Proof Suppose that $\|\widehat{f} - g\|_\infty \leq \epsilon$. Then for all indices i for which $\beta_i^{(g)} = 0$, $g(x_{m(I_i^{2\epsilon})}) = f(x_{m(I_i^{2\epsilon})})$, and thus

$$|\widehat{f}(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| = |\widehat{f}(x_{m(I_i^\epsilon)}) - g(x_{m(I_i^\epsilon)})| \leq \|\widehat{f} - g\|_\infty \leq \epsilon$$

and thus

$$\widehat{\beta}_i = \mathbb{I}(|\widehat{f}(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| > \epsilon) = 0 = \beta_i^{(g)}$$

On the other hand, if $\beta_i^{(g)} = 1$, then by the reverse triangle inequality

$$\begin{aligned}
|\widehat{f}(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| &\geq |g(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| - |\widehat{f}(x_{m(I_i^\epsilon)}) - g(x_{m(I_i^\epsilon)})| \\
&\geq |g(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| - \|\widehat{f} - g\|_\infty \\
&\geq |g(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| - \epsilon
\end{aligned}$$

By the definition of the packing, $|g(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| > 2\epsilon$, from which we conclude

$$|\widehat{f}(x_{m(I_i^\epsilon)}) - f(x_{m(I_i^\epsilon)})| > 2\epsilon - \epsilon = \epsilon \implies \widehat{\beta}_i = 1 = \beta_i^{(g)}$$

■

For each $i \in [n]$, let $g^{(i)}$ denote the alternative corresponding to $\beta_j^{(g)} = \mathbb{I}(i = j)$. Let $\mathcal{E}_{i,0}$ denote the event that Alg never samples in the interior of the interval $I_i^{2\epsilon}$, and that $\widehat{\beta}_i(\widehat{f}) = 0$, and let $\mathcal{E}_{i,1}$ denote the event that Alg never samples in the interior of the interval $I_i^{2\epsilon}$ and that $\widehat{\beta}_i(\widehat{f}) = 1$. Since $g^{(i)}$ and f coincide outside the interior of $I_i^{2\epsilon}$,

$$\mathbb{P}_{f,\mathsf{Alg}}[\mathcal{E}_{i,0}] = \mathbb{P}_{g^{(i)},\mathsf{Alg}}[\mathcal{E}_{i,0}] \quad \text{and} \quad \mathbb{P}_{f,\mathsf{Alg}}[\mathcal{E}_{i,1}] = \mathbb{P}_{g^{(i)},\mathsf{Alg}}[\mathcal{E}_{i,1}] \tag{24}$$

Moreover, since for all $g \in \mathcal{G}_{f,2\epsilon}$, $\widehat{\beta}(\widehat{f}) = \beta^{(g)}$ whenever $\|\widehat{f} - g\|_\infty \le \epsilon$, δ-correctness of Alg implies that

$$\mathbb{P}_{g^{(i)},\mathsf{Alg}}[\mathcal{E}_{i,0}] \le \mathbb{P}_{g^{(i)},\mathsf{Alg}}[\widehat{\beta}_i = 0] \le \delta \quad \text{and} \quad \mathbb{P}_{f,\mathsf{Alg}}[\mathcal{E}_{i,1}] \le \mathbb{P}_{f,\mathsf{Alg}}[\widehat{\beta}_i = 1] \le \delta \tag{25}$$

Putting these together,

$$\begin{aligned} 2\delta &\overset{\text{Equation (25)}}{\ge} \mathbb{P}_{g^{(i)},\mathsf{Alg}}[\mathcal{E}_{i,0}] + \mathbb{P}_{f,\mathsf{Alg}}[\mathcal{E}_{i,1}] \\ &\overset{\text{Equation (24)}}{=} \mathbb{P}_{f,\mathsf{Alg}}[\mathcal{E}_{i,0}] + \mathbb{P}_{f,\mathsf{Alg}}[\mathcal{E}_{i,1}] \\ &= \mathbb{P}_{f,\mathsf{Alg}}[\mathcal{E}_{i,0} \cup \mathcal{E}_{i,1}] \end{aligned}$$

where the last equality follows since $\mathcal{E}_{i,0}$ and $\mathcal{E}_{i,1}$ are disjoint. Noting that $\mathcal{E}_{i,0} \cup \mathcal{E}_{i,1}$ is the event that Alg never queries in the interior of $I_i^{2\epsilon}$, we have that

$$\mathbb{E}[\#\text{queries in interior of } I_i^{2\epsilon}] \ge \mathbb{P}[\mathsf{Alg} \text{ queries at least once in interior of } I_i^{2\epsilon}] \ge 1 - 2\delta \ .$$

Summing up over all n of the intervals $I_i^{2\epsilon}$, and noting that their interiors are disjoint, we conclude

$$\mathbb{E}[T] = \sum_{i=1}^{n} \mathbb{E}[\#\text{queries in interior of } I_i^{2\epsilon}] \ge (1 - 2\delta)n \ .$$

E Proof of Lemma 2.4

Lemma 2.4 will follow readily once we establish that $\omega(f, \cdot, \cdot)$ is "smooth" in the following sense:

Lemma E.1 *Let $[x - t, x + t] \subset [0, 1]$, and suppose that $\Delta(f, x, t) \ge \epsilon$. Then,*

$$\omega(f, x + \tau, \epsilon(1 - \frac{|\tau|}{t})) \le t + |\tau| \ . \tag{26}$$

Proof For any $\tau = [-t, t]$, we have

$$\begin{aligned} \mathrm{Sec}[f, [x - t, x + t]](x + \tau) - f(x + \tau) &\overset{\text{(Lemma B.1)}}{\ge} \min\{\frac{t - \tau}{t}, \frac{t + \tau}{t}\} \cdot (\mathrm{Sec}[f, [x - t, x + t]](x) - f(x)) \\ &= \min\{1 - \frac{\tau}{t}, 1 + \frac{\tau}{t}\} \cdot \Delta(f, x, t) \\ &= (1 - \frac{|\tau|}{t}) \cdot \epsilon \end{aligned} \tag{27}$$

First suppose that $\tau \ge 0$, then

$$\begin{aligned} \Delta(f, x + \tau, t + \tau) &= \mathrm{Sec}[f, [x - t, x + t + 2\tau]](x + \tau) - f(x + \tau) \\ &\overset{\text{(Lemma B.3)}}{\ge} \mathrm{Sec}[f, [x - t, x + t]](x + \tau) - f(x + \tau) \\ &\overset{\text{(Equation (27))}}{\ge} \epsilon(1 - \frac{|\tau|}{t}). \end{aligned}$$

On the other hand, if $\tau \leq 0$ then similarly we have

$$\Delta(f, x+\tau, t+|\tau|) = \text{Sec}[f, [x-t+2\tau, x+t]](x+\tau) - f(x+\tau) \geq \epsilon(1 - \frac{|\tau|}{t}).$$

By definition, and combining the above pieces, we get

$$\omega(f, x+\tau, \epsilon(1-\frac{|\tau|}{t})) = \inf\{s : \Delta(f, x+\tau, s) \geq \epsilon(1-\frac{|\tau|}{t})\} \leq t + |\tau|.$$

■

Proof [Proof of Lemma 2.4] Fix some α. For $u \in [x, x+\alpha t]$, we have $1-\alpha \leq 1 - \frac{u-x}{t}$, thus

$$\omega(f, u, \epsilon(1-\alpha)) \overset{Lemma\ H.6}{\leq} \omega(f, u, \epsilon(1-\tfrac{u-x}{t}))$$

and

$$\begin{aligned}
\int_x^{x+t} \omega(f,u,\epsilon(1-\alpha))^{-1} du &\geq \int_x^{x+\alpha t} \omega(f,u,\epsilon(1-\tfrac{u-x}{t}))^{-1} du \\
&= \int_0^{\alpha t} \omega(f, x+\tau, \epsilon(1-\tfrac{\tau}{t}))^{-1} d\tau \\
&\geq \int_0^{\alpha t} (t+\tau)^{-1} d\tau \\
&\geq \alpha t \cdot (t+\alpha t)^{-1}
\end{aligned}$$

which proves one side of the integral. A similar argument holds for $u \in [x-\alpha t, x]$ since $1-\alpha \leq 1 - \frac{|u-x|}{t}$. ■

F Proof of Noisy Lower Bound, Theorem 3.1

We begin by restating Birge's inequality, which will be used throughout.

Lemma F.1 (Birge's Inequality) *Let $\mathbb{P}_0, \mathbb{P}_1, \ldots, \mathbb{P}_n$ denote a family of probability distributions on a space $(\Omega, \mathcal{F})$, and let $A_0, A_1, \ldots, A_n$ denote pairwise disjoint events. If $p := \min_i \mathbb{P}_i(A_i) \geq 1/(n+1)$, then*

$$\frac{1}{n}\sum_{i=1}^{n} \text{KL}(\mathbb{P}_i; \mathbb{P}_0) \geq \text{kl}(p, \frac{1-p}{n}). \tag{28}$$

We also again employ the packing of Lemma 2.2 and Equation 4.

For $g, h \in \mathcal{G}_{f,2\epsilon}$, let $\mathbb{P}_g$ denote the law of $\{(X_s, Y_s)\}_{1 \leq s \leq T}$ induced by g and Alg, where the $Y_s \sim \mathcal{N}(g(X_s), \sigma^2)$. We also use $\mathbb{E}_g$ analogously (with dependence on Alg implicit). Let $\text{KL}(g(x), h(x))$ denote the KL between $\mathcal{N}(g(x), \sigma^2)$ and $\mathcal{N}(h(x), \sigma^2)$, which is equal to $(g(x)-h(x))^2/2\sigma^2$. For each $i \in [n]$, recall from Lemma 2.2 that $I_i = [m_i - \omega(m_i, f, 2\epsilon), m_i + \omega(m_i, f, 2\epsilon))]$ are disjoint. Let T_i denote the number of samples Alg collects from interval $\mathsf{int}(I_i)$. Finally, define the events $A_g := \{\|\hat{f} - g\|_\infty \leq \epsilon\}$. We then have

$$\|g - g'\|_\infty \geq 2\epsilon \implies A_g \cap A_{g'} = \emptyset \quad \text{and} \quad \mathbb{P}_g[A_g] \geq 1-\delta \quad \forall g \in \mathcal{G}_{2\epsilon, f}$$

where the first point follows from the triangle inequality and the second from (ϵ, δ)-correctness.

F.1 First part of theorem

For each $i \in [n]$, let $g^{(i)}$ denote the alternative corresponding to $\beta_j^{(g)} = \mathbb{I}(i = j)$. By construction, $g^{(i)}(m_i) - f(m_i) = \Delta(f, m_i, \omega(m_i, f, 2\epsilon)) = 2\epsilon$. Since $\mathrm{KL}(f(X_s), g^{(i)}(X_s)) = 0$ for $X_s \notin \mathrm{int}(I_i)$, we have

$$\begin{aligned}
\mathrm{KL}(\mathbb{P}_f, \mathbb{P}_{g^{(i)}}) = \mathbb{E}_f[\sum_{s=1}^{T} \mathrm{KL}(f(X_s), g^{(i)}(X_s))] &= \mathbb{E}_f[\sum_{s: X_s \in \mathrm{int}(I_i)} \mathrm{KL}(f(X_s), g^{(i)}(X_s))] \\
&\leq \mathbb{E}_f[T_i] \cdot \sup_{x \in I_i} \mathrm{KL}(f(x), g^{(i)}(x)) \\
&= \mathbb{E}_f[T_i] \cdot \frac{1}{2} \sup_{x \in I_i} (f(x) - g^{(i)}(x))^2 \\
&= \mathbb{E}_f[T_i] \cdot 2\epsilon^2/\sigma^2.
\end{aligned}$$

Birge's inequality with $n = 2$, $\mathbb{P}_1 = \mathbb{P}_f$ and $\mathbb{P}_0 = \mathbb{P}_{g^{(i)}}$, and $A_1 = A_f$ and $A_0 = A_{g^{(i)}}$ implies

$$\begin{aligned}
\frac{2\epsilon^2}{\sigma^2} \cdot \mathbb{E}_f[T_i] &\geq \mathbb{E}_f[\sum_{s=1}^{T} \mathrm{KL}(f(X_s), g^{(i)}(X_s))] \\
&\geq \mathrm{kl}(1 - \delta, \frac{\delta}{2})
\end{aligned}$$

We rearrange to get $\mathbb{E}_f[T_i] \geq \sigma^2 \mathrm{kl}(1 - \delta, \delta)/2\epsilon^2$, and summing over $i \in [n]$ obtains the claimed result.

F.2 Proof of second part

Recall the functions of $\mathcal{G}_{f,2\epsilon}$ defined in Equation 4. It will be convenient to introduce the notation $\beta \oplus i \in \{0,1\}^n$ to denote the vector which agrees with β except for flipping the i-th bit. As above, for any $\beta \neq \beta' \in \{0,1\}^n$,

$$\|g_\beta - g_{\beta'}\|_\infty = \max_{i \in [N]} \sup_{x \in I_i} |f(x) - g^{(i)}(x)| \geq 2\epsilon \tag{29}$$

In particular, $g_0 = f$ and since $g_\beta(x) = g_{\beta \oplus i}(x)$ for all $x \notin \mathrm{int}(I_i)$, following the arguments above yields that

$$\mathrm{KL}(\mathbb{P}_{g_{\beta \oplus i}}, \mathbb{P}_{g_\beta}) \leq \frac{2\epsilon^2}{\sigma^2} \mathbb{E}_{g_{\beta \oplus i}}[T_i]. \tag{30}$$

If we let $A_\beta := \{\|\hat{f} - g_\beta\| \leq \epsilon\}$, then since $\|g_\beta - g_{\beta \oplus i}\|_\infty \geq 2\epsilon$, the $\{A_\beta\}_{t \in \{0,1\}}$ are disjoint. By correctness, $\mathbb{P}_{g_\beta}[A_\beta] \geq 1 - \delta \geq 1/2 \geq 1/(n+1)$. Hence, Birge's inequality implies

$$\frac{2\epsilon^2}{n\sigma^2} \sum_{i=1}^{n} \mathbb{E}_{g_{\beta \oplus i}}[T_i] \geq \mathrm{kl}(1 - \delta, \delta/n) \quad \text{so that} \quad \sum_{i=1}^{n} \mathbb{E}_{g_{\beta \oplus i}}[T_i] \geq \frac{n\sigma^2}{2\epsilon^2} \cdot \mathrm{kl}(1 - \delta, \delta/n). \tag{31}$$

Next, notice that

$$\sum_{\beta \in \{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}_{g_{\beta \oplus i}}[T_i] = \sum_{i=1}^{n} \sum_{\beta \in \{0,1\}^n} \mathbb{E}_{g_{\beta \oplus i}}[T_i] = \sum_{i=1}^{n} \sum_{\beta \in \{0,1\}^n} \mathbb{E}_{g_\beta}[T_i] = \sum_{\beta \in \{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}_{g_\beta}[T_i].$$

Combining the above with Equation 31 implies that

$$\frac{1}{2^n} \sum_{\beta \in \{0,1\}^n} \sum_{i=1}^{n} \mathbb{E}_{g_\beta}[T_i] \geq \frac{n\sigma^2}{2\epsilon^2} \mathrm{kl}(1 - \delta, \delta/n) \tag{32}$$

and hence, since $T = \sum_{i=1}^{n} T_i$ we have

$$\sup_{g \in \mathcal{G}_{f,2\epsilon}} \mathbb{E}_g[T] \geq \sup_{\beta \in \{0,1\}^n} \mathbb{E}_{g_\beta}[T] \geq \frac{1}{2^n} \sum_{\beta \in \{0,1\}^n} \mathbb{E}_{g_\beta}[T] \geq \frac{n\sigma^2}{2\epsilon^2} \mathrm{kl}(1 - \delta, \delta/n). \tag{33}$$

G Proof of Noisy Upperbound, Theorem 3.2

For any function $\widetilde{f}$ and convex function f let $\widetilde{r}(x) = \widetilde{f}(x) - f(x)$. Note that on some interval $I \subseteq [0,1]$

$$\begin{aligned}
\text{Sec}[\widetilde{f}, I](x) - f(x) &= \text{Sec}[\widetilde{f}, I](x) - \text{Sec}[f, I](x) + \text{Sec}[f, I](x) - f(x) \\
&\leq \max\{\widetilde{r}(x_{l(I)}), \widetilde{r}(x_{r(I)})\} + 2\Delta(f, I) \\
&= \max\{\widetilde{r}(x_{l(I)}), \widetilde{r}(x_{r(I)})\} + 2(\Delta(\widetilde{f}, I) + \Delta(f, I) - \Delta(\widetilde{f}, I)) \\
&= \max\{\widetilde{r}(x_{l(I)}), \widetilde{r}(x_{r(I)})\} + 2(\Delta(\widetilde{f}, I) - \frac{\widetilde{r}(x_{l(I)}) + \widetilde{r}(x_{r(I)})}{2} + \widetilde{r}(x_{m(I)})) \\
&= 2\Delta(\widetilde{f}, I) - \min\{\widetilde{r}(x_{l(I)}), \widetilde{r}(x_{r(I)})\} + 2\widetilde{r}(x_{m(I)}))
\end{aligned}$$

and

$$\begin{aligned}
\text{Sec}[\widetilde{f}, I](x) - f(x) &= \text{Sec}[\widetilde{f}, I](x) - \text{Sec}[f, I](x) + \text{Sec}[f, I](x) - f(x) \\
&\geq \min\{\widetilde{r}(x_{l(I)}), \widetilde{r}(x_{r(I)})\}
\end{aligned}$$

yielding the claimed equation.

Define $\mathcal{X} = \{x : T(x) > 0\}$. As indicated just above, $\widetilde{\delta}$ is defined so that at any point during the algorithm we have $\sum_{x \in \mathcal{X}} \widetilde{\delta}(x) \leq \delta$. Thus, with probability at least $1 - \delta$, $|\widetilde{r}(x)| \leq \phi(T(x), \widetilde{\delta}(x))$ for all $x \in \mathcal{X}$, so in what follows assume these inequalities hold. Note that $\max\{\underline{B}(I^*, U), \Delta(\widetilde{f}, I^*) + \overline{B}(I^*, U)\} \leq \epsilon/2$ then we have $\sup_{x \in [0,1]} |\text{Sec}[f, \mathcal{T}](x) - f(x)| \leq \epsilon$ with probability at least $1 - \delta$.

Fix ϵ and assume t is the first time in which $\max\{\underline{B}(I^*, T, \widetilde{\delta}), \Delta(\widetilde{f}, I^*) + \overline{B}(I^*, T, \widetilde{\delta})\} \leq \epsilon$. Fix any $I \in \mathcal{L}(\mathcal{T})$.

Case 1: $I = I^*$ and $\underline{B}(I, T, \widetilde{\delta}) > \Delta(\widetilde{f}, I) + \overline{B}(I, T, \widetilde{\delta})$

This can only occur if $\underline{B}(I, T, \widetilde{\delta}) = \frac{1}{2} \max\left\{\phi(T(x_{l(I)}), \widetilde{\delta}(x_{l(I)})),\ \phi(T(x_{r(I)}), \widetilde{\delta}(x_{r(I)}))\right\} > \epsilon$. But if this occurs, the argument $x \in \{x_{l(I)}, x_{r(I)}\}$ that achieves the maximum is sampled, decreasing $\phi(T(x), \widetilde{\delta}(x))$. Thus, $T(x) \leq \min\{s : \phi(s, \widetilde{\delta}(x)) \leq \epsilon\}$ for $x \in \{x_{l(I)}, x_{r(I)}\}$.

Case 2: $I = I^*$ and $\underline{B}(I, T, \widetilde{\delta}) \leq \Delta(\widetilde{f}, I) + \overline{B}(I, T, \widetilde{\delta})$

Lines 5-9 insure that $\Delta(\widetilde{f}, I) \leq (1 + \beta)\overline{B}(I, T, \widetilde{\delta})$ so

$$\begin{aligned}
\Delta(\widetilde{f}, I) + \overline{B}(I, T, \widetilde{\delta}) &\leq (2 + \beta)\overline{B}(I, T, \widetilde{\delta}) \\
&= (2 + \beta)(\underline{B}(I, T, \widetilde{\delta}) + \phi(T(x_{m(I)}), \widetilde{\delta}(x_{m(I)})))
\end{aligned}$$

This implies that the number of times this case can occur while $\underline{B}(I, T, \widetilde{\delta}) > \phi(T(x_{m(I)}), \widetilde{\delta}(x_{m(I)}))$ is limited by the number of times $\underline{B}(I, T, \widetilde{\delta}) \geq \frac{\epsilon}{2(2+\beta)}$ because otherwise we would have $\Delta(\widetilde{f}, I) + \overline{B}(I, T, \widetilde{\delta}) \leq \epsilon$ by the above display. On the other hand, the same analogous argument holds for $\underline{B}(I, T, \widetilde{\delta}) \leq \phi(T(x_{m(I)}), \widetilde{\delta}(x_{m(I)}))$. Thus, $T(x) \leq \min\{s : \phi(s, \widetilde{\delta}(x)) \leq \frac{\epsilon}{2(2+\beta)}\}$ for $x \in \{x_{l(I)}, x_{m(I)}, x_{r(I)}\}$.

Considering both cases, we conclude that when the stopping condition is met, $T(x) \leq \min\{s : \phi(s, \widetilde{\delta}(x)) \leq \frac{\epsilon}{2(2+\beta)}\}$ for all $x \in \mathcal{X}$. Solving for this minimum s, we find $T(x) \leq (2 + \beta)^2 \sigma^2 \epsilon^{-2} \log(\log((2 + \beta)^2 \epsilon^{-2})/\widetilde{\delta}(x))$, ignoring constants. Since $\widetilde{\delta}(x)$ depends on $|\mathcal{X}|$, all that is left to do is bound the size of the tree.

Again, fix some $I \in \mathcal{L}(\mathcal{T})$ of the tree $\mathcal{T}$ formed when the stopping condition has been met. If I' is the parent of I then there exists some previous time such that

$$\begin{aligned}
(1 + \beta)\overline{B}(I', T, \widetilde{\delta}) &< \Delta(\widetilde{f}, I') \\
&\leq \Delta(f, I') + \overline{B}(I', T, \widetilde{\delta})
\end{aligned}$$

and

$$\begin{aligned}
\epsilon &< \Delta(\widetilde{f}, I') + \overline{B}(I', T, \widetilde{\delta}) \\
&\leq \Delta(f, I') + 2\overline{B}(I', T, \widetilde{\delta})
\end{aligned}$$

which together imply $\Delta(f, I') > \frac{\beta/2}{(2+\beta)}\epsilon$. This is remarkable because now the "largest-tree" argument of Section 2.3 follows identically, with the exception of $\epsilon/2 \mapsto \frac{\beta/2}{(2+\beta)}\epsilon$. Thus, we can immedietely apply Theorem 2.2 to bound the total number of points sampled by the algorithm, namely, $|\mathcal{X}| \leq \overline{N}(f, \frac{\beta}{(2+\beta)}\epsilon)$. Given an upper bound on $|\mathcal{X}|$ we obtain a lower bound: $\min_{x\in\mathcal{X}} \widetilde{\delta}(x) = \delta/2|\mathcal{X}|^2 \geq \delta/2N(f, \frac{\beta}{(2+\beta)}\epsilon)$. We showed above that the stopping condition is met if $\max_{x\in\mathcal{X}} \phi(T(x), \widetilde{\delta}(x)) \leq \frac{\epsilon}{2(2+\beta)}$. Using our lower bound on $\min_{x\in\mathcal{X}} \widetilde{\delta}(x)$ and inverting ϕ, we can apply the bound $T(x) \leq (2+\beta)^2\sigma^2\epsilon^{-2}\log(\overline{N}(f, \frac{\beta}{(2+\beta)}\epsilon)\log((2+\beta)^2\epsilon^{-2})/\delta)$ for all $x \in \mathcal{X}$. This completes the proof.

H Technical Properties of Λ, ω, and $\underline{N}$, t_{left}, and t_{right}

Proposition H.1 *Suppose that f is continuous. Then $\Lambda(f, \epsilon)$ is finite, and $t_{\text{left}}(f, \epsilon)$, $t_{\text{right}}(f, \epsilon)$, $\Lambda(f, \epsilon)$ and $\underline{N}(f, \epsilon)$ are all right-continuous in ϵ. Moreover, $\omega(f, x, \epsilon)$ is continuous in ϵ.*

Proposition H.2 (Restatement and Extension of Proposition 2.3) *For all $\lambda \geq 1$*

$$\Lambda(f, \epsilon) \in [1, \lambda]\left\{\Lambda(f, \lambda\epsilon) + \log\frac{t_{\text{right}}(f, \lambda\epsilon)t_{\text{left}}(f, \lambda\epsilon)}{t_{\text{right}}(f, \epsilon)t_{\text{left}}(f, \epsilon)})\right\} \tag{34}$$

Moreover,

$$\log(\frac{\omega_{\max}}{\omega_{\min}}) \leq \log(\|f\|_\infty/\epsilon) + \log(1/t_{\text{left}}(f, \epsilon)) + \log(1/t_{\text{right}}(f, \epsilon)) \tag{35}$$

and thus,

$$\underline{N}(f, \epsilon) \gtrsim \frac{\Lambda(f, \epsilon)}{\log(\|f\|_\infty/\epsilon) + \log(1/t_{\text{left}}(f, \epsilon)) + \log(1/t_{\text{right}}(f, \epsilon))} \tag{36}$$

H.1 Preliminaries

Lemma H.3 *Suppose that f is continuous. Then for all $t \in [0, \min\{x, 1-x\}]$, the function $\Delta(f, x, t)$ is convex, symmetric, continuous, non-negative and has $\Delta(f, x, 0) = 0$. Thus, satisfies $\Delta(f, x, ct) \leq c\Delta(f, x, t)$ for all $t \in [0, \min\{x, 1-x\}]$ and all $c \leq 1$. In particular, if $\Delta(f, x, t') > 0$, then $\Delta(f, x, t)$ is strictly increasing on $t \in [t', \min\{x, 1-x\}]$.*

Proof Writing $\Delta(f, x, t) = \frac{f(x+t)+f(x-t)}{2} + f(x)$, it is immediate that $\Delta(f, x, t)$ is symmetric in t. Being the sum of affine reparametrizations of a convex, continuous function f, $\Delta(f, x, t)$ is convex and continuous. By Jensen's inequality, $\frac{f(x+t)+f(x-t)}{2} \geq f(\frac{x+t}{2} + \frac{x-t}{2}) = f(x)$, so $\Delta(f, x, t) \geq 0$. Since $\Delta(f, x, 0) = 0$, convexity implies that 0 is a global minimimum of $t \mapsto \Delta(f, x, t)$, whence $\Delta(f, x, t)$ is non-decreasing for $t \geq 0$. ■

If we define $\epsilon^*(f, x) := \Delta(f, x, \min\{x, 1-x\})$, which represents the error of the midpoint approximation at x on interval I whose endpoints include either 0 or 1. The first lemma establishes that $\epsilon^*(f, x)$ does not grow as x approaches the endpoints of $[0, 1]$

Lemma H.4 *For any convex function $f : [0, 1] \to \mathbb{R}$, the functions $t \mapsto \Delta(f, x, t)$, $t \mapsto \Delta(f, x + t, t)$ and $t \mapsto \Delta(f, x - t, t)$ (defined on the approximate domains) are all non-decreasing in t. In particular, $\epsilon^*(f, x)$ is non-decreasing on $[0, 1/2]$ and non-increasing on $[1/2, 1]$*

Proof The function $t \mapsto \Delta(f, x, t)$ is addressed in Lemma H.3. Here, we will prove that $t \mapsto \Delta(f, x + t, t)$ is non-decreasing; that $t \mapsto \Delta(f, x - t, t)$ is non-decreasing will follow by a similar argument. Write $\Delta(f, x + t, t) = \frac{f(x)+f(x+2t)}{2} - f(x + t)$. Since $f(x + 2t)$ and $f(x + t)$ are convex in t, they admit right and left derivatives, which we denote by ∂_+ and ∂_-. Then for $\sigma \in \{+, -\}$, $\frac{\partial_\sigma}{\partial_t} f(x + t) = (\partial_\sigma f)(x + t)$, and $\frac{\partial_\sigma}{\partial_t}(f(x + 2t)) = 2(\partial_\sigma f)(x + 2t)$. Hence, $\Delta(f, x + t, t)$ is both-right and left-differentiable, which

$\frac{\partial_\sigma}{\partial_t}(\Delta(f,x+t,t)) = (\partial_\sigma f)(x+2t) - (\partial_\sigma f)(x+t)$. Since f is convex, both its right and left derivatives are non-decreasing in t. Hence $\frac{\partial_\sigma}{\partial_t}(\Delta(f,x+t,t)) \geq 0$, which implies that $\Delta(f,x+t,t)$ is non decreasing.

That $\epsilon^*(f,x)$ is non-decreasing on $[0,1/2]$ follows by consider $\Delta(f,t,t)$ and that it is non-increasing on $[1/2,1]$ follows by considering $\Delta(f,1-t,t)$. ■

Next, by using the sufficient decrease of $\Delta(f,x,t)$, we can express $\Delta(f,x,t)$ as the inverse of $\omega(f,x,\epsilon)$:

Lemma H.5 *For any $\epsilon \in (0,\epsilon^*(f,x)]$, $\omega(f,x,\epsilon)$ is equal to the unique t satisfying $\Delta(f,x,t) = \epsilon$. Moreover, $\Delta(f,x,\omega(f,x,0)] = 0$. In other words, for any $\epsilon \in [0,\epsilon^*(f,x)]$, $\omega(f,x,\epsilon)$ is given by the inverse of $\Delta(f,x,t)$ on $t \leq \min\{1-x,x\}$.*

Proof We may assume without loss of generality that $x \in [0,1/2]$. Fix $\epsilon \in (0,\epsilon^*(f,x)]$. Since $\epsilon^*(f,x) > 0$, and since $\lim_{t\to 0}\Delta(f,x,t) = 0$ and $\Delta(f,x,t)$ is continuous, there exists some $\epsilon' \in (0,\epsilon)$ and some t' such that $\Delta(f,x,t') = \epsilon'$. Since $\Delta(f,x,t') < \epsilon$ we have that $\omega(f,\epsilon,t) \geq t'$. Hence,

$$\omega(f,\epsilon,x) = \sup\{t \in [t', \min\{x,1-x\}] : \Delta(f,x,t) \leq \epsilon\} \tag{37}$$

Since $\Delta(f,x,t)$ is continuous and strictly increasing on $[t',\min\{u,1-u\}]$, we conclude that $\omega(f,\epsilon,x)$ is the unique t satisfying $\Delta(f,x,t) = \epsilon$. Moreover, $\Delta(f,x,\omega(f,x,0)) = 0$ follows from the definition of Δ and continuity of ω. ■

Lastly, Lemma H.3 also lets us relate $\omega(f,x,\epsilon)$ and $\omega(f,x,\epsilon')$ as follows.

Lemma H.6 *For any $\epsilon' \leq \epsilon$ and $x \in [t_{\text{left}}(f,\epsilon), 1-t_{\text{right}}(f,\epsilon)]$, we have*

$$\frac{\epsilon'}{\epsilon}\omega(f,x,\epsilon) \leq \omega(f,x,\epsilon') \leq \omega(f,x,\epsilon) \tag{38}$$

Proof To prove the first point, let $t = \omega(f,\epsilon,x)$ and $t' = \omega(f,\epsilon',x)$. It is clear from the definition of ω that $t \geq t'$. Thus, it suffices to show that $t' \geq \frac{\epsilon'}{\epsilon}t$. By the previous part, we have that $\Delta(f,x,t) = \epsilon$ and $\Delta(f,x,t') = \epsilon'$. Thus, by Lemma H.3, $\epsilon' = \Delta(f,x,t') \leq \frac{t'}{t}\Delta(f,x,t) = \frac{t'}{t}\epsilon$. Rearranging, $t' \geq \frac{\epsilon'}{\epsilon}t$. The second point now follows directly from the fact that convex functions ϕ satisfying $\phi(0) = 0$ satisfy $\phi(ct) \leq c\phi(t)$ for any $c \leq 1$ (Lemma B.2) ■

The second lemma verifies that, for continuous f, the key objects in the definition of $\Lambda(f,\epsilon)$ are bounded away from zero:

Lemma H.7 *If f is continuous, then for any $\epsilon > 0$, $t_{\text{left}}(f,\epsilon)$, $t_{\text{right}}(f,\epsilon)$, and $\inf_{x\in[t_{\text{left}}(f,\epsilon),1-t_{\text{right}}(f,\epsilon)]}\omega(f,x,\epsilon)$ are all strighlty postive.*

Proof If f is continuous, then $\lim_{x\to 0}\Delta(f,x,x) \downarrow 0$ and $\lim_{x\to 1}\Delta(f,x,1-x) \downarrow 0$, which implies that $t_{\text{left}}(f,\epsilon)$ and $t_{\text{right}}(f,\epsilon)$ are positive. Moroever, if f is continuous on $[0,1]$, it is uniformly continuous on $[0,1]$, which implies that there exists a t sufficiently small such that $|f(y)-f(x)| \leq \epsilon/2$ whenever $|y-x| \leq t$. We may assume without loss of generality that $t \leq \min\{t_{\text{right}}(f,\epsilon), t_{\text{left}}(f,\epsilon)\}$, so that $\Delta(f,I)[x-t,x+t]$ is well defined for all $x \in [t_{\text{left}}(f,\epsilon), 1-t_{\text{right}}(f,\epsilon)]$. Then, $\Delta(f,I)[x-t,x+t] \leq \frac{1}{2}|f(x+t)-f(x)| + \frac{1}{2}|f(x)-f(x-t)| \leq \epsilon/2$. Hence, $\omega(f,x,\epsilon) \geq t$ for all $x \in [t_{\text{left}}(f,\epsilon), 1-t_{\text{right}}(f,\epsilon)]$. ■

Corollary H.8 *For any $\lambda \geq 1$ and any $x_1 \in [t_{\text{left}}(f,\epsilon), t_{\text{left}}(f,\lambda\epsilon)]$*

$$x_1/\lambda \leq \omega(f,x_1,\epsilon) \leq x_1 \tag{39}$$

Analogous results hold for any $x_2 \in [1-t_{\text{right}}(f,\lambda\epsilon), 1-t_{\text{right}}(f,\epsilon)]$.

Proof Since $x_1 \geq t_{\text{left}}(f,\epsilon)$, we have $\omega(f,x_1,\epsilon) \leq x_1$ by definition. On the other hand, since $x \leq t_{\text{left}}(f,\lambda\epsilon)$, $\epsilon^*(f,x) \leq \lambda$. Thus, $\omega(f,x_1,\epsilon) = \omega(f,x_1,\epsilon^*(f,x)\cdot(\epsilon/\epsilon^*(f,x)) \geq \frac{\epsilon/\epsilon^*(f,x)}{\omega}(f,x_1,\epsilon^*(f,x)) \geq \lambda^{-1}\cdot\omega(f,x_1,\epsilon^*(f,x)) = x_1/\lambda$. ■

H.2 Proof of Proposition H.1

We begin following technical lemma, which relies on the technical preliminaries developed in the previous section.

Lemma H.9 *The maps $\epsilon \mapsto \omega(f,x,\epsilon)$ is continuous for $\epsilon \in [0,\epsilon^*(f,x)]$, and the maps $\epsilon \mapsto t_{\text{left}}(f,\epsilon)$ and $\epsilon \mapsto t_{\text{left}}(f,\epsilon)$ are all right-continuous in ϵ.*

Proof Since $\epsilon \mapsto \omega(f,x,\epsilon)$ on $[0,\epsilon^*(f,x)]$ is given by the inverse of a strictly increasing continuous function, it is continuous on $[0,\epsilon^*(f,x)]$. To see that is right continuous, $t_{\text{left}}(f,\epsilon)$, take a sequence $\epsilon_i \to \epsilon$. Since $t_{\text{left}}(f,\epsilon_i)$ is monotone, $t^* = \lim_{i\to\infty} t_{\text{left}}(f,\epsilon_i)$ exists, and satisfies $t^* \geq t_{\text{left}}(f,\epsilon)$. Since $\epsilon^*(f,x) = \Delta(f,x,x)$ is continuous in x and satisfies $\epsilon^*(f,t_{\text{left}}(\epsilon_i)) = \epsilon_i$, we also have $\Delta(f,t^*,t^*) = \epsilon$. Thus, if $t^* \neq t$, then we would have that $t^* > t_{\text{left}}(f,\epsilon)$ and $\Delta(f,t^*,t^*) = \epsilon$, contradicting the definition of $t_{\text{left}}(f,\epsilon)$. Right continuity of t_{right} follows analogously. ■

Next, we state a result which summarizes Lemma H.6 and Corollary H.8:

Corollary H.10 *Given $x \in \overline{I}(f,\epsilon)$ and $\lambda \geq 1$, let*

$$x_\lambda \in \overline{I}(f,\lambda\epsilon) = \begin{cases} x & x \in \overline{I}(f,\lambda\epsilon) \\ t_{\text{left}}(f,\lambda\epsilon) & x \in [t_{\text{left}}(f,\epsilon), t_{\text{left}}(f,\lambda\epsilon)] \\ 1 - t_{\text{right}}(f,\lambda\epsilon) & x \in [1-t_{\text{right}}(f,\lambda\epsilon), 1-t_{\text{right}}(f,\epsilon)] \end{cases} \tag{40}$$

Then,

$$1 \geq \frac{\omega(f,x,\epsilon)}{\omega(f,x_\lambda,\lambda\epsilon)} \geq \begin{cases} 1/\lambda & x \in \overline{I}(f,\lambda\epsilon) \\ \frac{t_{\text{left}}(f,\epsilon)}{t_{\text{left}}(f,\lambda\epsilon)\lambda} & x \in [t_{\text{left}}(f,\epsilon), t_{\text{left}}(f,\lambda\epsilon)] \\ \frac{t_{\text{right}}(f,\epsilon)}{t_{\text{right}}(f,\lambda\epsilon)\lambda} & x \in [1-t_{\text{right}}(f,\lambda\epsilon), 1-t_{\text{right}}(f,\epsilon)] \end{cases} \tag{41}$$

Proof When $x \in \overline{I}(f,\lambda\epsilon)$, we have by Lemma H.6

$$\frac{1}{\lambda}\omega(f,x,\lambda\epsilon) \leq \omega(f,x,\epsilon) \leq \omega(f,x,\lambda\epsilon)$$

In the second case, Corollary H.8 yields that

$$\lambda^{-1} t_{\text{left}}(x,f,\epsilon) \leq \frac{x}{\lambda} \leq \omega(f,x,\epsilon) \leq x \leq t_{\text{left}}(f,\lambda\epsilon)$$

The third case follow similarly. ■

As a direct consequence, we conclude that

$$\frac{1}{\lambda}\min\left\{1, \frac{t_{\text{left}}(f,\epsilon)}{t_{\text{left}}(f,\lambda\epsilon)}, \frac{t_{\text{right}}(f,\epsilon)}{t_{\text{right}}(f,\lambda\epsilon)}\right\} \leq \frac{\omega_{\max}(f,\epsilon)}{\omega_{\max}(f,\lambda\epsilon)}, \frac{\omega_{\min}(f,\epsilon)}{\omega_{\min}(f,\lambda\epsilon)} \leq 1 \tag{42}$$

so that, by right-continuity of $t_{\text{left}}(f,\epsilon)$ and $t_{\text{right}}(f,\epsilon)$, we have

Corollary H.11 *$\omega_{\min}(f,\epsilon)$ and $\omega_{\max}(f,\epsilon)$ are right-continuous.*

We can now prove Proposition H.1:

Proof $\Lambda(f,\epsilon)$ is finite since, by Lemma H.7 $\inf_{x\in[t_{\text{left}}(f,\epsilon),1-t_{\text{right}}(f,\epsilon)]}\omega(f,x,\epsilon) > 0$. To verify right-continuity, Proposition 2.3 implies that, for any $\lambda \geq 1$,

$$\frac{1}{\lambda}\Lambda(f,\epsilon) + \log\frac{t_{\text{left}}(f,\lambda\epsilon)t_{\text{right}}(f,\lambda\epsilon)}{t_{\text{left}}(f,\epsilon)t_{\text{right}}(f,\epsilon)} \leq \Lambda(f,\epsilon) \leq \left\{\Lambda(f,\lambda\epsilon) + \log\frac{t_{\text{left}}(f,\lambda\epsilon)t_{\text{right}}(f,\lambda\epsilon)}{t_{\text{left}}(f,\epsilon)t_{\text{right}}(f,\epsilon)}\right\}.$$

Thus, right-continuity follows as long as $\log\frac{t_{\text{left}}(f,\lambda\epsilon)t_{\text{right}}(f,\lambda\epsilon)}{t_{\text{left}}(f,\epsilon)t_{\text{right}}(f,\epsilon)}$ is right-continuous. By This follows since $\log(\cdot)$ is continuous on $(0,\infty)$, and and $t_{\text{left}}(f,\epsilon)$ and $t_{\text{right}}(f,\epsilon)$ are right-continuous (Lemma H.9) and strictly-positive (Lemma H.7).

To see that $\underline{N}(f,\epsilon)$ is right continuous, we express $\underline{N}(f,\epsilon)$ in terms of $\Lambda(f,\epsilon)$, $\omega_{\max}(f,\epsilon)$, $\omega_{\min}(f,\epsilon)$, the first of which is right-continuous by the above, and the second two of which are right-continuous by Corollary H.11 . ■

H.3 Proof of Proposition H.2

For the bound on $\Lambda(f,\epsilon)$, we have

$$\begin{aligned}\Lambda(f,\epsilon) &= \int_{t_{\text{left}}(f,\epsilon)}^{1-t_{\text{right}}(f,\epsilon)} \omega^{-1}(f,x,\epsilon) \\ = &\int_{t_{\text{left}}(f,\epsilon)}^{t_{\text{left}}(f,\lambda\epsilon)} \omega^{-1}(f,x,\epsilon)dx + \int_{t_{\text{left}}(f,\lambda\epsilon)}^{1-t_{\text{right}}(f,\lambda\epsilon)} \omega^{-1}(f,x,\epsilon)dx + \int_{1-t_{\text{right}}(f,\lambda\epsilon)}^{1-t_{\text{right}}(f,\epsilon)} \omega^{-1}(f,x,\epsilon)dx\end{aligned}$$

By Lemma H.6, we have

$$\int_{t_{\text{left}}(f,\lambda\epsilon)}^{1-t_{\text{right}}(f,\lambda\epsilon)} \omega^{-1}(f,x,\epsilon)dx \in [1,\lambda]\cdot\int_{t_{\text{left}}(f,\lambda\epsilon)}^{1-t_{\text{right}}(f,\lambda\epsilon)} \omega^{-1}(f,x,\lambda\epsilon)dx = \Lambda(f,\lambda\epsilon)$$

For $x \in [t_{\text{left}}(f,\epsilon), t_{\text{left}}(f,\lambda\epsilon)]$, Corollary H.8 implies

$$1/x \le \omega^{-1}(f,x,\epsilon) \le \frac{\lambda}{x}$$

So that

$$\int_{t_{\text{left}}(f,\epsilon)}^{t_{\text{left}}(f,\lambda\epsilon)} \omega^{-1}(f,x,\epsilon)dx \in [1,\lambda]\log(\frac{t_{\text{left}}(f,\lambda\epsilon)}{t_{\text{left}}(f,\epsilon)})$$

The case $x \in [1-t_{\text{right}}(f,\lambda\epsilon), 1-t_{\text{right}}(f,\epsilon)]$ similarly yields

$$\int_{1-t_{\text{right}}(f,\epsilon)}^{1-t_{\text{right}}(f,\lambda\epsilon)} \omega^{-1}(f,x,\epsilon)dx \in [1,\lambda]\log(\frac{t_{\text{right}}(f,\lambda\epsilon)}{t_{\text{right}}(f,\epsilon)})$$

Putting everything together,

$$\Lambda(f,\epsilon) \in [1,\lambda]\left\{\Lambda(f,\lambda\epsilon) + \log\frac{t_{\text{right}}(f,\lambda\epsilon)t_{\text{left}}(f,\lambda\epsilon)}{t_{\text{right}}(f,\epsilon)t_{\text{left}}(f,\epsilon)})\right\}$$

To control $\log(\omega_{\max}(f,\epsilon)/\omega_{\min}(f,\epsilon))$, we we the coarse bound

$$\log(\omega_{\max}(f,\epsilon)/\omega_{\min}(f,\epsilon)) \le \log(1/\omega_{\min}(f,\epsilon))$$

Thus, by Equation (42) with $\lambda = 2\|f\|_\infty/\epsilon$,

$$\begin{aligned}\omega_{\min}(f,\epsilon) &\ge \frac{\epsilon}{2\|f\|_\infty}\min\{\frac{t_{\text{left}}(f,\epsilon)}{t_{\text{left}}(f,2\|f\|_\infty)}, \frac{t_{\text{right}}(f,\epsilon)}{t_{\text{right}}(f,2\|f\|_\infty)}\}\omega_{\min}(f,\infty) \\ &\ge \frac{\epsilon t_{\text{right}}(f,\epsilon)t_{\text{left}}(f,\epsilon)}{2\|f\|_\infty t_{\text{left}}(f,\|f\|_\infty)t_{\text{right}}(f,\|\infty\|)}\omega_{\min}(f,\infty)\end{aligned}$$

Finally, since $\|\frac{f(0)+f(1)}{2} - f(1/2)\|_\infty \le 2\|f\|_\infty$, we have that $t_{\text{left}}(f,2\|f\|_\infty) = t_{\text{right}}(f,2\|\infty\|) = \omega_{\min}(f,2\|f\|_\infty) = 1/2$. Hence,

$$\omega_{\min}(f,\epsilon) \ge \frac{\epsilon\cdot t_{\text{right}}(f,\epsilon)\cdot t_{\text{left}}(f,\epsilon)}{\|f\|_\infty}\ . \tag{43}$$

Concluding, we find

$$\log(1/\omega_{\min}(f,\epsilon)) \le \log(\|f\|_\infty/\epsilon) + \log(1/t_{\text{right}}(f,\epsilon)) + \log(1/t_{\text{left}}(f,\epsilon))\ . \tag{44}$$