Improved Bounds about On-line Learning of Smooth Functions of a Single Variable

Philip M. Long
ISCS Department
National University of Singapore
Singapore 119260, Republic of Singapore
plong@iscs.nus.sg

November 25, 1997

Abstract

We consider the complexity of learning classes of smooth functions formed by bounding different norms of a function's derivative. The learning model is the generalization of the mistake-bound model to continuous-valued functions. Suppose F_q is the set of all absolutely continuous functions f from $[0,1]$ to $\mathbf{R}$ such that $||f'||_q \leq 1$, and $\mathrm{opt}(F_q, m)$ is the best possible bound on the worst-case sum of absolute prediction errors over sequences of m trials. We show that for all $q \geq 2$, $\mathrm{opt}(F_q, m) = \Theta(\sqrt{\log m})$, and that $\mathrm{opt}(F_2, m) \leq \frac{\sqrt{\log_2 m}}{2} + O(1)$, matching a known lower bound of $\frac{\sqrt{\log_2 m}}{2} - O(1)$ to within an additive constant.

1 Introduction

In this paper, we continue a line of research investigating the complexity of learning, in the on-line model, classes of functions intended to capture the idea of similar inputs tending to yield similar outputs.

In the model that we will consider here [6, 1, 7], an algorithm is trying to learn a real-valued function f, given the a priori knowledge that f comes from some class F. Learning proceeds in *trials*, where, in the tth trial, the algorithm

- gets $x_t \in [0,1]$,
- outputs a prediction $\hat{y}_t$ of $f(x_t)$, and
- discovers $f(x_t)$.

An algorithm A is evaluated by the worst-case sum of its absolute prediction errors, i.e.[1] by its worst-case value of $\sum_{t=1}^{m} |\hat{y}_t - f(x_t)|$. We refer to the best possible bound on this quantity as a function of m as $\mathrm{opt}(F, m)$. This is defined formally in Section 2.

[1] We number our trials from 0, but, as in [4], we start counting errors on trial number 1. This is for technical reasons: we could obtain similar results without this if we set the range to be $[0,1]$, or required that $f(0) = 0$.

	F_∞		F_2	
	previous	current	previous	current
upper bounds	$O(\log m)$	$O(\sqrt{\log m})$	$O(\log m)$	$\frac{\sqrt{\log(m+2)}}{2}+1$
lower bounds	$\Omega(\psi(m))$ for some unbounded ψ	$\Omega(\sqrt{\log m})$	$\frac{\sqrt{\lfloor \log m \rfloor}}{2}$	$\frac{\sqrt{\lfloor \log m \rfloor}}{2}$

Table 1: Comparison between the current and previous state of knowledge about $\mathrm{opt}(F_\infty, m)$ and $\mathrm{opt}(F_2, m)$. All previous results are due to Kimber and Long [4].

Since the derivative measures the rate that the output is changing with the input, a norm of the derivative measures the overall tendency of similar inputs to yield similar outputs. For this reason, for various q, we will study the set F_q of all absolutely continuous functions f from $[0,1]$ to $\mathbf{R}$ such that $\int |f'(x)|^q \, dx \leq 1$.

The set F_∞ is defined analogously using the limit as q goes to infinity. This set can be defined in a simpler way (see [8]) as the set of functions with a Lipschitz bound of 1, i.e. the set of functions f for which for all $a, b \in [0,1]$, $|f(a) - f(b)| \leq |a - b|$. Informally, this is the set of functions for which the outputs are never more dissimilar than the inputs.

In this paper, we show that for all $q \geq 2$,

$$\mathrm{opt}(F_q, m) = \Theta(\sqrt{\log m}). \tag{1}$$

We also show that $\mathrm{opt}(F_2, m) \leq \frac{\sqrt{\log_2 m}}{2} + O(1)$. Together with a known lower bound [4], this implies that

$$\mathrm{opt}(F_2, m) = \frac{\sqrt{\log_2 m}}{2} \pm O(1). \tag{2}$$

Since if $p \leq q$, $F_q \subseteq F_p$, which implies that $\mathrm{opt}(F_q, m) \leq \mathrm{opt}(F_p, m)$, (1) can be established by proving an $O(\sqrt{\log m})$ upper bound on $\mathrm{opt}(F_2, m)$, and an $\Omega(\sqrt{\log m})$ lower bound on $\mathrm{opt}(F_\infty, m)$. Upper and lower bounds on $\mathrm{opt}(F_\infty, m)$ and $\mathrm{opt}(F_2, m)$ were implicit[2] in the work of Kimber and Long [4]. The state of knowledge about these classes before and after this paper is summarized in Table 1.

In addition to the work from [4] described above, F_2 was studied in an analogous model using the quadratic loss ($(\hat{y}_t - f(x_t))^2$) by Faber and Mycielski [3] and in [4]. Cesa-Bianchi, Long, and Warmuth [2] extended this work to the noisy case.

As mentioned in [4], these results can be trivially generalized via scaling, both to allow any bounded interval as the domain, and to allow bounds other than 1 on whatever norm of the derivative.

2 Definitions

Denote the reals by $\mathbf{R}$. We refer the reader to [8] for the definitions and facts from elementary real analysis used here.

[2] For their proof of the upper bounds, they used slightly stronger assumptions than that the functions were absolutely continuous. To get the bounds listed in Table 1 under "previous" from their results, all that is needed is Lemma 3 of the present paper, which is easily proved.

For some set $A \subseteq \mathbf{R}$, define floor_A and ceil_A by

$$\begin{aligned}\text{floor}_A(x) &= \sup(A \cap (-\infty, x]) \\ \text{ceil}_A(x) &= \inf(A \cap [x, \infty)).\end{aligned}$$

For finite A, $\text{floor}_A(x)$ is the greatest element of A no bigger than x, and $\text{ceil}_A(x)$ is the least element of A at least as big as x, so if the points of $A \cup \{x\}$ are plotted on the number line, $\text{floor}_A(x)$ and $\text{ceil}_A(x)$ will be the two points plotted on either side of x.

In the model considered in this paper [6, 7], learning proceeds in *trials*. The algorithm is trying to learn a function $f : [0, 1] \rightarrow \mathbf{R}$. In each trial $t = 0, 1, 2, ...$ an algorithm

- is given $x_t \in [0, 1]$,
- outputs $\hat{y}_t \in \mathbf{R}$, and
- receives $f(x_t) \in \mathbf{R}$.

For a learning algorithm A, we define

$$L(A, F, m) = \sup_{f \in F, x_0, \ldots, x_m \in [0,1]} \sum_{t=1}^{m} |\hat{y}_t - f(x_t)|,$$

where the $\hat{y}_t$'s are generated from A, f, and the x_t's as described above. We then define

$$\text{opt}(F, m) = \inf_A L(A, F, m)$$

where the infimum ranges over learning algorithms.

Choose $q \geq 1$. Define F_q to be the set of all absolutely continuous functions $f : [0, 1] \rightarrow \mathbf{R}$ such that

$$\int |f'(x)|^q \, dx \leq 1.$$

Since any absolutely continuous function is differentiable almost everywhere, the left hand side is always well-defined for such functions.

The following is the first of this paper's main results.

Theorem 1 *For all $q \geq 2$,*

$$\text{opt}(F_q, m) = \Theta(\sqrt{\log m}).$$

Putting our upper bound on $\text{opt}(F_2, m)$ (Theorem 7) together with [4, Theorem 21], we obtain the other main result.

Theorem 2

$$\text{opt}(F_2, m) = \frac{\sqrt{\log m}}{2} \pm O(1).$$

3 The upper bound

Suppose $S = \{(u_i, v_i) : 1 \le i \le m\}$ is a finite subset of $[0,1] \times \mathbf{R}$ such that

$$u_1 < u_2 < \cdots < u_m.$$

Define $f_S : [0,1] \to \mathbf{R}$ to be the function which linearly interpolates the points in S and extrapolates with the constants v_1 and v_m respectively. That is, for all x, $f_\emptyset(x) = 0$, and

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

if $|S| \ge 1$.

For $f : [0,1] \to \mathbf{R}$, define the *action* of f, denoted by $J[f]$, to be

$$J[f] = \int_0^1 f'(x)^2 dx. \tag{3}$$

Note that F_2 is the set of absolutely continuous functions whose action is at most 1.

Facts similar to the following lemma are known (see [5]), but we include a proof in an appendix since we do not know a reference for precisely this statement.

Lemma 3 *Choose $m \in \mathbf{N}$. Choose $(u_1, v_1), ..., (u_m, v_m) \in [0,1] \times \mathbf{R}$ such that the u_i's are distinct. Let $S = \{(u_i, v_i) : 1 \le i \le m\}$. If f is an absolutely continuous function such that for all $i \le m$, $f(u_i) = v_i$, then $J[f] \ge J[f_S]$.*

Proof: In Appendix A. □

Next, we record a lemma implicit in the analysis of [4] that describes the change in the action of f_S when a pair is added to S.

Lemma 4 ([4]) *Choose $m \in \mathbf{N}$. Let $(u_1, v_1), ..., (u_m, v_m)$ be a sample with $0 \le u_1 < u_2 < \cdots < u_m \le 1$. Let $S = \{(u_i, v_i) : 1 \le i \le m\}$ and let $U = \{u_i : 1 \le i \le m\}$. Choose an example $(x, y) \in [0,1] \times \mathbf{R}$ such that $x \notin U$. If $x \in [u_1, u_m]$, then*

$$J[f_{S \cup \{(x,y)\}}] = J[f_S] + \frac{(\mathrm{ceil}_U(x) - \mathrm{floor}_U(x))(y - f_S(x))^2}{(\mathrm{ceil}_U(x) - x)(x - \mathrm{floor}_U(x))}.$$

If $x \notin [u_1, u_m]$, then

$$J[f_{S \cup \{(x,y)\}}] = J[f_S] + \frac{(y - f_S(x))^2}{\min_i |x - u_i|}. \tag{4}$$

Finally, we establish some technical lemmas, whose proofs are given in appendices.

Lemma 5 *For any $m \in \mathbf{N}$, $q_1, ..., q_m \in \mathbf{R}$ and $r_1, ..., r_m, z > 0$, if $\sum_{i=1}^m q_i^2/r_i \le 1$ and $\sum_{i=1}^m r_i \le z$, then*

$$\sum_{i=1}^m q_i \le \sqrt{z}.$$

Proof: In Appendix B. □

Lemma 6 *For all $q, r \geq 0$ for which $q \geq r$,*

$$r \log_2 \frac{1}{r} + (q - r) \log_2 \frac{1}{q - r} - q \log_2 \frac{1}{q} \geq \frac{4r(q - r)}{q}.$$

Proof: In Appendix C. □

Now we are ready for the main result of this section.

Theorem 7 *For any $m \geq 1$,*

$$\mathrm{opt}(F_2, m) \leq \frac{\sqrt{\log_2(m + 2)}}{2} + 1.$$

Proof: Consider the algorithm, call it A, that interpolates linearly and extrapolates using the nearest neighbor. Specifically, algorithm A, on the tth trial, gets x_t from the environment, outputs $f_{\{(x_i, f(x_i)) : i < t\}}(x_t)$, and gets $f(x_t)$.

Choose $x_0, ..., x_m \in [0, 1], f \in F_2$. Let $\hat{y}_1, ..., \hat{y}_m$ be the predictions generated from these by A in the obvious way. Assume without loss of generality that the x_t's are distinct. For each $t \in \mathbf{N}, t \leq m$ let $X_t = \{x_s : 0 \leq s < t\}$. Define

$$\mathrm{IN} = \{t \in \{1, ..., m\} : x_t \in [(\min X_t), (\max X_t)]\}$$

and

$$\mathrm{OUT} = \{1, ..., m\} - \mathrm{IN}.$$

Note that the elements of X_t can be viewed as the dividers of a partition of $[0, 1]$ into subintervals, and that such a partition can in turn be viewed as a probability distribution. Define H_t to be the entropy of that probability distribution. In other words, if $u_0 < ... < u_{t-1}$ are the elements of X_t in sorted order, define

$$H_t = u_0 \log_2 \frac{1}{u_0} + \left(\sum_{s=1}^{t-1} (u_s - u_{s-1}) \log_2 \frac{1}{u_s - u_{s-1}} \right) + (1 - u_{t-1}) \log_2 \frac{1}{1 - u_{t-1}}.$$

We will bound the total error of algorithm A by bounding the errors incurred in trials in IN and trials in OUT separately.

We begin with the trials in IN. Lemma 4 implies that for each $t \in \mathrm{IN}$, the action of A's hypothesis increases by

$$\frac{(\mathrm{ceil}_{X_t}(x_t) - \mathrm{floor}_{X_t}(x_t))(f(x_t) - \hat{y}_t)^2}{(\mathrm{ceil}_{X_t}(x_t) - x_t)(x_t - \mathrm{floor}_{X_t}(x_t))}.$$

Since

- A's original hypothesis has action zero,
- Lemma 3 implies that the action of A's hypothesis is at most that of f which is in turn at most 1, and
- Lemma 4 implies that the action of A's hypothesis does not decrease after trials in OUT,

we have

$$\sum_{t \in \mathrm{IN}} \frac{(\mathrm{ceil}_{X_t}(x_t) - \mathrm{floor}_{X_t}(x_t))(f(x_t) - \hat{y}_t)^2}{(\mathrm{ceil}_{X_t}(x_t) - x_t)(x_t - \mathrm{floor}_{X_t}(x_t))} \leq 1. \tag{5}$$

By inspection, for $t \in \mathrm{IN}$,

$$\begin{aligned} H_{t+1} - H_t = & (x_t - \mathrm{floor}_{X_t}(x_t)) \log_2 \tfrac{1}{x_t - \mathrm{floor}_{X_t}(x_t)} \\ & + (\mathrm{ceil}_{X_t}(x_t) - x_t) \log_2 \tfrac{1}{\mathrm{ceil}_{X_t}(x_t) - x_t} \\ & - (\mathrm{ceil}_{X_t}(x_t) - \mathrm{floor}_{X_t}(x_t)) \log_2 \tfrac{1}{\mathrm{ceil}_{X_t}(x_t) - \mathrm{floor}_{X_t}(x_t)}, \end{aligned}$$

so Lemma 6 implies that for $t \in \mathrm{IN}$,

$$H_{t+1} - H_t \geq \frac{4(\mathrm{ceil}_{X_t}(x_t) - x_t)(x_t - \mathrm{floor}_{X_t}(x_t))}{\mathrm{ceil}_{X_t}(x_t) - \mathrm{floor}_{X_t}(x_t)}. \tag{6}$$

Since $H_1 = 0$, since H_t is nondecreasing in t, and since for all t, $H_t \leq \log_2(t+1)$, (6) implies that

$$\sum_{t \in \mathrm{IN}} \frac{(\mathrm{ceil}_{X_t}(x_t) - x_t)(x_t - \mathrm{floor}_{X_t}(x_t))}{\mathrm{ceil}_{X_t}(x_t) - \mathrm{floor}_{X_t}(x_t)} \leq \frac{\log_2(m+2)}{4}.$$

Putting this together with (5) and Lemma 5, we have

$$\sum_{t \in \mathrm{IN}} |f(x_t) - \hat{y}_t| \leq \frac{\sqrt{\log_2(m+2)}}{2}. \tag{7}$$

Now we turn to the trials in OUT. Here, applying Lemma 4, for each $t \in \mathrm{OUT}$, the action of A's hypothesis increases by at least

$$\frac{(f(x_t) - \hat{y}_t)^2}{\min\{|x_t - u| : u \in X_t\}}.$$

Arguing as above, this implies that

$$\sum_{t \in \mathrm{OUT}} \frac{(f(x_t) - \hat{y}_t)^2}{\min\{|x_t - u| : u \in X_t\}} \leq 1. \tag{8}$$

Since for each $t \in \mathrm{OUT}$,

$$\max X_{t+1} - \min X_{t+1} \geq (\max X_t - \min X_t) + \min\{|x_t - u| : u \in X_t\},$$

the fact that $X_{m+1} \subseteq [0,1]$ implies that

$$\sum_{t \in \mathrm{OUT}} \min\{|x_t - u| : u \in X_t\} \leq 1.$$

Putting this together with (8) and Lemma 5, we have

$$\sum_{t \in \mathrm{OUT}} |f(x_t) - \hat{y}_t| \leq 1.$$

Putting this together with (7) completes the proof. □

4 The lower bound

To prove Theorem 1, all that remains is to prove a lower bound for F_∞. This proof builds on a lower bound argument for F_2 [4].

Theorem 8 *For $m \in \mathbf{N}$,*

$$\mathrm{opt}(F_\infty, m) \geq \frac{\sqrt{\lfloor \log_2(1+m) \rfloor}}{8}.$$

Proof: Let $k = \lfloor \log_2(1+m) \rfloor$. Let $x_0 = 1$ and $y_0 = 0$. For $i \in \mathbf{N}, j \in \mathbf{Z}, 0 \leq j < 2^{i-1}$, let

$$x_{2^{i-1}+j} = \frac{1}{2^i} + \frac{j}{2^{i-1}}.$$

Consider trials 2^{i-1} through $2^i - 1$ to be part of stage i. For example, for large m, we have

stage 1:	$x_1 = 1/2,$
stage 2:	$x_2 = 1/4, x_3 = 3/4,$
stage 3:	$x_4 = 1/8, x_5 = 3/8, x_6 = 5/8, x_7 = 7/8$
$\vdots$	$\vdots$

Choose an algorithm A for learning F_∞. We will construct, using algorithm A, a sequence $f_0, f_1, ..., f_{2^k-1} \in F_\infty$ and $y_1, ..., y_{2^k-1} \in \mathbf{R}$ where if f_{2^k-1} is the target function, then f_{2^k-1} is consistent with the x_t's and y_t's and algorithm A has total error at least $\sqrt{k}/8$.

For the sake of the argument, we will also define

$$g_{1,0}, g_{1,1}, g_{2,0}, ..., g_{2,2}, ..., g_{k,0}, ..., g_{k,2^{k-1}} \in F_2$$

and $v_1, ..., v_{2^k-1} \in \mathbf{R}$.

Set f_0 to be the constant 0 function.

Choose a stage i. Let $g_{i,0} = f_{2^{i-1}-1}$, that is, f_t for the last trial t before the beginning of stage i. Choose a trial t in some stage i. Set $v_t = f_{t-1}(x_t) \pm \frac{1}{2^{i+1}\sqrt{k}}$, whichever is furthest from $\hat{y}_t$, and let $g_{i,t-2^{i-1}+1}$ be the function which linearly interpolates $\{(0,0),(1,0)\} \cup \{(x_s, y_s) : s < 2^{i-1}\} \cup \{(x_s, v_s) : 2^{i-1} \leq s \leq t\}$.

Let u_{left} and u_{right} be the two elements of $\{0,1\} \cup \{x_s : s < t\}$ that are closest to x_t. Then if $|v_t - f_{t-1}(u_{\mathrm{left}})| \leq 2^{-i}$ and $|v_t - f_{t-1}(u_{\mathrm{right}})| \leq 2^{-i}$ then set $y_t = v_t$. Otherwise, set $y_t = f_{t-1}(x_t)$; in this case, we say that we *pass* on trial t. Informally, we set $y_t = v_t$, unless doing so would make any function consistent with $(x_1, y_1), ..., (x_t, y_t)$ violate the Lipschitz condition. Let f_t be the function which linearly interpolates $\{(0,0),(1,0)\} \cup \{(x_s, y_s) : s \leq t\}$.

By construction, each $f_t \in F_\infty$. We claim that, for each $g_{i,j}$, $J[g_{i,j}] \leq 1/4$. This is proved by double induction, first on the index of the stage. We claim that for each i,

$$J[f_{2^{i-1}-1}] \leq \frac{i-1}{4k}. \tag{9}$$

When $i = 1$, this is true since $J[f_0] = 0$.

Choose a stage $i \geq 1$. We assume that (9) holds for i, and will prove that it holds for $i+1$. We claim that for each $j = 0, ..., 2^{i-1}$, that

$$J[g_{i,j}] \leq \frac{i-1}{4k} + \frac{j}{k2^{i+1}}. \tag{10}$$

When $j = 0$, this is true by (9) and the definition of $g_{i,0}$. Choose $j \in \{0, ..., 2^{i-1} - 1\}$, and assume (10) holds for j. Applying Lemma 4,

$$J[g_{i,j+1}] = J[g_{i,j}] + \frac{2\left(\frac{1}{2^{i+1}\sqrt{k}}\right)^2}{2^{-i}} = J[g_{i,j}] + \frac{1}{k2^{i+1}}.$$

Applying the induction hypothesis, we get

$$J[g_{i,j+1}] \leq \frac{i-1}{4k} + \frac{j}{k2^{i+1}} + \frac{1}{k2^{i+1}}.$$

This completes the proof of the induction step for the induction over j. Plugging in $j = 2^{i-1}$, we get

$$J[g_{i,2^{i-1}}] \leq \frac{i}{4k}. \tag{11}$$

But, since Lemma 4 implies that for all $j = 0, ..., 2^{i-1}$

$$J[f_{2^{i-1}-1+j}] \leq J[g_{i,j}],$$

(11) implies

$$J[f_{2^i-1}] \leq \frac{i}{4k}.$$

This completes the proof of the induction step for the induction over i. Applying (9) with $i = k+1$ implies that for all i, $J[g_{i,2^{i-1}}] \leq 1/4$, and since Lemma 4 implies that the action of $g_{i,j}$ is nondecreasing in j, this implies that for all i, j, $J[g_{i,j}] \leq 1/4$.

We claim that, for each stage i, we pass on at most half of the trials in stage i. Note that for each trial j of the ith stage in which we pass, $g_{i,j}$ has (absolute) slope at least 1 on one of the subintervals on either side of the domain element presented on that trial, thus for all $j' \geq j$ during the ith stage, $g_{i,j'}$ also has slope at least 1 on that subinterval. At the end of the ith stage, there are 2^i subintervals. If at least p trials were passed, then, integrating only over the subintervals of absolute slope at least 1 resulting from these passed trials yields

$$J[g_{i,2^{i-1}}] \geq p/2^i.$$

But $J[g_{i,2^{i-1}}] \leq 1/4$. Hence, $p \leq 2^{i-2}$. Therefore, during stage i, there must have been at least $2^{i-1} - 2^{i-2} = 2^{i-2}$ trials that were not skipped. Since, on those trials, we force A to have error at least $\frac{1}{2^{i+1}\sqrt{k}}$, the total error of algorithm A is at least

$$\sum_{i=1}^{k} 2^{i-2}\left(\frac{1}{2^{i+1}\sqrt{k}}\right) = \sqrt{k}/8.$$

This completes the proof. □

Acknowledgement

We are very grateful to an anonymous referee for finding and fixing a mistake in an earlier version of this paper, and for other helpful comments. This work was supported by United States Office of Naval Research grant N00014–94–1–0938 and National University of Singapore Academic Research Fund Grant RP960625.

References

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

[2] 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(3):604–619, 1996.

[3] V. Faber and J. Mycielski. Applications of learning theorems. *Fundamenta Informaticae*, 15(2):145–167, 1991.

[4] D. Kimber and P.M. Long. On-line learning of smooth functions of a single variable. *Theoretical Computer Science*, 148(1):141–156, 1995.

[5] G. Leitmann. *The Calculus of Variations and Optimal Control.* Plenum Press, 1981.

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

[7] J. Mycielski. A learning algorithm for linear operators. *Proceedings of the American Mathematical Society*, 103(2):547–550, 1988.

[8] H.L. Royden. *Real Analysis.* Macmillan, 1963.

A Proof of Lemma 3

We will make use of the following lemma, known as Jensen's inequality.

Lemma 9 *Choose a random variable Y and a convex function ψ. Then*

$$\mathbf{E}(\psi(Y)) \geq \psi(\mathbf{E}(Y)).$$

Proof (of Lemma 3): Assume without loss of generality that $u_1 < u_2 < ... < u_m$. Define $u_0 = 0$ and $u_{m+1} = 1$. By definition, $J[f] = \int_0^1 f'(x)^2 \, dx$, which implies

$$J[f] = \sum_{i=0}^{m} \int_{u_i}^{u_{i+1}} f'(x)^2 \, dx,$$

which in turn implies

$$J[f] = \sum_{i=0}^{m} (u_{i+1} - u_i) \left(\frac{1}{u_{i+1} - u_i} \int_{u_i}^{u_{i+1}} f'(x)^2 \, dx \right).$$

Applying Lemma 9 yields

$$J[f] \geq \sum_{i=0}^{m} (u_{i+1} - u_i) \left(\frac{1}{u_{i+1} - u_i} \int_{u_i}^{u_{i+1}} f'(x) \, dx \right)^2 .$$

Since f is absolutely continuous, this implies

$$J[f] \geq \sum_{i=0}^{m} (u_{i+1} - u_i) \left(\frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} \right)^2 . \tag{12}$$

However, since for any $x \in (u_i, u_{i+1})$,

$$f'_S(x) = (v_{i+1} - v_i)/(u_{i+1} - u_i) = (f(u_{i+1}) - f(u_i))/(u_{i+1} - u_i)$$

and $f'_S(x) = 0$ for all $x \notin [u_1, u_m]$, we have

$$J[f_S] = \sum_{i=1}^{m-1} (u_{i+1} - u_i) \left(\frac{f(u_{i+1}) - f(u_i)}{u_{i+1} - u_i} \right)^2 .$$

Combining this with (12) completes the proof. □

B Proof of Lemma 5

Assume without loss of generality that $\sum_{i=1}^m r_i = z$ and $\sum_{i=1}^m q_i^2/r_i = 1$.

Fix $r_1, ..., r_m > 0$ such that $\sum_{i=1}^m r_i = z$, and consider the problem of maximizing $\sum_{i=1}^m q_i$ subject to $\sum_{i=1}^m q_i^2/r_i = 1$. Applying Lagrange multipliers, a necessary condition for a maximum is that there is a λ such that for all i,

$$1 - 2\lambda q_i/r_i = 0.$$

Solving, we get that for each i, $q_i = r_i/(2\lambda)$, and therefore, that

$$\sum_{i=1}^m q_i = \frac{1}{2\lambda} \sum_{i=1}^m r_i. \tag{13}$$

However, substituting into the constraint yields $\sum_{i=1}^m (r_i/(2\lambda))^2/r_i = 1$, which implies $\sum_{i=1}^m r_i = 4\lambda^2$. Since $\sum_{i=1}^m r_i = z$, this implies $\lambda = \pm\sqrt{z}/2$. In (13), replacing λ with each of $\pm\sqrt{z}/2$, replacing $\sum_{i=1}^m r_i$ with z and simplifying, we see that the maximum is one of $\pm\sqrt{z}$, and therefore is $\sqrt{z}$, completing the proof. □

C Proof of Lemma 6

First, we need the following.

Claim 10 *For all $r \in [0, 1/2], \ln \frac{1}{1-r} \geq (4 \ln 2) r^2$.*

Proof: Define $g : [0, 1/2] \to \mathbf{R}$ by

$$g(r) = \ln \frac{1}{1-r} - (4\ln 2)r^2.$$

Then

$$g''(r) = \frac{1}{(1-r)^2} - 8\ln 2,$$

which is negative for all $r \in [0, 1/2]$. Thus g is minimized at 0 and 1/2, where it takes the value 0. □

Proof (of Lemma 6): By symmetry, we may assume without loss of generality that $r \le q/2$. Fix r. Define $f : [2r, \infty) \to \mathbf{R}$ by

$$f(q) = r \ln \frac{1}{r} + (q-r) \ln \frac{1}{q-r} - q \ln \frac{1}{q} - \frac{(4\ln 2)r(q-r)}{q}.$$

Then

$$f'(q) = \ln \frac{1}{1 - r/q} - (4\ln 2)(r/q)^2.$$

Applying Claim 10, we have that f' is nonnegative over the domain of f, and therefore that f is minimized when $q = 2r$, where it takes a value of 0. Dividing through by $\ln 2$ completes the proof. □