Worst-case Error Bounds for Online Learning of Smooth Functions

Weian (Andrew) Xie

February 14, 2025

Abstract

Online learning is a model of machine learning where the learner is trained on sequential feedback. We investigate worst-case error for the online learning of real functions that have certain smoothness constraints. Suppose that $\mathcal{F}_q$ is the class of all absolutely continuous functions $f : [0, 1] \to \mathbb{R}$ such that $\|f'\|_q \leq 1$, and $\mathrm{opt}_p(\mathcal{F}_q)$ is the best possible upper bound on the sum of the p^{th} powers of absolute prediction errors for any number of trials guaranteed by any learner. We show that for any $\delta, \epsilon \in (0, 1)$, $\mathrm{opt}_{1+\delta}(\mathcal{F}_{1+\epsilon}) = O(\min(\delta, \epsilon)^{-1})$. Combined with the previous results of Kimber and Long (Theoretical Computer Science, 1995) and Geneson and Zhou (Theoretical Computer Science, 2023), we achieve a complete characterization of the values of $p, q \geq 1$ that result in $\mathrm{opt}_p(\mathcal{F}_q)$ being finite, a problem open for nearly 30 years.

We study the learning scenarios of smooth functions that also belong to certain special families of functions, such as polynomials. We prove a conjecture by Geneson and Zhou (Theoretical Computer Science, 2023) that it is not any easier to learn a polynomial in $\mathcal{F}_q$ than it is to learn any general function in $\mathcal{F}_q$. We also define a noisy model for the online learning of smooth functions, where the learner may receive incorrect feedback up to $\eta \geq 1$ times, denoting the worst-case error bound as $\mathrm{opt}^{\mathrm{nf}}_{p,\eta}(\mathcal{F}_q)$. We prove that $\mathrm{opt}^{\mathrm{nf}}_{p,\eta}(\mathcal{F}_q)$ is finite if and only if $\mathrm{opt}_p(\mathcal{F}_q)$ is. Moreover, we prove for all $p, q \geq 2$ and $\eta \geq 1$ that $\mathrm{opt}^{\mathrm{nf}}_{p,\eta}(\mathcal{F}_q) = \Theta(\eta)$.

Keywords: online learning, smooth functions, noisy labels, single-variable functions, worst-case error

1 Introduction

Imagine a situation where a learner makes daily predictions for stock price values, given relevant inputs such as industry performance, market-wide economic trends, or recent company performance. The next day, the learner obtains some form of feedback on its previous prediction—receiving the actual price value, for example. The learner then uses its increased knowledge of the price value at certain inputs to generate better-informed future predictions—on the basis that similar inputs yield similar price values. The question of how fast the learner can use its accumulated information to generate better predictions in the worst case scenario arises naturally, and is the original motivation for the model of online learning previously studied in [1, 9, 12, 14, 11, 15].

Worst-case error bounds for online learning were first studied on functions over a discrete domain and with output values among a finite set, in [1, 12, 15]. The model over real-valued single-variable smooth functions $f : [0, 1] \to \mathbb{R}$ that we investigate was first introduced in [11], later studied in [14] and most recently studied [9]. The last paper also extended the problem to multi-variable smooth functions $f : [0, 1]^d \to \mathbb{R}$.

As our problem investigates how much accuracy the learner can guarantee in the worst-case scenario, we naturally represent the learning process as a game between the learner and an adversary, the latter of which is trying to force as much error as possible, with both players playing optimally. In the discrete model, different forms of feedback have been studied, including the standard model, where the adversary tells the learner the precise function output at the queried input; the bandit model, where the adversary only tells the learner YES or NO based on whether the learner's answer is correct; and a model where the adversary is allowed to give incorrect reinforcement up to η times, for some $\eta \geq 1$. The smooth function problem has so far only been studied in the context of the standard model, and all results so far pertain to this model. Indeed, the bandit model would not be interesting to study in the context of smooth functions, as the adversary can guarantee infinite error each time—through answering NO each time and simply setting the function $f \equiv C$ for some C sufficiently far away from all of the learner's guesses. However, a noisy model is interesting to study. For the noisy model in the smooth function setting, the adversary may provide erroneous output values to the learner.

1.1 Definitions

In the standard model of online learning, an algorithm A tries to learn a function from a given class of functions $\mathcal{F}$. The class $\mathcal{F}$ consists of functions $f : S \to \mathbb{R}$ for some fixed input set S. In the t^{th} trial, the algorithm is repeatedly given an input x_t within S and queried on the value of $f(x_t)$, whereupon it outputs a prediction $\hat{y}_t$. Afterwards, the true value of $f(x_t)$ is revealed to A, and the raw error of the round $e_t = |\hat{y}_t - f(x_t)|$ is recorded. The total error function is calculated as the sum of the p^{th} powers of the raw errors, for some parameter $p \geq 1$. Specifically, as per the notation of [9], for a fixed learning algorithm A operating on a finite sequence of inputs $\sigma = (x_0, x_1, \ldots, x_m) \in S^{m+1}$, and function $f \in \mathcal{F}$, this sum is

$$\mathscr{L}_p(A, f, \sigma) = \sum_{t=1}^{m} e_t^p = \sum_{t=1}^{m} |\hat{y}_t - f(x_t)|^p.$$

Subsequently, we define the worst-case scenario learning error for a fixed algorithm A over a class $\mathcal{F}$, as

$$\mathscr{L}_p(A, \mathcal{F}) = \sup_{f \in \mathcal{F}, \sigma \in \bigcup_{m \in \mathbb{Z}^+} S^m} \mathscr{L}_p(A, f, \sigma).$$

We then define the worst-case error for the best possible learning algorithm:

$$\mathrm{opt}_p(\mathcal{F}) = \inf_A \mathscr{L}_p(A, \mathcal{F}).$$

In [11], [14], and [9], the family $\mathcal{F}$ studied was the class of absolutely continuous single-variable functions $f : [0,1] \to \mathbb{R}$ that satisfy certain smoothness constraints, in the form of their derivatives having bounded norms. Namely, for all $q \geq 1$, define $\mathcal{F}_q$ to be the class of absolutely continuous functions $f : [0,1] \to \mathbb{R}$ such that $\int_0^1 |f'(x)|^q \leq 1$. Naturally, an extension of this definition is $\mathcal{F}_\infty$, denoting the class of absolutely continuous functions $f : [0,1] \to \mathbb{R}$ such that $\sup_{x \in (0,1)} |f'(x)| \leq 1$. As noted in [9] and [14], $\mathcal{F}_\infty$ contains precisely the functions $f : [0,1] \to \mathbb{R}$ that satisfy $|f(x) - f(y)| < |x - y|$ for all $x, y \in [0,1]$. The paper [9] notes that for any $q \geq 1$, the range of any function $f \in \mathcal{F}_q$ is at most 1. Furthermore, [9] also notes that $\mathcal{F}_\infty \subseteq \mathcal{F}_q \subseteq \mathcal{F}_r$ for all $1 \leq r \leq q$ by Jensen's Inequality, from which it follows that $\mathrm{opt}_p(\mathcal{F}_\infty) \leq \mathrm{opt}_p(\mathcal{F}_q) \leq \mathrm{opt}_p(\mathcal{F}_r)$ for all $1 \leq r \leq q$.

We also carry over some notation from [11], [14], and [9]. Let the q-action of a function $f : [0,1] \to \mathbb{R}$, denoted by $J_q[f]$, be defined as

$$J_q[f] = \int_0^1 |f'(x)|^q \mathrm{d}x.$$

As such, $\mathcal{F}_q$ is the set of functions whose q-action is less than or equal to 1. Furthermore, given a set $S = \{(u_i, v_i) : 1 \leq i \leq m\}$, where each $(u_i, v_i) \in [0,1] \times \mathbb{R}$ such that $u_i < u_{i+1}$ for all $1 \leq i \leq m-1$, define $f_S : [0,1] \to \mathbb{R}$ such that $f_\emptyset \equiv 0$ and

$$f_S(x) = \begin{cases} v_1 & x \leq u_1 \\ v_i + \frac{(x-u_i)(v_{i+1}-v_i)}{u_{i+1}-u_i} & x \in (u_i, u_{i+1}] \\ v_m & x > u_m \end{cases}$$

if $|S|$ is nonzero. Therefore, graphically, f_S is a continuous piecewise function composed of various line segments.

We now state two facts regarding f_S that will come into use later. Given a set $S = \{(u_i, v_i) : 1 \leq i \leq m\}$, with $(u_i, v_i) \in [0,1] \times \mathbb{R}$ for each $1 \leq i \leq m$ and $u_1 < \ldots < u_m$, one useful fact about f_S is that

$$J_1[f_S] = \sum_{i=1}^{m-1} \left(\int_{u_i}^{u_{i+1}} \left| \frac{v_{i+1} - v_i}{u_{i+1} - u_i} \right| \mathrm{d}x \right) = \sum_{i=1}^{m-1} |v_{i+1} - v_i|.$$

Another useful fact about f_S is that it has the minimum q-action out of all functions f passing through all points in S:

Lemma 1.1 ([11], [9])**.** *Let $S = \{(u_1, v_1), \ldots, (u_m, v_m)\}$ be a set of m points with $(u_i, v_i) \in [0,1] \times \mathbb{R}$ for each i, such that $u_1 < \ldots < u_m$. Then, for any $q \geq 1$ and any absolutely continuous function $f : [0,1] \to \mathbb{R}$ with $f(u_i) = v_i$ for $1 \leq i \leq m$, we have $J_q[f] \geq J_q[f_S]$.*

As per [11], we define the learning algorithm LININT using f_S. In particular, on trial 0, LININT's prediction is $\hat{y}_0 = \mathrm{LININT}(\emptyset, x_1) = 0$, and in any subsequent trial $t > 0$, its prediction is

$$\hat{y}_t = \mathrm{LININT}((x_0, f(x_0)), \ldots, (x_{t-1}, f(x_{t-1})), x_t) = f_{\{(x_0, f(x_0)), \ldots, (x_{t-1}, f(x_{t-1}))\}}(x_t).$$

Intuitively, this means that LININT makes a prediction either based on linear interpolation on the two closest surrounding points (one on each side of the input) which the algorithm knows the true value of, or simply based on its closest neighbor, if the requested input is to the left or to the right of all known points.

1.2 Smooth Functions in the Standard Model

The problem of determining the value of $\text{opt}_p(\mathcal{F}_q)$ across all $p, q \geq 1$ has been extensively explored in [11], [14], and [9]. Kimber and Long [11] proved that if we set $q = 1$, then for any $p \geq 1$, $\text{opt}_p(\mathcal{F}_1) = \infty$. That is, no matter what parameter p is chosen, the learner can never guarantee finite error when learning a function from $\mathcal{F}_1$. Furthermore, the same paper [11] also established that $\text{opt}_1(\mathcal{F}_\infty) = \infty$, from which it follows that for any $q \geq 1$, we have $\text{opt}_1(\mathcal{F}_q) = \infty$ as well. On the other hand, they proved that $\text{opt}_p(\mathcal{F}_q) = 1$ for all $p, q \geq 2$, from which it follows that $\text{opt}_p(\mathcal{F}_\infty) = 1$ for all $p \geq 2$ as well. This result was also proved in [6] with a different learning algorithm.

The paper [11] also established that $\text{opt}_{1+\epsilon}(\mathcal{F}_q) = O(\epsilon^{-1})$ for all $\epsilon \in (0, 1)$ and all $q \geq 2$. Geneson and Zhou [9] improved this bound and established a lower bound differing by a constant factor, proving that $\text{opt}_{1+\epsilon}(\mathcal{F}_\infty) = \Theta(\epsilon^{-\frac{1}{2}})$ and $\text{opt}_{1+\epsilon}(\mathcal{F}_q) = \Theta(\epsilon^{-\frac{1}{2}})$ for all $\epsilon \in (0, 1)$ and all $q \geq 2$. They also established that $\text{opt}_2(\mathcal{F}_{1+\epsilon}) = \Theta(\epsilon^{-1})$ for any $\epsilon \in (0, 1)$. From this, the upper bound $\text{opt}_p(\mathcal{F}_{1+\epsilon}) = O(\epsilon^{-1})$ for all $p \geq 2$ and $\epsilon \in (0, 1)$ follows. Furthermore, for any $q > 1$, they established that for sufficiently large p, the learner can guarantee an error of at most 1. Specifically, for any $q > 1$ and $p \geq 2 + \frac{1}{q-1}$, $\text{opt}_p(\mathcal{F}_q) = 1$.

As such, upper bounds for $\text{opt}_p(\mathcal{F}_q)$ have been established for all $p, q > 1$ whenever at least one of $p \geq 2$ and $q \geq 2$ holds. However, in the entirety of previous literature, no upper bounds for any instance of $p, q \in (1, 2)$ have been proved. In fact, there was no proof of finiteness for $\text{opt}_p(\mathcal{F}_q)$ for any choice of $p, q \in (1, 2)$. In this paper, we establish an upper bound on $\text{opt}_p(\mathcal{F}_q)$ for all values of $p, q \in (1, 2)$.

Theorem 1.2. *For $\delta, \epsilon \in (0, 1)$, we have $\text{opt}_{1+\delta}(\mathcal{F}_{1+\epsilon}) = O(\min(\delta, \epsilon)^{-1})$.*

As a corollary, we have the following result, which was also a conjecture by Geneson and Zhou [9].

Theorem 1.3. *For all $p > 1$ and $q > 1$, $\text{opt}_p(\mathcal{F}_q)$ is finite.*

Combined with the results from [11] that $\text{opt}_p(\mathcal{F}_q) = \infty$ whenever either $p = 1$ or $q = 1$, we now achieve a complete characterization of the values of (p, q) for which $\text{opt}_p(\mathcal{F}_q)$ is finite (i.e. precisely when $p, q > 1$), answering a question that has been open since the model of the online learning was first defined for smooth functions in 1995 by Kimber and Long [11].

1.3 Special Families of Smooth Functions

Geneson and Zhou [9] first explored the online learning of special families of smooth functions with further imposed restrictions. They suggested studying families such as polynomials, sums of exponential functions, sums of trigonometric functions, and piecewise combinations of these functions. Placing extra restrictions on smooth functions is natural as most functions modeling real-life phenomena belong to families that have additional properties beyond merely than the smoothness constraints that were extensively studied. As such, this direction can provide new results that are more specifically applicable to real-life learning scenarios.

Carrying over the notation from [9], let $\mathcal{P}_q \subseteq \mathcal{F}_q$ denote the family of polynomials f such that $f \in \mathcal{F}_q$. We prove a conjecture from [9] in the affirmative, establishing that given a fixed q-action restriction on a smooth function, having the extra restriction of the function being a polynomial does not decrease the learner's error.

Theorem 1.4. *For all $p > 0$ and $q \geq 1$, we have $\text{opt}_p(\mathcal{P}_q) = \text{opt}_p(\mathcal{F}_q)$.*

1.4 Smooth Functions with Noisy Feedback

The online learning of functions with noisy feedback has previously been studied in the discrete setting for classifiers, in [5, 2, 7], and we extend it here to the context of smooth functions as well. This model would be more applicable than the previously studied standard model. Suppose that a certain phenomenon to be learned cannot be perfectly represented by any function within a class, and the goal is to simply find a function from the class that can model the phenomenon as well as possible. As such, these inaccuracies correspond to the lies made by the adversary in our model.

Suppose that in the same format as in the standard model, an algorithm A attempts to learn a real-valued function $f : [0,1] \to \mathbb{R}$ from a certain class $\mathcal{F}$. In the t^{th} trial, the algorithm is queried on the value of f at an input x_t, whereupon it outputs a prediction $\hat{y}_t$ and the adversary subsequently reveals the actual value of $f(x_t)$. However, for a fixed positive integer $\eta \geq 1$, the adversary is allowed to lie about the value of $f(x_t)$ for up to η trials t. We define the error function similarly to our definition in the standard case, as the sum of the p^{th} powers of the raw errors of each trial $e_t = |\hat{y}_t - f(x_t)|$. However, as the real value of the function at x_t, $f(x_t)$, may differ from the adversary's reported value, the learner may possibly not know the exact value of the error function. Regardless, we are interested in how much accuracy the learner can guarantee in the worst-case scenario in the presence of possible lies from the adversary.

In the standard model of the learning of smooth functions, the error made by the learner on the first trial, e_0, is not counted in the error function, as otherwise the adversary can simply set the function f identically equal to some constant C sufficiently far away from the learner's prediction to generate an arbitrarily large error. As such, the learner needs to know the value of the function for at least one input to guarantee finite error on its first prediction. In this noisy version, where the learner is allowed to lie up to η times, getting feedback at one input is not sufficient to guarantee finite error on the next prediction for the learner, as the adversary can simply lie. There is no point in studying the worst-case error of the learner when the adversary can always force infinite error for the learner on its first counted trial. As such, it is necessary that we give the learner more initial rounds of feedback. We establish the following result about the number of initial rounds of adversary feedback necessary for the learner to guarantee finite error on its first real prediction.

Theorem 1.5. *For any integer $\eta \geq 1$, if incorrect feedback can be given up to η times, then at least $2\eta + 1$ initial rounds must be thrown out for the learner to guarantee finite error on its first prediction that counts toward the error evaluation.*

Accordingly, let the error function calculate the sum of the p^{th} powers of the raw errors starting from the $2\eta+2^{\text{nd}}$ trial. In particular, for a fixed learning algorithm A operating on a finite sequence of inputs $\sigma = (x_0, x_1, \ldots, x_m) \in [0,1]^{m+1}$ and fixed function $f \in \mathcal{F}$, define the error function for the noisy model to be

$$\mathscr{L}^{\text{ag}}_{p,\eta}(A, f, \sigma) = \sum_{t=2\eta+1}^{m} e_t^p = \sum_{t=2\eta+1}^{m} |\hat{y}_t - f(x_t)|^p.$$

Similar to the definition of the standard model, we denote the worst-case error for a fixed algorithm A over a class $\mathcal{F}$ of functions $f : [0,1] \to \mathbb{R}$ as

$$\mathscr{L}^{\text{ag}}_{p,\eta}(A, \mathcal{F}) = \sup_{f \in \mathcal{F}, \sigma \in \bigcup_{m \in \mathbb{Z}^+} [0,1]^m} \mathscr{L}^{\text{ag}}_{p,\eta}(A, f, \sigma).$$

Finally, we define the worst-case error for the best possible learning algorithm over a class $\mathcal{F}$, where the adversary is allowed to lie up to η times:

$$\text{opt}^{\text{ag}}_{p,\eta}(\mathcal{F}) = \inf_{A} \mathscr{L}^{\text{ag}}_{p,\eta}(A, \mathcal{F}).$$

We first establish a characterization for when $\text{opt}^{\text{ag}}_{p,\eta}(\mathcal{F}_q)$ is finite.

Theorem 1.6. *For any integer $\eta \geq 1$, the value of $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q)$ is finite if and only if $p > 1$ and $q > 1$.*

For $p, q \geq 2$ and any $\eta \geq 1$, we show that the noisy worst-case error is precisely on the order of η.

Theorem 1.7. *For any $\eta \geq 1$, $p, q \geq 2$ we have $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q) = \Theta(\eta)$.*

1.5 Order of Results

In Section 2, we work with the standard single-variable model, proving Theorem 1.2 and its corollary Theorem 1.3. In Section 3, we study the online learning of smooth polynomials, proving Theorem 1.4. In Section 4, we focus on the noisy learning scenario, defining its setup and establishing some preliminary bounds on $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q)$. In Section 5, we discuss open problems, conjectures, and future directions of research.

2 An Upper Bound on $\text{opt}_p(\mathcal{F}_q)$ for $p, q \in (1, 2)$

In this section, we prove that for all $\delta, \epsilon \in (0, 1)$, we have $\text{opt}_{1+\delta}(\mathcal{F}_{1+\epsilon}) = O(\min(\delta, \epsilon)^{-1})$. To prove this upper bound, we will use the LININT learning algorithm first defined by [11], which has previously been used to prove various upper bounds on $\text{opt}_p(\mathcal{F}_q)$ in [11], [14], and [9]. Specifically, we will first prove that $\mathscr{L}_{1+\epsilon}(\text{LININT}, \mathcal{F}_{1+\epsilon}) = O(\epsilon^{-1})$ for $\epsilon \in (0, 1)$. We will then use this to bound $\text{opt}_{1+\delta}(\mathcal{F}_{1+\epsilon})$ for all $\epsilon, \delta \in (0, 1)$.

The main idea for the first part of our proof that $\mathscr{L}_{1+\epsilon}(\text{LININT}, \mathcal{F}_{1+\epsilon}) = O(\epsilon^{-1})$ is similar to the main idea of previous proofs of upper bounds on $\mathscr{L}_p(\text{LININT}, \mathcal{F}_q)$. Specifically, we will compare changes in $J_{1+\epsilon}[f_S]$ as new points are added to the input-output set S to powers of $|y - f_S(x)|$, the raw absolute errors generated by the corresponding rounds, culminating in Lemma 2.3. This approach is similar to that of Lemma 10 in [11] and Lemma 2.10 in [9]. The main difference is that we will use novel inequality tactics, such as the binomial series expansion, to prove stronger inequalities than the inequalities previously proved. In the second part of our proof, we establish Lemma 2.6, a novel inequality that holds for an arbitrary number of rounds. Our result then follows upon combining Lemmas 2.3 and 2.6.

From this point onwards, we assume without loss of generality that the learning algorithm is never queried on the same input more than once, as the algorithm can always guarantee zero error upon being queried on the same input the second time. First, we prove a modification of the inequality that is Corollary 2.9 from Geneson and Zhou [9]. In particular, we specify that $0 < a \leq b < 1$ and expand the domain of x that the inequality works for from all $x \notin (a, b)$ to all $x \notin (-a, a)$, at the cost of weakening the inequality by a constant factor.

Lemma 2.1. *For reals $0 < a \leq b < 1$ such that $a + b \leq 1$, $q \in (1, 2)$, and any $|x| \geq a$, we have*

$$a\left|\frac{x}{a} + 1\right|^q + b\left|\frac{x}{b} - 1\right|^q - (a + b) \geq \frac{(q-1)|x|^q}{3}.$$

Proof. Note that for all $x \in (-\infty, -a] \cup [b, \infty)$, the inequality directly follows from Corollary 2.9 from Geneson and Zhou [9]. Therefore, we only have to deal with the case where $x \in [a, b)$. For

fixed a and b, let

$$\begin{aligned} f(x) &= a\left|\frac{x}{a}+1\right|^q + b\left|\frac{x}{b}-1\right|^q - (a+b) - \frac{(q-1)|x|^q}{3} \\ &= a\left(\frac{x}{a}+1\right)^q + b\left(1-\frac{x}{b}\right)^q - (a+b) - \frac{(q-1)x^q}{3} \end{aligned}$$

when $x \in [a,b]$. We prove that $f'(x) > 0$ for all $x \geq a$, so that it suffices to check that $f(a) \geq 0$. Indeed,

$$\begin{aligned} f'(x) &= q\left(\frac{x}{a}+1\right)^{q-1} - q\left(1-\frac{x}{b}\right)^{q-1} - \frac{q(q-1)x^{q-1}}{3} \\ &\geq q\cdot 2^{q-1} - q - \frac{q(q-1)}{3} = q\left(2^{q-1} - 1 - \frac{q-1}{3}\right) > 0 \end{aligned}$$

for $q \in (1,2)$, when $a \leq x \leq b$. Consider the function $g(x) = b\left(1-\frac{x}{b}\right)^q$. Note that $g'(x) = -q\left(1-\frac{x}{b}\right)^{q-1} \geq -q$ when $x \in [0,b]$. As $g(0) = b$, we have $g(x) \geq b - qx$ for all $x \in [0,b]$. As such, $b\left(1-\frac{a}{b}\right)^q = g(a) \geq b - qa$. Now, we have

$$\begin{aligned} f(a) &= a\cdot 2^q + b\left(1-\frac{a}{b}\right)^q - (a+b) - \frac{(q-1)a^q}{3} \geq a\cdot 2^q + (b-qa) - (a+b) - \frac{(q-1)a^q}{3} \\ &\geq a\cdot 2^q + (b-qa) - (a+b) - \frac{(q-1)a}{3} = a(2^q - q - 1) - \frac{(q-1)a}{3} \\ &\geq a\left(\frac{q-1}{3}\right) - \frac{(q-1)a}{3} = 0. \end{aligned}$$

Hence, we have that $f(x) \geq 0$ for all $x \in [a,b)$, as desired. ■

We now prove a modification of another inequality, Lemma 2.7, from [9]. Specifically, under the assumption that $a \leq b$, we will improve the factor of $a+b$ to a factor of a on the denominator of the right hand side, at the cost of a constant factor, which is a significant improvement from the bound in [9] when a is significantly smaller than b. However, we will restrict the domain from $x \in (-a,b)$ to $x \in (-a,a)$, given that we have already dealt with the case where $x \in [a,b)$ in Lemma 2.1.

Lemma 2.2. *For reals $0 < a \leq b$, $q \in (1,2)$, and $x \in (-a,a)$, we have*

$$a\left(1+\frac{x}{a}\right)^q + b\left(1-\frac{x}{b}\right)^q - (a+b) \geq \frac{q(q-1)}{3a}\cdot x^2.$$

Proof. We use the generalized binomial series expansion of $\left(1+\frac{x}{a}\right)^q$ and $\left(1-\frac{x}{b}\right)^q$ to rewrite the expression on the left hand side, which we will call Δ for convenience. Doing this is valid as by assumption $\left|\frac{x}{a}\right| < 1$ and $\left|\frac{x}{b}\right| < 1$, so the series expansions converge correctly. Then,

$$\begin{aligned} \Delta &= a\left(\sum_{k=0}^{\infty}\binom{q}{k}\left(\frac{x}{a}\right)^k\right) + b\left(\sum_{k=0}^{\infty}\binom{q}{k}\left(-\frac{x}{b}\right)^k\right) - (a+b) \\ &= a\left(\sum_{k=1}^{\infty}\binom{q}{k}\left(\frac{x}{a}\right)^k\right) + b\left(\sum_{k=1}^{\infty}\binom{q}{k}\left(-\frac{x}{b}\right)^k\right) \\ &= qx + a\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(\frac{x}{a}\right)^k\right) - qx + b\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(-\frac{x}{b}\right)^k\right) \\ &= a\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(\frac{x}{a}\right)^k\right) + b\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(-\frac{x}{b}\right)^k\right). \end{aligned}$$

We claim that $\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(\frac{x}{a}\right)^k\right) \geq \frac{q(q-1)}{3}\left(\frac{x}{a}\right)^2$. The inequality is trivial if $\frac{x}{a}=0$. If $\frac{x}{a}<0$, note that every term of the sum is nonnegative, because $\binom{q}{k}$ is negative precisely when $k\geq 2$ is odd, from the condition that $q\in(1,2)$, and $\left(\frac{x}{a}\right)^k$ is also negative precisely when k is odd. Therefore, we have $\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(\frac{x}{a}\right)^k\right) \geq \frac{q(q-1)}{2}\left(\frac{x}{a}\right)^2$ clearly. If $\frac{x}{a}>0$, note that $\binom{q}{k}\left(\frac{x}{a}\right)^k>0$ when $k\geq 2$ is even, and $\binom{q}{k}\left(\frac{x}{a}\right)^k<0$ when $k\geq 2$ is odd. Furthermore, for any $k\geq 2$, we have $\left|\binom{q}{k}\left(\frac{x}{a}\right)^k\right| > \left|\binom{q}{k+1}\left(\frac{x}{a}\right)^{k+1}\right|$ because $\left|\binom{q}{k+1}\right| = \left|\binom{q}{k}\cdot\frac{q-k}{k+1}\right| = \left|\binom{q}{k}\right|\cdot\left|\frac{k-q}{k+1}\right| < \left|\binom{q}{k}\right|$ when $q\in(1,2)$ and $\left|\left(\frac{x}{a}\right)^{k+1}\right| < \left|\left(\frac{x}{a}\right)^k\right|$. As such, because $\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(\frac{x}{a}\right)^k\right)$ is an alternating series with the magnitude of summands decreasing, the sum is bounded below by

$$\binom{q}{2}\left(\frac{x}{a}\right)^2+\binom{q}{3}\left(\frac{x}{a}\right)^3 > \left(\binom{q}{2}+\binom{q}{3}\right)\left(\frac{x}{a}\right)^2 \geq \frac{2}{3}\binom{q}{2}\left(\frac{x}{a}\right)^2 \geq \frac{q(q-1)}{3}\left(\frac{x}{a}\right)^2$$

when $q\in(1,2)$, as claimed.

Similarly, we can conclude that $\left(\sum_{k=2}^{\infty}\binom{q}{k}\left(-\frac{x}{b}\right)^k\right) \geq \frac{q(q-1)}{3}\left(-\frac{x}{b}\right)^2$, which is nonnnegative. Using these inequalities, we get the lower bound on our desired expression:

$$\Delta \geq a\left(\frac{q(q-1)}{3}\left(\frac{x}{a}\right)^2\right)+b\cdot 0 \geq \frac{q(q-1)}{3a}\cdot x^2.$$

■

Combining the above two inequalities, we make the following key claim, which is essentially a strengthened and more specific version of Lemma 2.10 in [9].

Lemma 2.3. *For a fixed $q\in(1,2)$, a nonempty set $S=\{(u_i,v_i):1\leq i\leq k\}$ of points in $[0,1]\times\mathbb{R}$ with $u_i<u_{i+1}$ for each $1\leq i\leq k-1$, and another point $(x,y)\in[0,1]\times\mathbb{R}$ such that $x\neq u_i$ for any i, we must have either*

$$J_q[f_{S\cup\{(x,y)\}}]-J_q[f_S]\geq\frac{q-1}{3}\cdot|y-f_S(x)|^q$$

or

$$J_q[f_{S\cup\{(x,y)\}}]-J_q[f_S]\geq\frac{(q-1)}{3|m|^{2-q}\cdot d}\cdot(y-f_S(x))^2,$$

where we let $d=\min_i|x-u_i|$ and $m=f'_S(x)$, the slope of the linear interpolation function at x.

Proof. First, if $x<u_1$, as established in the proof of Lemma 2.10 in the paper [9],

$$J_q[f_{S\cup\{(x,y)\}}]-J_q[f_S]=(u_1-x)\left|\frac{v_1-y}{u_1-x}\right|^q.$$

Because $|u_1-x|\leq 1$, the above expression is at least $|v_1-y|^q=|y-f_S(x)|^q\geq(q-1)|y-f_S(x)|^q$, hence satisfying the first inequality. The case of $x>u_k$ is similar.

Meanwhile, if $u_i<x<u_{i+1}$ for some integer $1\leq i\leq k-1$, substituting $a=x-u_i, b=u_{i+1}-x, c=y-f_S(x)$, and $m=f'_S(x)=\frac{v_{i+1}-v_i}{u_{i+1}-u_i}=\frac{f_S(x)-v_i}{a}=\frac{v_{i+1}-f_S(x)}{b}$, we have

$$J_q\left[f_{S\cup\{(x,y)\}}\right]-J_q[f_S]=|m|^q\left(a\left|1+\frac{c}{ma}\right|^q+b\left|1-\frac{c}{mb}\right|^q-(a+b)\right),$$

as established in the proof of Lemma 2.10 in [9]. Without loss of generality, assume that $a = |x - u_i| \le |x - u_{i+1}| = b$, as the other case follows from symmetry. Note that $a = d = \min_i |x - u_i|$. If $\frac{c}{m} \notin (-a, a)$, applying Lemma 2.1, we have

$$J_q\left[f_{S\cup\{(x,y)\}}\right] - J_q[f_S] \ge |m|^q \left(\frac{q-1}{3}\left|\frac{c}{m}\right|^q\right) = \frac{q-1}{3}\cdot|c|^q = \frac{q-1}{3}\cdot|y - f_S(x)|^q,$$

satisfying the first inequality. On the other hand, if $\frac{c}{m} \in (-a, a)$, applying Lemma 2.2, we have

$$\begin{aligned} J_q\left[f_{S\cup\{(x,y)\}}\right] - J_q[f_S] &\ge |m|^q \left(\frac{q(q-1)}{3a}\cdot\left|\frac{c}{m}\right|^2\right) \\ &\ge |m|^q \left(\frac{(q-1)}{3d}\cdot\left|\frac{c}{m}\right|^2\right) = \frac{q-1}{3|m|^{2-q}\cdot d}\cdot(y - f_S(x))^2. \end{aligned}$$

As a result, in the case where $\frac{c}{m} \in (-a, a)$, the second inequality is satisfied. Hence, in any case, either the first or the second inequality needs to hold. ■

Now, we proceed to the second part of our proof, where we prove another bound, Lemma 2.6, which we will combine at the end with Lemma 2.3 using Hölder's Inequality to prove our desired result. We first establish an inequality, which will be used in proving Lemma 2.6.

Lemma 2.4. *For all reals $p > 1$ and $x \ge 2$, we have*

$$x^p - (x-1)^{p-1}\cdot x \ge p - 1.$$

Proof. We consider two cases, dependent on whether $p \ge 2$ or $1 < p < 2$. If $p \ge 2$, then for all $x \ge 2$,

$$\begin{aligned} x^p - (x-1)^{p-1}\cdot x = x\left(x^{p-1} - (x-1)^{p-1}\right) &\ge x(2^{p-1} - 1^{p-1}) \\ &\ge 2\cdot(2^{p-1} - 1^{p-1}) = 2^p - 2 \ge p - 1, \end{aligned}$$

where the second step follows from the function $f(x) = x^{p-1} - (x-1)^{p-1}$ being non-decreasing for $x \ge 2$, which can be verified by differentiation. If $1 < p < 2$, then for all $x \ge 2$, we have

$$\begin{aligned} x^p - (x-1)^{p-1}\cdot x &= x^p\left(1 - \left(1 - \frac{1}{x}\right)^{p-1}\right) = x^p\left(1 - \left(\sum_{k=0}^{\infty}\binom{p-1}{k}\left(-\frac{1}{x}\right)^k\right)\right) \\ &= x^p\left(\sum_{k=1}^{\infty}(-1)\binom{p-1}{k}\left(-\frac{1}{x}\right)^k\right) = x^p\left(\sum_{k=1}^{\infty}(-1)^{k+1}\binom{p-1}{k}\left(\frac{1}{x}\right)^k\right) \end{aligned}$$

using the generalized binomial notation. The binomial series expansion in the second step is valid as $\left|\frac{1}{x}\right| < 1$. Furthermore, as $0 < p - 1 < 1$, $\binom{p-1}{k}$ is negative for all even positive integers k and positive for all odd positive integers k, so the expression $(-1)^{k+1}\binom{p-1}{k}\left(\frac{1}{x}\right)^k$ is positive for all positive integers k. As such, our expression is bounded below by

$$x^p\left(\binom{p-1}{1}\cdot\frac{1}{x}\right) = x^{p-1}(p-1) \ge p - 1$$

when $x \ge 2$. ■

Now, we make the following definition.

Definition 2.1. For a fixed $p \geq 1$ and a set $S = \{(u_i, v_i) : 1 \leq i \leq k\}$ of at least two points in $[0,1] \times \mathbb{R}$ with $u_i < u_{i+1}$ for each $1 \leq i \leq k-1$, define

$$H_{p,S} = \sum_{i=1}^{k-1} |v_{i+1} - v_i| \left(1 - |u_{i+1} - u_i|^{p-1}\right).$$

This seemingly contrived expression can be rewritten in various ways. For example, if we let $m_i = \frac{v_{i+1}-v_i}{u_{i+1}-u_i}$ be the slope of the line segment between u_i and u_{i+1} in f_S for all $1 \leq i \leq k-1$, we have that

$$H_{p,S} = \sum_{i=1}^{k-1} \left(|v_{i+1} - v_i| - |m_i||u_{i+1} - u_i|^p\right) = J_1[f_S] - \left(\sum_{i=1}^{k-1} |m_i||u_{i+1} - u_i|^p\right).$$

A key feature of $H_{p,S}$ which will be of later use is that whenever $f_S \in \mathcal{F}_1$, we have $0 \leq H_{p,S} \leq 1$. Indeed, when $J_1[f_S] \leq 1$, we have

$$H_{p,S} = J_1[f_S] - \left(\sum_{i=1}^{k-1} |m_i||u_{i+1} - u_i|^p\right) \leq J_1[f_S] \leq 1$$

and

$$H_{p,S} = J_1[f_S] - \left(\sum_{i=1}^{k-1} |m_i||u_{i+1} - u_i|^p\right) \geq J_1[f_S] - \left(\sum_{i=1}^{k-1} |m_i|\right) = 0.$$

We prove that $H_{p,S}$ never decreases when we add new points into set S, and prove a lower bound on the amount that $H_{p,S}$ increases by upon the addition of a new point into S in terms of p, the slope of f_S at the x-coordinate of the new point, and the closest x-coordinate difference between the new point and a point in S. Note that these are precisely the quantities involved in the second inequality in Lemma 2.3.

Lemma 2.5. *Consider a fixed real parameter $p \geq 1$, a set $S = \{(u_i, v_i) : 1 \leq i \leq k\}$ of at least two points in $[0,1] \times \mathbb{R}$ with $u_i < u_{i+1}$ for each $1 \leq i \leq k-1$, and another $(x, y) \in [0,1] \times \mathbb{R}$ with $x \neq u_i$ for any $1 \leq i \leq k$. If we let $d = \min_{1 \leq i \leq k} |x - u_i|$ and let $m = f_S'(x)$, which is the slope of the linear interpolation of S at x, then we have*

$$H_{p,S\cup(x,y)} - H_{p,S} \geq (p-1)|m|d^p.$$

Proof. First, if $x < u_1$, it suffices to prove that $H_{p,S\cup(x,y)} - H_{p,S} \geq 0$ as $m = 0$. Indeed, we have

$$H_{p,S\cup(x,y)} - H_{p,S} = |y - v_1|(1 - |x - u_1|^{p-1}) \geq 0.$$

The case where $x > u_k$ resolves similarly. Now, let $u_i < x < u_{i+1}$ for some $1 \leq i \leq k-1$. Without loss of generality assume that $v_i \leq v_{i+1}$. For convenience, define

$$\Delta_1 = \sum_{j=1}^{i-1} |v_{j+1} - v_j|(1 - |u_{j+1} - u_j|^{p-1}), \quad \Delta_2 = \sum_{j=i+1}^{k-1} |v_{j+1} - v_j|(1 - |u_{j+1} - u_j|^{p-1}).$$

First, we claim that we can reduce to only proving the case where $y \in [v_i, v_{i+1}]$. Indeed, note that if $y > v_{i+1}$, we can reduce it to the case where $y = v_{i+1}$, as

$$\begin{aligned} H_{p,S\cup(x,y)} &= \Delta_1 + \Delta_2 + \sum_{j=i}^{i+1} |y - v_j|(1 - |x - u_j|^{p-1}) \\ &\geq \Delta_1 + \Delta_2 + \sum_{j=i}^{i+1} |v_{i+1} - v_j|(1 - |x - u_j|^{p-1}) = H_{p,S\cup(x,v_{i+1})}, \end{aligned}$$

where the second step follows from the fact that $|y-v_i| \geq |v_{i+1}-v_i|$ and $|y-v_{i+1}| \geq 0 = |v_{i+1}-v_{i+1}|$. In the same way, the case where $y < v_i$ can be reduced to $y = v_i$. Assume from now on that $y \in [v_i, v_{i+1}]$. Also, assume that $d = |x-u_i| \leq |x-u_{i+1}|$. The case where x is nearer to u_{i+1} than u_i can be handled similarly. As $|y-v_i|+|y-v_{i+1}| = |v_{i+1}-v_i|$ and $1-|x-u_{i+1}|^{p-1} \leq 1-|x-u_i|^{p-1}$,

$$\begin{aligned} H_{p,S\cup(x,y)} &= \Delta_1 + \Delta_2 + \sum_{j=i}^{i+1} |y-v_j|(1-|x-u_j|^{p-1}) \\ &\geq \Delta_1 + \Delta_2 + (|y-v_i| + |y-v_{i+1}|)\left(1-|x-u_{i+1}|^{p-1}\right) \\ &= \Delta_1 + \Delta_2 + |v_{i+1}-v_i|\left(1-|x-u_{i+1}|^{p-1}\right) \\ &= \Delta_1 + \Delta_2 + \sum_{j=i}^{i+1} |v_i-v_j|(1-|x-u_j|^{p-1}) = H_{p,S\cup(x,v_i)}, \end{aligned}$$

so it suffices to check that $H_{p,S\cup(x,v_i)} - H_{p,S} \geq (p-1)|m|d^p$. Let $\frac{|u_i-u_{i+1}|}{|x-u_i|} = \frac{|u_i-u_{i+1}|}{d} = \lambda$. As such, we can substitute $|x-u_i|, |x-u_{i+1}|$, and $|u_i-u_{i+1}|$ with d, $\lambda d - d$, and λd, respectively. We have $\lambda \geq 2$, from x being closer to u_i than u_{i+1}. Note that as

$$\begin{aligned} H_{p,S\cup(x,v_i)} &= \Delta_1 + \Delta_2 + |v_{i+1}-v_i|\left(1-|x-u_{i+1}|^{p-1}\right) \\ &= \Delta_1 + \Delta_2 + \left|\frac{v_{i+1}-v_i}{u_{i+1}-u_i}\right|\left(|u_{i+1}-u_i| - |x-u_{i+1}|^{p-1}|u_{i+1}-u_i|\right) \\ &= \Delta_1 + \Delta_2 + |m|(\lambda d - (\lambda d - d)^{p-1}(\lambda d)) \end{aligned}$$

and

$$\begin{aligned} H_{p,S} &= \Delta_1 + \Delta_2 + |v_{i+1}-v_i|\left(1-|u_i-u_{i+1}|^{p-1}\right) \\ &= \Delta_1 + \Delta_2 + |m|\left(|u_i-u_{i+1}| - |u_i-u_{i+1}|^p\right) \\ &= \Delta_1 + \Delta_2 + |m|\left(\lambda d - (\lambda d)^p\right), \end{aligned}$$

we have that indeed

$$\begin{aligned} H_{p,S\cup(x,v_i)} - H_{p,S} &= |m|(\lambda d)^p - |m|(\lambda d - d)^{p-1}(\lambda d) \\ &= |m|d^p\left(\lambda^p - (\lambda-1)^{p-1}\cdot\lambda\right) \geq |m|d^p(p-1), \end{aligned}$$

where the last step follows from Lemma 2.4.

■

Now, we establish the following key inequality.

Lemma 2.6. *Choose a sequence $(u_1,v_1),(u_2,v_2),\ldots$ of points in $[0,1]\times\mathbb{R}$ with $u_i \neq u_j$ for any $i \neq j$, and define the set $S_n = \{(u_i,v_i) : 1 \leq i \leq n\}$ for every positive integer n. For each $i > 1$, let $d_i = \min_{j<i}|u_i-u_j|$ and let $m_i = f'_{S_{i-1}}(u_i)$, the slope of the linear interpolation of the first $i-1$ points of the sequence at u_i. If the inequality $J_1[f_{S_i}] \leq 1$, or equivalently $f_{S_i} \in \mathcal{F}_1$, holds for every positive integer i, then for any $p > 1$, we have*

$$\sum_{i=2}^{\infty} |m_i| d_i^p \leq \frac{1}{p-1}.$$

Proof. Indeed, we have

$$\sum_{i=2}^{\infty}(p-1)|m_i|d_i^p = (p-1)|m_i|d_i^p + \sum_{i=3}^{\infty}(p-1)|m_i|d_i^p$$
$$= \sum_{i=3}^{\infty}(p-1)|m_i|d_i^p \leq \sum_{i=3}^{\infty}\left(H_{p,S_i} - H_{p,S_{i-1}}\right) \leq 1.$$

Specifically, the second step follows from the fact $|m_2| = 0$, the third step follows from Lemma 2.5, and the last step follows from $0 \leq H_{p,S_i} \leq 1$ for any i, as $J_1[f_{S_i}] \leq 1$ for any i. ■

We now combine Lemma 2.6 with Lemma 2.3 to prove our desired upper bound on the error of the LININT learning algorithm when $(p,q) = (1+\epsilon, 1+\epsilon)$ for $\epsilon \in (0,1)$.

Theorem 2.7. *For $\epsilon \in (0,1)$, we have $\mathscr{L}_{1+\epsilon}(\text{LININT}, \mathcal{F}_{1+\epsilon}) = O(\epsilon^{-1})$ for $\epsilon \in (0,1)$.*

Proof. Let $1+\epsilon = q \in (1,2)$ and suppose that $f \in \mathcal{F}_q$ is the function to be learned by LININT. Let $\sigma = (x_0, x_1, \ldots, x_n)$ be the sequence of inputs, all different from each other, with $n \geq 1$. Similar to [9], we define $S_i = \{(x_0, f(x_0)), \ldots, (x_n, f(x_n))\}$ for each $0 \leq i \leq n$, and let $\hat{y}_i \in \mathbb{R}$ for every $1 \leq i \leq n$ be the prediction of LININT corresponding to each x_i. Define the raw error of each round $e_i = |\hat{y}_i - f(x_i)|$, for each $1 \leq i \leq n$. Also, let $d_i = \min_{j<i}|x_i - x_j|$ and $m_i = f'_{S_{i-1}}(x_i)$ for each i. By Lemma 2.3, for each round $1 \leq i \leq n$, we have either

$$J_q[f_{S_i}] - J_q[f_{S_{i-1}}] \geq \frac{q-1}{3} \cdot e_i^q$$

or

$$J_q[f_{S_i}] - J_q[f_{S_{i-1}}] \geq \frac{q-1}{3|m_i|^{2-q} \cdot d_i} \cdot e_i^2.$$

Let set $I_1 \subseteq \{1, 2, \ldots, n\}$ consist of all positive integers $1 \leq i \leq n$ such that e_i satisfy the first inequality. Analogously, let set $I_2 \subseteq \{1, 2, \ldots, n\}$ consist of all positive integers $1 \leq i \leq n$ such that e_i, m_i, d_i satisfy the second inequality. Note that $I_1 \cup I_2 = \{1, 2, \ldots n\}$.

By Lemma 1.1 and 2.3, we have

$$1 \geq J_q[f] \geq J_q[f_{S_n}] = \sum_{i=1}^{n}(J_q[f_{S_i}] - J_q[f_{S_{i-1}}]) \geq \sum_{i \in I_1}(J_q[f_{S_i}] - J_q[f_{S_{i-1}}]) \geq \frac{q-1}{3} \cdot \sum_{i \in I_1} e_i^q.$$

Dividing both sides, we get $\sum_{i \in I_1} e_i^q \leq \frac{3}{q-1}$. Similarly, we also have

$$1 \geq \sum_{i \in I_2}(J_q[f_{S_i}] - J_q[f_{S_{i-1}}]) \geq \frac{q-1}{3} \sum_{i \in I_2} \frac{e_i^2}{|m_i|^{2-q} \cdot d_i}.$$

Thus, $\sum_{i \in I_2} \frac{e_i^2}{|m_i|^{2-q} \cdot d_i} \leq \frac{3}{q-1}$. Now, note that by Hölder's Inequality,

$$\sum_{i \in I_2} e_i^q = \sum_{i \in I_2} \left(\frac{e_i^q}{|m_i|^{\frac{q(2-q)}{2}} \cdot d_i^{\frac{q}{2}}} \right) \cdot \left(|m_i|^{\frac{q(2-q)}{2}} \cdot d_i^{\frac{q}{2}} \right)$$
$$\leq \left(\sum_{i \in I_2} \frac{e_i^2}{|m_i|^{2-q} \cdot d_i} \right)^{\frac{q}{2}} \left(\sum_{i \in I_2} |m_i|^q d_i^{\frac{q}{2-q}} \right)^{\frac{2-q}{2}} \leq \left(\frac{3}{q-1} \right)^{\frac{q}{2}} \left(\sum_{i \in I_2} |m_i|^q d_i^{\frac{q}{2-q}} \right)^{\frac{2-q}{2}}.$$

Note that $|m_i| \cdot d_i^{\frac{1}{2-q}} \leq |m_i| \cdot d_i$, as $\frac{1}{2-q} > 1$. Furthermore, if $1 \leq j \leq i-1$ satisfies that $|x_i - x_j| = \min_{1 \leq k \leq i-1} |x_i - x_k|$, then

$$|m_i| \cdot d_i = \left| \frac{f_{S_{i-1}}(x_i) - f_{S_{i-1}}(x_j)}{x_i - x_j} \right| \cdot |x_i - x_j| = |f_{S_{i-1}}(x_i) - f_{S_{i-1}}(x_j)| \leq 1,$$

as the range of $f_{S_{i-1}}$ is at most 1 as $f_{S_{i-1}} \in \mathcal{F}_q$. As such, $|m_i| \cdot d_i^{\frac{1}{2-q}} \leq 1$ for each i, so

$$\begin{aligned}\sum_{i \in I_2} |m_i|^q d_i^{\frac{q}{2-q}} = \sum_{i \in I_2} \left(|m_i| d_i^{\frac{1}{2-q}} \right)^q \leq \sum_{i \in I_2} |m_i| d_i^{\frac{1}{2-q}} \leq |m_1| d_1^{\frac{1}{2-q}} + \sum_{i=2}^{n} |m_i| d_i^{\frac{1}{2-q}} \\ \leq 1 + \frac{1}{\frac{1}{2-q} - 1} = 1 + \frac{2-q}{q-1} = \frac{1}{q-1},\end{aligned}$$

where the fourth step follows from Lemma 2.6, which is valid because $f_{S_i} \in \mathcal{F}_q$ which is in $\mathcal{F}_1$ for all i. Therefore, we have

$$\sum_{i \in I_2} e_i^q \leq \left(\frac{3}{q-1} \right)^{\frac{q}{2}} \left(\frac{1}{q-1} \right)^{\frac{2-q}{2}} \leq \frac{3}{q-1}.$$

Putting everything together, we have

$$\mathscr{L}_q(\text{LININT}, f, \sigma) = \sum_{i=1}^{n} e_i^q \leq \sum_{i \in I_1} e_i^q + \sum_{i \in I_2} e_i^q \leq \frac{6}{q-1},$$

for any choice of $f \in \mathcal{F}_q$, any positive integer n, and any sequence of inputs $\sigma \in [0,1]^{n+1}$. As such, we have $\mathscr{L}_{1+\epsilon}(\text{LININT}, \mathcal{F}_{1+\epsilon}) \leq \frac{6}{\epsilon}$, for any $\epsilon \in (0,1)$. ■

Note that from Geneson and Zhou [9], we have the following inequality.

Lemma 2.8 ([9])**.** *For any $q > 1$ and $p' > p > 1$, we have $\mathscr{L}_{p'}(\text{LININT}, \mathcal{F}_q) \leq \mathscr{L}_p(\text{LININT}, \mathcal{F}_q)$.*

Combining this with Theorem 2.7, we can get the following.

Theorem 2.9. *For $0 < \epsilon \leq \delta < 1$, we have $\mathscr{L}_{1+\delta}(\text{LININT}, \mathcal{F}_{1+\epsilon}) = O(\epsilon^{-1})$ for $\epsilon \in (0,1)$.*

As such, we have the following upper bound.

Theorem 2.10. *For $0 < \epsilon \leq \delta < 1$, we have $\text{opt}_{1+\delta}(\mathcal{F}_{1+\epsilon}) = O(\epsilon^{-1})$.*

On the other hand, as $\mathcal{F}'_q \subseteq \mathcal{F}_q$ for all $q' > q \geq 1$ by Jensen's inequality, we have $\text{opt}_p(\mathcal{F}_{q'}) \leq \text{opt}_p(\mathcal{F}_q)$ for all $q' > q \geq 1$. Thus, for all $0 < \delta \leq \epsilon < 1$, we have $\text{opt}_{1+\delta}(\mathcal{F}_{1+\epsilon}) \leq \text{opt}_{1+\delta}(\mathcal{F}_{1+\delta}) = O(\delta^{-1})$. Combining this with Theorem 2.10, we have the following upper bound.

Theorem 2.11. *For $\delta, \epsilon \in (0,1)$, we have $\text{opt}_{1+\delta}(\mathcal{F}_{1+\epsilon}) = O(\min(\delta, \epsilon)^{-1})$.*

As a corollary, we have now completely classified the values of $p, q \geq 1$ such that $\text{opt}_p(\mathcal{F}_q)$ is finite, confirming a conjecture by [9].

Corollary 2.12. *The worst-case learning error $\text{opt}_p(\mathcal{F}_q)$ is finite if and only if $p, q > 1$.*

3 Online Learning of Polynomials

We prove the conjecture, raised by [9], that for all $p > 0$ and $q \geq 1$, $\text{opt}_p(\mathcal{P}_q) = \text{opt}_p(\mathcal{F}_q)$. Intuitively, this means that given a fixed q-action restriction on a smooth function, having the extra restriction of the function being a polynomial does not decrease the learner's error. The main idea of our proof is that the adversary can replicate any strategy it uses against the learner in the scenario of all smooth functions to the scenario of all smooth polynomials. As such, it can guarantee the same worst-case error in the latter scenario as in the former. Note that in this section, we are not concerned with the learner's strategy, whether it be LININT or another algorithm; we focus only on strategies used by the adversary against the learner, adapting based on whatever predictions the learner makes to maximize error for the learner.

Our proof invokes the following well-known result on the polynomial approximation of continuous real-valued functions.

Theorem 3.1 (Weierstrass Approximation Theorem)**.** *For any continuous real-valued function f defined on a real interval $[a, b]$ and any $\epsilon > 0$, there exists a polynomial P such that for all $x \in [a, b]$, $|f(x) - P(x)| < \epsilon$.*

Using the Weierstrass Approximation Theorem, we prove that given any set of known points S, the adversary can find a polynomial P that is arbitrarily close to the points in S and has q-action bounded above by ϵ more than the q-action of f_S for any $\epsilon > 0$.

Lemma 3.2. *Given any $q \geq 1$, $\epsilon > 0$, and a set $S = \{(u_i, v_i) : 1 \leq i \leq m\}$ of points in $[0, 1] \times \mathbb{R}$ such that $u_1 < \cdots < u_m$ and its corresponding linear interpolation function f_S, there exists a polynomial $P : [0, 1] \to \mathbb{R}$ such that $|P(u_i) - v_i| = |P(u_i) - f_S(u_i)| < \epsilon$ for all $1 \leq i \leq m$, and*

$$J_q[P] < J_q[f_S] + \epsilon.$$

Proof. Consider the derivative of f_S, which when considered as a graph, consists of disjoint horizontal line segments, broken off at each u_i, where it is not defined. We modify f'_S to make it continuous, so Theorem 3.1 can be used. Define $d_0, d_1, \ldots, d_m$ to be the slopes of the segments of f_S in order:

$$d_i = \begin{cases} 0 & i = 0 \\ \frac{v_{i+1} - v_i}{u_{i+1} - u_i} & 1 \leq i \leq m - 1 \\ 0 & i = m. \end{cases}$$

Now, for a sufficiently small positive $\epsilon_2 < \min_{1 \leq i < j \leq m} |u_i - u_j|$, define the function $F'_S : [0, 1] \to \mathbb{R}$ such that

$$F'_S(x) = \begin{cases} d_0 & x \leq u_1 - \epsilon_2 \\ d_{i-1} + \frac{(x - (u_i - \epsilon_2))(d_i - d_{i-1})}{\epsilon_2} & x \in (u_i - \epsilon_2, u_i] \\ d_i & x \in (u_i, u_{i+1} - \epsilon_2] \\ d_m & x > u_m. \end{cases}$$

It is easy to see that F'_S is continuous for all $x \in (0, 1)$. Indeed, intuitively, the graph of F'_S is essentially the graph of f'_S but with the disjoint horizontal segments "connected".

We claim that for sufficiently small ϵ_2, $\int_0^x |F'_S(t)|^q \mathrm{d}t$ gets arbitrarily close to $\int_0^x |f'_S(t)|^q \mathrm{d}t$ for any $x \in [0, 1]$ and any $q \geq 1$. Indeed, $F'_S(t)$ only differs from $f'_S(t)$ when $t \in (u_i - \epsilon_2, u_i]$ for some $1 \leq i \leq m$. For a fixed i, when t is in this range, $f'_S(t)$ is precisely d_{i-1}, whereas $F'_S(t)$ is between d_{i-1} and d_i. As such, $||f'_S(t)|^q - |F'_S(t)|^q|$ is bounded above by $c = \max_{1 \leq i < j \leq m} (|d_i|^q - |d_j|^q)$, which is finite, for any q. As such, for any $x \in [0, 1]$,

$$\left|\int_0^x |F_S'(t)|^q \mathrm{d}t - \int_0^x |f_S'(t)|^q \mathrm{d}t\right| \leq \int_0^x \left||f_S'(t)|^q - |F_S'(t)|^q\right| \mathrm{d}t$$
$$\leq \sum_{i=1}^m \int_{u_i-\epsilon_2}^{u_i} \left||f_S'(t)|^q - |F_S'(t)|^q\right| \mathrm{d}t$$
$$\leq \sum_{i=1}^m \int_{u_i-\epsilon_2}^{u_i} c \mathrm{d}t = mc\epsilon_2,$$

which gets arbitrarily close to 0 if we set ϵ_2 sufficiently small. We will use this inequality later on. Note that as a corollary, setting $x = 1$ above yields that $\int_0^1 |F_S'(t)|^q \mathrm{d}t$ gets arbitrarily close to $J_q[f_S]$ when we set ϵ_2 sufficiently small—we will also use this fact later on.

Now, for sufficiently small $\epsilon_3 > 0$ (which we will specify later), consider a polynomial Q such that for all $x \in [0,1]$, $|F_S'(x) - Q(x)| < \epsilon_3$, which is possible by Theorem 3.1 as F_S' is continuous for all $x \in [0,1]$. Subsequently, let $P(x) \equiv \int_0^x Q(t)\mathrm{d}t + v_1$, which is clearly a polynomial.

We claim that $J_q[P] < J_q[f_S] + \epsilon$ for any $\epsilon > 0$, given that our choices of ϵ_2 and ϵ_3 are sufficiently small. In particular, define $c_1 = \max_{x\in[0,1]} |F_S'(x)| = \max_{0\leq i\leq m} |d_i|$. Now, pick $\epsilon_2 < \frac{\epsilon}{2mc}$ and let ϵ_3 be sufficiently small such that $(c_1+\epsilon_3)^q - c_1^q < \frac{\epsilon}{2}$. Note that this implies that $(|F_S'(x)| + \epsilon_3)^q - (|F_S'(x)|)^q < \frac{\epsilon}{2}$ for all $x \in [0,1]$, as $0 \leq |F_S'(x)| \leq c_1$ and the function $(y+\epsilon_3)^q - y^q$ is increasing for all $y \geq 0$ and any positive ϵ.

As such, we have the upper bound

$$J_q[P] = \int_0^1 |Q(x)|^q \mathrm{d}x \leq \int_0^1 (|F_S'(x)| + \epsilon_3)^q \mathrm{d}x \leq \int_0^1 \left(|F_S'(x)|^q + \frac{\epsilon}{2}\right) \mathrm{d}x$$
$$= \int_0^1 |F_S'(x)|^q \mathrm{d}x + \frac{\epsilon}{2} \leq J_q[f_S] + \left(\int_0^1 |F_S'(x)|^q \mathrm{d}x - J_q[f_S]\right) + \frac{\epsilon}{2}$$
$$= J_q[f_S] + \left(\int_0^1 |F_S'(x)|^q \mathrm{d}x - \int_0^1 |f_S'(x)|^q \mathrm{d}x\right) + \frac{\epsilon}{2}$$
$$\leq J_q[f_S] + mc\epsilon_2 + \frac{\epsilon}{2} < J_q[f_S] + \frac{\epsilon}{2} + \frac{\epsilon}{2} = J_q[f_S] + \epsilon.$$

Furthermore, we also claim that $|P(u_i) - v_i| < \epsilon$ for any $\epsilon > 0$, given that our choices of ϵ_2 and ϵ_3 are sufficiently small. Indeed, as long as $\epsilon_3 < \frac{\epsilon}{2}$ and $\epsilon_2 < \frac{\epsilon}{2mc}$, we have for each $1 \leq i \leq m$ that

$$P(u_i) = \int_0^{u_i} Q(t)\mathrm{d}t + v_1 \leq \int_0^{u_i} \left(F_S'(t) + \epsilon_3\right) \mathrm{d}t + v_1 \leq \int_0^{u_i} \left(F_S'(t)\right) \mathrm{d}t + \epsilon_3 + v_1$$
$$\leq \int_0^{u_i} \left(f_S'(t)\right) \mathrm{d}t + mc\epsilon_2 + \epsilon_3 + v_1 = (v_i - v_1) + v_1 + mc\epsilon_2 + \epsilon_3$$
$$= v_i + mc\epsilon_2 + \epsilon_3 < v_i + \frac{\epsilon}{2} + \frac{\epsilon}{2} = v_i + \epsilon$$

and

$$P(u_i) = \int_0^{u_i} Q(t)\mathrm{d}t + v_1 \geq \int_0^{u_i} \left(F_S'(t) - \epsilon_3\right) \mathrm{d}t + v_1 \geq \int_0^{u_i} \left(F_S'(t)\right) \mathrm{d}t - \epsilon_3 + v_1$$
$$\geq \int_0^{u_i} \left(f_S'(t)\right) \mathrm{d}t - mc\epsilon_2 - \epsilon_3 + v_1 = (v_i - v_1) + v_1 - mc\epsilon_2 - \epsilon_3$$
$$= v_i - mc\epsilon_2 - \epsilon_3 < v_i - \frac{\epsilon}{2} - \frac{\epsilon}{2} = v_i - \epsilon.$$

Therefore, for any $\epsilon > 0$, we can find a polynomial P such that $|P(u_i) - v_i| < \epsilon$ for all $1 \leq i \leq m$ and $J_q[P] < J_q[f_S] + \epsilon$.

■

For a finite set of points $S \subseteq [0,1] \times \mathbb{R}$, we now prove a fact about constructing a function that passes through all points in S by taking a weighted average of a special set of $2^{|S|}$ functions that each do not necessarily pass through the points in S. We will later combine this with Lemma 3.2 to implicitly construct a polynomial passing through all points in S by taking a weighted average of $2^{|S|}$ polynomials that each approximate but do not necessarily pass through the points in S. We use the convention that $[m] := \{1, 2, \ldots, m\}$.

Lemma 3.3. *Suppose there is a set $S = \{(u_i, v_i) : 1 \leq i \leq m\}$ of m points in $[0,1] \times \mathbb{R}$ such that $u_1 < \ldots < u_m$. If we have 2^m functions $f_X : [0,1] \times \mathbb{R}$, corresponding to each subset $X \subseteq [m]$, such that for any $X \subseteq [m]$, it holds for any integer $1 \leq i \leq m$ with $i \in X$ that $f_X(u_i) > v_i$, and it holds for any integer $1 \leq i \leq m$ with $i \notin X$ that $f_X(u_i) < v_i$, then there exists a weighted average of the functions $f \equiv \sum_{X \subseteq [m]} w_X f_X$, with $0 \leq w_X \leq 1$ for all $X \subseteq [m]$ and $\sum_{X \subseteq [m]} w_X = 1$, such that $f(u_i) = v_i$ for all $1 \leq i \leq m$.*

Proof. We proceed by induction. For the base case of $m = 1$, we are given two functions $f_\emptyset, f_{\{1\}} : [0,1] \to \mathbb{R}$, and one point (u_1, v_1) that the desired weighted average should pass through. As $f_{\{1\}}(u_1) > v_1$ and $f_\emptyset(u_1) < v_1$, we can clearly assign two weights $0 \leq w_{\{1\}}, w_\emptyset \leq 1$ with sum 1 such that $w_{\{1\}} f_{\{1\}}(u_1) + w_\emptyset f_\emptyset = v_1$.

Proceeding with the inductive step, we assume that the result is true for $m-1$, and prove that it is true for m as well. Let the set of points be $S = \{(u_i, v_i) : 1 \leq i \leq m\}$, and functions be $f_X : [0,1] \times \mathbb{R}$ where $X \subseteq [m]$. Partition the 2^m functions into two groups, A and B, such that f_X goes into A if $m \in X$, and goes into B if $m \notin X$. Clearly, each of the two groups have 2^{m-1} functions. Note that for every subset $Y \subseteq [m-1]$, there exists a function in each of A and B such that the output at u_i is greater than v_i precisely when $i \in Y$, and the output at u_i is less than v_i precisely when $i \notin Y$ (specifically, the function $f_{Y \cup \{m\}}$ in A, and the function f_Y in B). By the inductive hypothesis, there exists a weighted average $F_1 \equiv \sum_{X \subseteq [m], m \in X} w_X f_X$ of the functions in A such that $F_1(u_i) = v_i$ for all $1 \leq i \leq m-1$. Similarly, there exists a weighted average $F_2 \equiv \sum_{X \subseteq [m], m \notin X} w_X f_X$ of the functions in B such that $F_2(u_i) = v_i$ for all $1 \leq i \leq m-1$.

Note that as all of the functions in group A satisfy $f(u_m) > v_m$, the weighted average of these functions, F_1, must satisfy $F_1(u_m) > v_m$ as well. Similarly, as all of the functions in the functions in group B satisfy $f(u_m) < v_m$, the weighted average, F_2, must satisfy $F_2(u_m) < v_m$. As such, there exists a weighted average of the functions F_1 and F_2, $f \equiv W_1 F_1 + W_2 F_2$, with $0 \leq W_1, W_2 \leq 1$ and $W_1 + W_2 = 1$, such that $f(u_m) = v_m$. Furthermore, note that as $F_1(u_i) = F_2(u_i) = v_i$ for all $1 \leq i \leq m-1$, $f(u_i) = v_i$ for all $1 \leq i \leq m-1$ as well. Thus, $f(u_i) = v_i$ for all $1 \leq i \leq m$. Furthermore, as f is the weighted average of two weighted averages of functions f_X where $X \subseteq [m]$, f is a weighted average of functions f_X where $X \subseteq [m]$, and is hence our desired function. ■

We now establish that any function that is a weighted average of a set of functions that each have q-action less than 1 also has q-action less than 1.

Lemma 3.4. *For any $q \geq 1$, if we have $n \geq 1$ continuous functions $f_1, f_2, \ldots, f_n : [0,1] \to \mathbb{R}$ such that $J_q[f_i] < 1$ for all $1 \leq i \leq n$, then for any weighted average of the functions $f \equiv \sum_{i=1}^n w_i f_i$, with $0 \leq w_i \leq 1$ for all $1 \leq i \leq n$ and $\sum_{i=1}^n w_i = 1$, we have $J_q[f] < 1$.*

Proof. We can see that

$$\begin{aligned} J_q[f] &= \int_0^1 |f'(x)|^q \mathrm{d}x = \int_0^1 \left|\sum_{i=1}^n w_i f_i'(x)\right|^q \mathrm{d}x \le \int_0^1 \left(\sum_{i=1}^n |w_i| \left|f_i'(x)\right|\right)^q \mathrm{d}x \\ &\le \int_0^1 \left(\left(\sum_{i=1}^n w_i \left|f_i'(x)\right|^q\right)^{\frac{1}{q}}\right)^q \mathrm{d}x = \int_0^1 \left(\sum_{i=1}^n w_i \left|f_i'(x)\right|^q\right) \mathrm{d}x \\ &= \sum_{i=1}^n \left(\int_0^1 w_i \left|f_i'(x)\right|^q \mathrm{d}x\right) = \sum_{i=1}^n w_i J_q[f_i] < \sum_{i=1}^n w_i = 1, \end{aligned}$$

where the fourth step follows by the Weighted Power Mean Inequality, as $q \ge 1$. ■

Putting everything together, we have the following key result.

Lemma 3.5. *Given a set $S = \{(u_i, v_i) : 1 \le i \le m\}$ of m points in $[0,1] \times \mathbb{R}$ with $u_1 < \cdots < u_m$ and any $q \ge 1$, if there exists an absolutely continuous function $f : [0,1] \to \mathbb{R}$ with $J_q[f] < 1$ such that $f(u_i) = v_i$ for all $1 \le i \le m$, then there exists a polynomial $P : [0,1] \to \mathbb{R}$ with $J_q[P] < 1$ such that $P(u_i) = v_i$ for all $1 \le i \le m$.*

Proof. Assume that $m \ge 2$, as the case where there is only one point in S is trivial. Note that by Lemma 1.1, the condition that $J_q[f] < 1$ implies that $J_q[f_S] < 1$. Now, pick a small $\epsilon_1 > 0$ (which we will specify later), and define the set S_X, for each subset $X \subseteq [m]$, as follows:

$$S_X = \{(u_i, v_i + \epsilon_1) : 1 \le i \le m, i \in X\} \cup \{(u_i, v_i - \epsilon_1) : 1 \le i \le m, i \notin X\}.$$

As such, each S_X contains exactly m points, each of whose y-values are within ϵ_1 from the y-values of the corresponding points in S. For each X, we claim that setting $\epsilon_1 > 0$ sufficiently small, we can guarantee $J_q[f_{S_X}] < 1$. Indeed, if we let $d_0, d_1, \ldots, d_m$ denote the slopes of the segments of f_S in order, and define $d_0', d_1', \ldots, d_m'$ to be the slopes of the segments of f_{S_X} in order, notice that $d_0' = d_m' = 0$, and for each $1 \le i \le m-1$, we have either $d_i' = d_i$, $d_i' = d_i + \frac{2\epsilon_1}{u_{i+1} - u_i}$, or $d_i' = d_i - \frac{2\epsilon_1}{u_{i+1} - u_i}$. As such, in any case, we have

$$|d_i'| \le |d_i| + \frac{2\epsilon_1}{u_{i+1} - u_i} \le |d_i| + \frac{2\epsilon_1}{\min_{1 \le i \le m-1} |u_{i+1} - u_i|} = |d_i| + C\epsilon_1,$$

if we substitute $C = \frac{2}{\min_{1 \le i \le m-1} |u_{i+1} - u_i|}$, which is fixed for a set S. Notice that for each i, we can set ϵ_1 sufficiently small such that the inequality $(|d_i| + C\epsilon_1)^q - |d_i|^q < \frac{1 - J_q[f_S]}{m}$ holds, as $\frac{1 - J_q[f_S]}{m}$ is positive. Hence, let ϵ_1 be sufficiently small such that the inequality $(|d_i| + C\epsilon_1)^q - |d_i|^q < \frac{1 - J_q[f_S]}{m}$ holds for each $1 \le i \le m-1$.

Now, notice that

$$\begin{aligned} J_q[f_{S_X}] - J_q[f_S] &= \sum_{i=1}^{m-1} |u_{i+1} - u_i| |d_i'|^q - \sum_{i=1}^{m-1} |u_{i+1} - u_i| |d_i|^q = \sum_{i=1}^{m-1} |u_{i+1} - u_i| (|d_i'|^q - |d_i|^q) \\ &\le \sum_{1 \le i \le m-1, |d_i'| > |d_i|} |u_{i+1} - u_i| \left(|d_i'|^q - |d_i|^q\right) \\ &\le \sum_{1 \le i \le m-1, |d_i'| > |d_i|} \left(|d_i'|^q - |d_i|^q\right) \\ &\le \sum_{1 \le i \le m-1, |d_i'| > |d_i|} ((|d_i| + C\epsilon_1)^q - |d_i|^q) < m \cdot \frac{1 - J_q[f_S]}{m} = 1 - J_q[f_S], \end{aligned}$$

implying that $J_q[f_{S_X}] < 1$. Therefore, we can pick sufficiently small ϵ_1 such that for each $X \subseteq [m]$, we have $J_q[f_{S_X}] < 1$.

Now, for each $X \subseteq [m]$, pick a sufficiently small $\epsilon_2 > 0$ such that $\epsilon_2 < \epsilon_1$ and $\epsilon_2 < 1 - J_q[f_{S_X}]$, and let $P_X : [0,1] \to \mathbb{R}$ be a polynomial such that

$$|P_X(u_i) - f_{S_X}(u_i)| < \epsilon_2 \text{ for all } 1 \leq i \leq m, \text{ and } J_q[P_X] < J_q[f_{S_X}] + \epsilon_2,$$

which must be possible by Lemma 3.2. By the first condition, we can see that whenever $1 \leq i \leq m$ satisfies $i \in X$, we have

$$P_X(u_i) > f_{S_X}(u_i) - \epsilon_2 > f_{S_X}(u_i) - \epsilon_1 = v_i,$$

and whenever $i \notin m$, we have

$$P_X(u_i) < f_{S_X}(u_i) + \epsilon_2 < f_{S_X}(u_i) + \epsilon_1 = v_i.$$

Furthermore, the second condition yields that $J_q[P_X] < 1$ as well.

As such, we have 2^m polynomials P_X, corresponding to each subset $X \subseteq [m]$, such that for each X, it holds for any integer $1 \leq i \leq m$ with $i \in X$ that $P_X(u_i) > v_i$, and it holds for any integer $1 \leq i \leq m$ with $i \notin X$ that $P_X(u_i) < v_i$. By Lemma 3.3, there exists a weighted average of the polynomials, another polynomial, $P \equiv \sum_{X \subseteq [m]} w_X P_X$, with weights $0 \leq w_X \leq 1$ for each X and $\sum_{X \subseteq [m]} w_X = 1$, such that $P(u_i) = v_i$ for all $1 \leq i \leq m$. Furthermore, by Lemma 3.4, as $J_q[P_X] < 1$ for all X, we know that $J_q[P] < 1$ as well. ■

Now, for any $q \geq 1$, let $\mathcal{F}'_q$ denote the class of functions $f : [0,1] \to \mathbb{R}$ such that $J_q[f] < 1$. As such, $\mathcal{F}'_q \subseteq \mathcal{F}_q$. Similarly, let $\mathcal{P}'_q$ denote the class of polynomials $P : [0,1] \to \mathbb{R}$ such that $J_q[P] < 1$. As such, $\mathcal{P}'_q \subseteq \mathcal{P}_q$ and $\mathcal{P}'_q \subseteq \mathcal{F}'_q$. We now establish the following key result that looks very close to our desired final result.

Lemma 3.6. *For all $p > 0$ and $q \geq 1$, we have $\text{opt}_p(\mathcal{P}'_q) = \text{opt}_p(\mathcal{F}'_q)$.*

Proof. Note that by Lemma 3.5, given any finite set S of points in $[0,1] \times \mathbb{R}$, there exists a polynomial $P \in \mathcal{P}'_q$ passing through all the points in S if and only if there exists a general smooth function $f \in \mathcal{F}'_q$ passing through all the points in S.

We claim that this means that the adversary can replicate any strategy that it uses in the learning scenario of $\mathcal{F}'_q$ to maximize the learner's error to the learning scenario of $\mathcal{P}'_q$, and vice versa. This would imply that the worst-case learning error is equal in the two scenarios.

To prove this, note that for any sequence of points $(x_0, y_0), (x_1, y_1), \ldots, (x_m, y_m)$ that the adversary reveals in the learning scenario of $\mathcal{F}'_q$, the same sequence of points can be revealed in the learning scenario of $\mathcal{P}'_q$ (and vice versa as well), as the existence of a general function $f \in \mathcal{F}'_q$ passing through the sequence of points implies the existence of a polynomial $P \in \mathcal{P}'_q$ passing through the same sequence of points (and vice versa as well, trivially). From here, it follows that $\text{opt}_p(\mathcal{P}'_q) = \text{opt}_p(\mathcal{F}'_q)$. ■

Now, we prove that the following equality holds.

Lemma 3.7. *For all $p > 0$ and $q \geq 1$, we have $\text{opt}_p(\mathcal{F}'_q) = \text{opt}_p(\mathcal{F}_q)$.*

Proof. For a real number $c > 0$, define the family of functions $c\mathcal{F}_q$ to contain precisely the functions $g : [0,1] \to \mathbb{R}$ such that $g \equiv cf$, where $f \in \mathcal{F}_q$. Note that $c\mathcal{F}_q$ contains exactly the continuous functions $g : [0,1] \to \mathbb{R}$ with $J_q[g] \leq c$. As such, for every $0 < c < 1$, we have that $c\mathcal{F}_q \in \mathcal{F}'_q$. Thus, $\text{opt}_p(c\mathcal{F}_q) \leq \text{opt}_p(\mathcal{F}'_q)$ for any $0 < c < 1$.

Furthermore, we claim that

$$\operatorname{opt}_p(c\mathcal{F}_q) = c^p \operatorname{opt}_p(\mathcal{F}_q)$$

for any $c > 0$. Indeed, as the family $c\mathcal{F}_q$ is precisely the family $\mathcal{F}_q$ scaled by a factor of c, the raw prediction errors of each round in the learning scenario of $c\mathcal{F}_q$ are effectively scaled by a factor of c from the raw prediction errors of the learning scenario of $\mathcal{F}_q$. Yet, as the raw prediction errors are raised to the p^{th} power, the total error function gets scaled by a factor of c^p, yielding the equation. Thus, we have

$$c^p \operatorname{opt}_p(\mathcal{F}_q) \leq \operatorname{opt}_p(\mathcal{F}'_q)$$

for all $0 < c < 1$. Taking $c \to 1$, we can see that $\operatorname{opt}_p(\mathcal{F}'_q)$ cannot be any less than $\operatorname{opt}_p(\mathcal{F}_q)$. Yet, as $\mathcal{F}'_q \in \mathcal{F}_q$, we also have $\operatorname{opt}_p(\mathcal{F}'_q) \leq \operatorname{opt}_p(\mathcal{F}_q)$. We thus have $\operatorname{opt}_p(\mathcal{F}'_q) = \operatorname{opt}_p(\mathcal{F}_q)$. ■

Using the same idea, we can get a similar result concerning $\mathcal{P}'_q$ and $\mathcal{P}_q$.

Lemma 3.8. *For all $p > 0$ and $q \geq 1$, we have $\operatorname{opt}_p(\mathcal{P}'_q) = \operatorname{opt}_p(\mathcal{P}_q)$.*

Combining all of the above, we get our result.

Theorem 3.9. *For any $p > 0$ and $q \geq 1$, we have $\operatorname{opt}_p(\mathcal{P}_q) = \operatorname{opt}_p(\mathcal{F}_q)$.*

4 Online Learning of Smooth Functions with Noisy Feedback

We start with the proof of Theorem 1.5, a fundamental result on the number of initial feedback rounds necessary in the noisy model.

Theorem 4.1. *For any integer $\eta \geq 1$, if incorrect feedback can be given up to η times, then at least $2\eta + 1$ initial rounds must be thrown out for the learner to guarantee finite error in its first prediction that counts towards error evaluation.*

Proof. We first show that after receiving 2η initial rounds of feedback, the learner cannot guarantee finite error on its next trial. Consider any sequence of inputs $\sigma = (x_0, x_1, \ldots, x_{2\eta-1}) \in [0,1]^{2\eta}$. Let the adversary claim that $f(x_i) = 0$ for all $0 \leq i \leq \eta - 1$, and that $f(x_i) = C$ for all $\eta \leq i \leq 2\eta - 1$ for some arbitrarily large C. Now, suppose the learner is queried on the value of $f(x_{2\eta})$. If $\hat{y}_{x_{2\eta}} \geq \frac{C}{2}$, then let the function be $f(x) \equiv 0$, which is valid because the adversary would have lied exactly η times. Similarly, if $\hat{y}_{x_{2\eta}} \leq \frac{C}{2}$, let the function be $f(x) \equiv C$, which would also be valid as it induces η lies. In any case, the raw error $|\hat{y}_{x_{2\eta}} - f(x_{2\eta})| \geq \frac{C}{2}$, which can be made arbitrarily large.

Now, we establish that throwing out $2\eta+1$ initial rounds suffices. We prove a stronger statement, that after $2\eta+1$ initial rounds, the learner is able to bound the value of the function f at any input within a closed interval of length 2. Suppose that in the first $2\eta+1$ rounds, the learner receives the set of points $S = \{(x_i, y_i) : 1 \leq i \leq 2\eta+1\}$, where x_i and y_i denote the queried input and adversary feedback, respectively, in the i^{th} round.

Let $(v_1, v_2, \ldots, v_{2\eta+1})$ be a permutation of $(y_1, y_2, \ldots, y_{2\eta+1})$, such that $v_1 \leq \cdots \leq v_{2\eta+1}$. Now, note that as the adversary may lie at most η times, at least $\eta+1$ points among the $2\eta+1$ points in S must align with the actual function. As such, at least $\eta+1$ elements of $\{v_1, \ldots, v_{2\eta+1}\}$ must be in the range of f. Note that any $\eta+1$-element subset of $\{v_1, \ldots, v_{2\eta+1}\}$ must contain a member v_j with $j \geq \eta+1$, so the range of f must contain some $v_j \geq v_{\eta+1}$. As the difference between the greatest and least values of f is at most 1, from the fact that $f \in \mathcal{F}_q$, we must have $f(x) \geq v_j - 1 \geq v_{\eta+1} - 1$ for all $x \in [0,1]$. On the other hand, as any $\eta+1$-element subset of $\{v_1, \ldots, v_{2\eta+1}\}$ must also contain a member v_k with $k \leq \eta+1$, the range of f must also contain some $v_k \leq v_{\eta+1}$, from which it follows that $f(x) \leq v_{\eta+1} + 1$ for all $x \in [0,1]$. As such, we have that $f(x)$ must be within $[v_{\eta+1} - 1, v_{\eta+1} + 1]$ for any $x \in [0,1]$, as promised. ■

This fundamental result motivates our definition of the worst-case error $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q)$ in Section 1.4 to only count from the $2\eta + 2^{\text{th}}$ round.

We now move on to proving some bounds on $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q)$. We first establish the following fundamental result, characterizing the values of η, p, q, such that the worst-case error is finite.

Theorem 4.2. *For any integer $\eta \geq 1$, the value of $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q)$ is finite if and only if $p, q > 1$.*

Proof. Clearly, if $p = 1$ or $q = 1$, $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q) \geq \text{opt}_p(\mathcal{F}_q) = \infty$. When $p, q > 1$, the key idea is to replicate the optimal strategy in the standard scenario for the same p, q, as if the adversary cannot lie, until something wrong occurs, indicating a lie—the total apparent error exceeds $\text{opt}_p(\mathcal{F}_q)$ (which is finite by Theorem 2.12), or the learner gives feedback not in the allocated range $[v_{\eta+1} - 1, v_{\eta+1} + 1]$. Then, the learner forgets all previous feedback, starts all over again, and repeats. As this can happen at most $\eta + 1$ times, and we can bound the error at every stage, the total error is bounded. ■

We now prove Theorem 1.7. Our proof is split into two parts: an upper bound through Lemma 4.3, and a lower bound through Lemma 4.4. The upper bound uses a similar but more sophisticated strategy as the previous result.

Lemma 4.3. *For any $\eta \geq 1$, $p, q \geq 2$ we have $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q) \leq 12\eta + 6$.*

Proof. Split the domain $[0, 1]$ into four intervals: $I_1 = [0, \frac{1}{4})$, $I_2 = [\frac{1}{4}, \frac{1}{2})$, $I_3 = [\frac{1}{2}, \frac{3}{4})$, $I_4 = [\frac{3}{4}, 1]$. We first claim that whenever the learner has adversary feedback on at least $2\eta + 1$ points in any interval I_j, it can determine a constant c such that for all $x \in I_j$, $f(x) \in [c - \frac{1}{2}, c + \frac{1}{2}]$.

To prove this, suppose that the learner receives the feedback set $S = \{(x_i, y_i) : 1 \leq i \leq 2\eta + 1\}$, where each $x_i \in I_j$ and $y_1 < \cdots < y_{2\eta+1}$. Indeed, letting $c = y_{\eta+1}$, as there must be a true point in S with output at least c, and a true point in S with output at most c, there must exist a $X_1 \in I_j$ with $f(X_1) = c$ by the Intermediate Value Theorem.

Now, if there exists any point $X_2 \in I_j$ with $f(X_2) > c + \frac{1}{2}$, then

$$J_q[f] > J_q\left[f_{\{(X_1,c),(X_2,c+\frac{1}{2})\}}\right] = |X_2 - X_1| \left|\frac{\frac{1}{2}}{X_2 - X_1}\right|^q \geq \left(\frac{1}{4}\right)^{1-q} \left(\frac{1}{2}\right)^q \geq 1$$

as $|X_2 - X_1| \leq \frac{1}{4}$ and $q \geq 2$; contradiction. Similarly there cannot be a point $X_2 \in I_j$ with $f(X_2) < c - \frac{1}{2}$, so for all $X_2 \in I_j$, $f(X_2) \in [c - \frac{1}{2}, c + \frac{1}{2}]$, as claimed.

Now, let A be an optimal learning algorithm in the standard scenario for the same p, q, so that it always guarantees a total error value of at most $\text{opt}_p(\mathcal{F}_q) = 1$. Consider the learning algorithm $\mathcal{A}$ trying to learn a function $f \in \mathcal{F}_q$ in the noisy scenario, following this strategy:

Upon the end of the initial $2\eta + 1$ rounds, first record the value v such that f can be bounded within $[v - 1, v + 1]$, previously shown to be possible. Now, let $\mathcal{A}$ enter Stage 1. For any trial t requesting the value of $f(x_t)$ for $x_t \in I_j$ for some j, if the learner knows at least $2\eta + 1$ points already in I_j, predict exactly as algorithm A would; otherwise, predict v. Furthermore, call the trials of the former kind *mimicked trials*.

Now, let Stage 1 continue until a *mimicked trial*, with $x_t \in I_j$ for some j, where one of the following events happen: 1) The adversary reveals a point (x_t, y) such that $y \notin [c - \frac{1}{2}, c + \frac{1}{2}]$, where c is the constant such that f is bounded within $[c - \frac{1}{2}, c + \frac{1}{2}]$ over I_j; 2) the algorithm A predicts $\hat{y}$ with $\hat{y} \notin [c - \frac{1}{2}, c + \frac{1}{2}]$; or 3) the total error the learner perceives throughout all *mimicked trials* of the Stage exceeds $\text{opt}_p(\mathcal{F}_q) = 1$.

Once one of the three events happen, let $\mathcal{A}$ exit Stage 1, forget all previous adversary feedback, predict v for the immediate next round, and enter Stage 2, repeating the same process until one of

the three events happen again, whereupon it enters Stage 3, etc. Clearly, whenever one of the three events occur, a lie must have occurred since the end of the last stage. Thus, the learning process includes at most $\eta + 1$ stages.

Now, onto error bounding. As there are 4 intervals, each with at most $2\eta+1$ *non-mimicked trials* which generate at most 1 error, the total error from *non-mimicked trials* is at most $4(2\eta+1) = 8\eta+4$.

Now, onto bounding *mimicked trials*. Note that every trial generates at most 1 error. For each stage besides stage $\eta+1$, the total perceived error from the *mimicked trials*, not counting the stage's last trial, is at most $1 + \mathrm{opt}_p(\mathcal{F}_q) = 2$. The last trial of every stage generates an actual error of at most 1.

Over all stages besides stage $\eta + 1$, there are at most η *mimicked trials* where the perceived error is not the actual error. The perceived error differs from the actual error by at most 1 in these trials, since the correct value is in $[c - \frac{1}{2}, c + \frac{1}{2}]$. Thus, the actual error from the first η stages is at most $(2 + 1)\eta + \eta = 4\eta$.

Finally, if the learning process reaches stage $\eta + 1$, the adversary cannot lie any more, so the error from that stage is at most $1+\mathrm{opt}_p(\mathcal{F}_q) = 2$. Therefore, in total, the sum of p^{th} powers of actual errors for *mimicked trials* over all stages is at most $4\eta + 2$. Adding the error from *non-mimicked trials*, we get at most $(4\eta + 2) + (8\eta + 4) = 12\eta + 6$. ■

Lemma 4.4. *For any $\eta \geq 1$, $p, q \geq 2$, we have* $\mathrm{opt}^{\mathrm{nf}}_{p,\eta}(\mathcal{F}_q) \geq 2\eta + 1$.

Proof. Fix any algorithm A for learning $\mathcal{F}_q$. Consider the following adversary strategy. For the first $2\eta + 1$ rounds, repeatedly query the learner at 0, and reveal the output to be 0. Let $f(0) = 0$, so that no lies have been used up yet. For the next $2\eta + 1$ rounds, the adversary repeatedly queries the learner at 1. For the first η of these rounds, reveal the value of $f(1)$ to be -1. Then, for the next η rounds, reveal the value of $f(1)$ to be 1. Then, on the next round, reveal the true value of $f(1)$ to be either 1 or -1, whichever makes the total error function larger. Note that no matter what we choose, exactly η lies have been used up. We claim that we can always guarantee a total error of at least $2\eta + 1$. Indeed, note that

$$\sum_{i=1}^{2\eta+1} |\hat{y}_i + 1|^p + \sum_{i=1}^{2\eta+1} |\hat{y}_i - 1|^p = \sum_{i=1}^{2\eta+1} (|\hat{y}_i + 1|^p + |\hat{y}_i - 1|^p) \geq 2(2\eta + 1),$$

where the last step comes from the fact that for any $p \geq 2$ and any $\hat{y}_i \in \mathbb{R}$, $|\hat{y}_i + 1|^p + |\hat{y}_i - 1|^p \geq 2$. To see why this is true, notice that if $\hat{y}_i \notin [-1, 1]$ the result is obvious; otherwise, Jensen's inequality gives the result.

This means at least one of $\sum_{i=1}^{2\eta+1} |\hat{y}_i+1|^p$ and $\sum_{i=1}^{2\eta+1} |\hat{y}_i-1|^p$ is at least $2\eta+1$. As the adversary can force an error of at least $2\eta + 1$ regardless of the learner's predictions, $\mathrm{opt}^{\mathrm{nf}}_{p,\eta}(\mathcal{F}_q) \geq 2\eta + 1$. ■

Combining the bounds, we obtain a precise bound on $\mathrm{opt}^{\mathrm{nf}}_{p,\eta}(\mathcal{F}_q)$ for any $p, q \geq 2$ and $\eta \geq 1$.

Theorem 4.5. *For any $\eta \geq 1$, $p, q \geq 2$, we have that* $\mathrm{opt}^{\mathrm{nf}}_{p,\eta}(\mathcal{F}_q) = \Theta(\eta)$.

5 Discussion

With the results of this paper, we have completed the characterization of the ordered pairs (p, q) for which $\mathrm{opt}_p(\mathcal{F}_q)$ is finite. This problem has been open since Kimber and Long [11] defined the model of online learning of smooth functions. For the standard learning scenario, a natural next step would be to establish lower and upper bounds on $\mathrm{opt}_p(\mathcal{F}_q)$ that match up to a constant factor for every $p, q \geq 1$.

The paper [9] established precisely that $\text{opt}_p(\mathcal{F}_q) = 1$ for all (p, q) with either $p, q \geq 2$ or $1 < q < 2$ and $p \geq 2 + \frac{1}{q-1}$. Another direction would be to identify more pairs of (p, q) such that $\text{opt}_p(\mathcal{F}_q) = 1$.

Having studied the online learning of polynomials and established that polynomials in $\mathcal{F}_q$ are not any easier to learn than general functions in $\mathcal{F}_q$, it remains to investigate the worst-case error of learning other special subsets of $\mathcal{F}_q$. In particular, we invite future research on generalizing the ideas for our proof that $\text{opt}_p(\mathcal{P}_q) = \text{opt}_p(\mathcal{F}_q)$ to other special subsets $\mathcal{A}_q$ of $\mathcal{F}_q$. Indeed, the crux of our proof for $\mathcal{P}_q$ was a connection with the Weierstrass Approximation Theorem, which allowed us to use polynomials to uniformly approximate any continuous real-valued function. Yet, there exists a generalization of the Weierstrass Approximation Theorem, the Stone-Weierstrass Theorem. This theorem offers a condition on general subsets of the set of continuous functions over $[0, 1]$, which if satisfied implies that the functions in the subset can uniformly approximate every continuous real function over $[0, 1]$. As such, the Stone-Weierstrass Theorem may be a gateway to proving that more special subsets $\mathcal{A}_q$ of $\mathcal{F}_q$ are as hard to learn as $\mathcal{F}_q$ itself. Specifically, we conjecture that it is not easier to learn sums of exponential functions in $\mathcal{F}_q$, or trigonometric polynomials in $\mathcal{F}_q$, than $\mathcal{F}_q$ in general.

In the noisy learning scenario, we have shown that $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q) = \Theta(\eta)$ for any η, and all $p, q \geq 2$. However, we have yet to obtain precise values of $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q)$ at any particular values of p, q, η. This could be an interesting research direction, alongside establishing upper bounds on $\text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q)$ when $p < 2$ or $q < 2$.

Motivated by our lower bound of $2\eta + 1$ for all $p, q \geq 2$, we make the following conjecture about the behavior of the noisy learning scenario as the parameter p approaches infinity.

Conjecture 5.1. *For any $\eta \geq 1$, $q \geq 1$, we have $\lim_{p\to\infty} \text{opt}^{\text{nf}}_{p,\eta}(\mathcal{F}_q) = 2\eta + 1$.*

It would also be interesting to investigate online learning of smooth functions with other modes of adversary feedback. In particular, delayed ambiguous reinforcement has been studied for the online learning of classifiers [2, 8]. Another possibility would be feedback on the absolute value of prediction error.

6 Acknowledgements

I would like to thank my PRIMES mentor, Dr. Jesse Geneson, for introducing me to this research topic, providing new ideas for research directions, and offering useful feedback on my proofs. I am immensely grateful for his continued patience throughout the entirety of my work on this paper. I would also like to thank the MIT PRIMES-USA program for providing this wonderful research opportunity.

References

[1] D. Angluin, Queries and concept learning. *Machine Learning* **2** (1988) 319–342.

[2] P. Auer and P.M. Long. Structural results about on-line learning models with and without queries. *Machine Learning* **36** (1999) 147-181.

[3] A. Barron, Approximation and estimation bounds for artificial neural networks. *Workshop on Computational Learning Theory* (1991).

[4] N. Cesa-Bianchi, P.M. Long, and M.K. Warmuth, Worst-case quadratic loss bounds for prediction using linear functions and gradient descent. *IEEE Transactions on Neural Networks* 7 (1996) 604–619.

[5] N. Cesa-Bianchi, Y. Freund, D. Helmbold, and M. Warmuth. Online prediction and conversion strategies. *Machine Learning*, 25:71–110, 1996.

[6] V. Faber and J. Mycielski, Applications of learning theorems. *Fundamenta Informaticae* **15** (1991) 145–167.

[7] Y. Filmus, S. Hanneke, I. Mehalel, and S. Moran. Optimal prediction using expert advice and randomized littlestone dimension. *Proceedings of Thirty Sixth Conference on Learning Theory*, pages 773–836, 2023.

[8] R. Feng, J. Geneson, A. Lee, and E. Slettnes, Sharp bounds on the price of bandit feedback for several models of mistake-bounded online learning. *Theoretical Computer Science* **965**: 113980 (2023).

[9] J. Geneson and E. Zhou. Online learning of smooth functions. *Theoretical Computer Science* **979C**: 114203 (2023).

[10] W. Hardle, Smoothing techniques. Springer Verlag (1991).

[11] D. Kimber and P. M. Long, On-line learning of smooth functions of a single variable. *Theoretical Computer Science* **148** (1995) 141–156.

[12] N. Littlestone, Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning* **2** (1988) 285–318.

[13] N. Littlestone and M.K. Warmuth, The weighted majority algorithm. *Proceedings of the 30th Annual Symposium on the Foundations of Computer Science* (1989)

[14] P. M. Long, Improved bounds about on-line learning of smooth functions of a single variable. *Theoretical Computer Science* **241** (2000) 25–35.

[15] J. Mycielski, A learning algorithm for linear operators. *Proceedings of the American Mathematical Society* **103** (1988) 547–550.