Estimating the Average of a Lipschitz-Continuous Function from One Sample

Abhimanyu Das and David Kempe

University of Southern California
{abhimand,dkempe}@usc.edu

Abstract. We study the problem of estimating the average of a Lipschitz continuous function f defined over a metric space, by querying f at only a single point. More specifically, we explore the role of randomness in drawing this sample. Our goal is to find a distribution minimizing the expected estimation error against an adversarially chosen Lipschitz continuous function. Our work falls into the broad class of estimating aggregate statistics of a function from a small number of carefully chosen samples. The general problem has a wide range of practical applications in areas such as sensor networks, social sciences and numerical analysis. However, traditional work in numerical analysis has focused on asymptotic bounds, whereas we are interested in the *best* algorithm. For arbitrary discrete metric spaces of bounded doubling dimension, we obtain a PTAS for this problem. In the special case when the points lie on a line, the running time improves to an FPTAS. For Lipschitz-continuous functions over $[0, 1]$, we calculate the precise achievable error as $1-\frac{\sqrt{3}}{2}$, which improves upon the $\frac{1}{4}$ which is best possible for deterministic algorithms.

1 Introduction

One of the fundamental problems in data-driven sciences is to estimate some aggregate statistic of a real-valued function f, by sampling f in few places. Frequently, obtaining samples incurs a cost in terms of computation, energy or time. Thus, researchers face an inherent tradeoff between the accuracy of estimating the aggregate statistic and the number of samples required. With samples a scarce resource, it becomes an important problem to determine where to sample f, and how to post-process the samples.

Naturally, there are many mathematical formulations of this estimation problem, depending on the aggregate statistic that we wish to estimate (such as the average, median or maximum value), the error objective that we wish to minimize (such as worst-case absolute error, average-case squared error, etc.), and on the conditions imposed on the function. In this paper, we study algorithms optimizing a worst-case error objective, i.e., we assume that f is chosen adversarially. Motivated by the applications described below, we use Lipschitz-continuity to impose a "smoothness" condition on f. (Note that without any smoothness conditions on f, we cannot hope to approximate any aggregate function in an adversarial setting without learning all function values.) That is, we assume that

M. de Berg and U. Meyer (Eds.): ESA 2010, Part I, LNCS 6346, pp. 219–230, 2010.

the domain of f is a metric space, and that f is Lipschitz-continuous over its domain. Thus, nearby points are guaranteed to have similar function values.

Here, we focus on perhaps the simplest aggregation function: the average $\overline{f}$. Despite its simplicity, it has many applications. For example, in sensor networks covering a geographical area, the average of a natural phenomenon (such as temperature or humidity) is frequently one of the most interesting quantities. Here, nearby locations tend to yield similar measurements. Since energy is a scarce resource, it is desirable to sample only a few of the sensors. Another application is in numerical analysis, where one of the fundamental problems is numerical integration of a function. If the domain is continuous, this corresponds precisely to computing the average. If the function to be integrated is costly to evaluate, then again, it is desirable to sample a small number of points.

If f is to be evaluated at k points, chosen deterministically and non-adaptively, then previous work [4] shows that the optimum sampling locations for estimating the average of f form a k-median of the metric space. However, the problem becomes significantly more complex when the algorithm gets to randomize its choice of sampling locations. In fact, even the seemingly trivial case of $k = 1$ turns out to be highly non-trivial, and is the focus of this paper. Addressing this case is an important step toward the ultimate goal of understanding the tradeoffs between the number of samples and the estimation error.

Formally, we thus study the following question: Given a metric space $\mathcal{M}$, a randomized *sampling algorithm* is described by (1) a method for sampling a location $x \in \mathcal{M}$ from a distribution $\mathbf{p}$; (2) a function g for predicting the average $\overline{f}$ of the function f over $\mathcal{M}$, using the sample $(x, f(x))$. The expected estimation error is then $E(\mathbf{p}, g, f) = \sum_{x \in \mathcal{M}} p_x \cdot |g(x, f(x)) - \overline{f}|$. (The sum is replaced by an integral, and $\mathbf{p}$ by a density, if $\mathcal{M}$ is continuous.) The worst-case error is $\hat{E}(\mathbf{p}, g) = \sup_{f \in L} E(\mathbf{p}, g, f)$, where L is the set of all 1-Lipschitz continuous functions defined on $\mathcal{M}$. Our goal is to find a randomized sampling algorithm (i.e., a distribution $\mathbf{p}$ and function g, computable in polynomial time) that (approximately) minimizes $\hat{E}(\mathbf{p}, g)$.

In this paper, we provide a PTAS for minimizing $\hat{E}(\mathbf{p}, g)$, for any discrete metric space $\mathcal{M}$ with constant doubling dimension. (This includes constant-dimensional Euclidean metric spaces.) For discrete metric spaces $\mathcal{M}$ embedded on a line, we improve this result to an FPTAS. Both of these algorithms are based on a linear program with infinitely many constraints, for which an approximate separation oracle is obtained.

We next study the perhaps simplest variant of this problem, in which the metric space is the interval $[0, 1]$. While the worst-case error of any deterministic algorithm is obviously $\frac{1}{4}$ in this case, we show that for a randomized algorithm, the bound improves to $1 - \frac{\sqrt{3}}{2}$. We prove this by providing an explicit distribution, and obtaining a matching lower bound using Yao's Minimax Principle. Our result can also be interpreted as showing how "close" a collection of Lipschitz-continuous functions on $[0, 1]$ can be.

Due to space constraints for this version, several discussions and proofs are relegated to a full version, which will be available on the authors' web sites.

1.1 Related Work

Estimating the integral of a smooth function f using its values at a discrete set of points is one of the core problems in numerical analysis. The tradeoffs between the number of samples needed and the estimation error bounds have been investigated in detail under the name of *Information Based Complexity (IBC)* [10,11]. More generally, IBC studies the problem of computing approximations to an operator $S(f)$ on functions f from a set F (with certain "smoothness" properties) using a finite set of samples $N(f) = [L_1(f), L_2(f), \ldots, L_n(f)]$. The L_i are functionals. For a given algorithm U, its error is $E(U) = \sup_{f \in F} \|S(f) - U(f)\|$. The goal in IBC is to find an ϵ-approximation U (i.e., ensuring that $E(U) \leq \epsilon$) with least information cost $c(U) = n$.

One of the common problems in IBC is multivariate integration of real-valued functions with a smoothness parameter r over d-dimensional unit balls. For such problems, Bakhvalov [2] designed a randomized algorithm providing an ϵ-approximation with cost $\Theta(\frac{1}{n^{2d/(d+2r)}})$. Bakhvalov [2], and later, Novak [9] also show that this cost is asymptotically optimal. The papers by Novak [9] and Mathe [7] show that if $r = 0$, then simple Monte-Carlo integration algorithms (which sample from the uniform distribution) have an asymptotically optimal cost of $\frac{1}{\epsilon^2}$. In [12,13], Wozniakowski studied the average case complexity of linear multivariate IBC problems, and proved conditions under which the problems are tractable, i.e., have cost polynomial in $\frac{1}{\epsilon}$ and d.

In [3], Baran et al. study the IBC problem for univariate integration of Lipschitz continuous functions, in an adaptive setting. That is, the sampling strategy can change adaptively based on the previously sampled values. They provide a deterministic and a randomized ϵ-approximation algorithm, which use $O(\log(\frac{1}{\epsilon \cdot \mathrm{OPT}}) \cdot \mathrm{OPT})$ and $O(\mathrm{OPT}^{4/3} + \mathrm{OPT} \cdot \log(\frac{1}{\epsilon}))$ samples, respectively. Here, OPT is the optimal number of samples for the problem instance. They prove that their algorithms are asymptotically optimal.

There are two main differences between the results in IBC and our work: first, IBC treats the target approximation as given and minimizes the number of samples. Our goal is to minimize the expected worst-case error with a fixed number of samples (one). More importantly, results in IBC are traditionally *asymptotic*, ignoring constants. For a single sample, this would trivialize the problem: it is implicit in our proofs that sampling at the metric space's median is a constant-factor approximation to the best randomized algorithm.

The deterministic version of our problem was studied previously in [4]. There, it was shown that the best sampling locations for reading k values non-adaptively are the optimal k-median of the metric space. Thus, the algorithm of Arya et al. [1] gives a polynomial-time $(3 + \epsilon)$-approximation algorithm.

2 Preliminaries

We are interested in real-valued Lipschitz-continuous functions over metric spaces of constant doubling dimension (e.g., [6]). Let $(\mathcal{M}, d)$ be a compact metric space with distances $d(x, y)$ between pairs of points. W.l.o.g., we assume that

$\max_{x,y\in\mathcal{M}} d(x,y) = 1$. We require $(\mathcal{M}, d)$ to have constant doubling dimension β, i.e., for every δ, each ball of diameter δ can be covered by at most c^β balls of diameter δ/c, for any $c \geq 2$.

A function f is Lipschitz continuous (with constant 1) if for all x, y, we have $|f(x)-f(y)| \leq d(x,y)$. Let L be the set of all such Lipschitz-continuous functions f, i.e., $L = \{f \mid |f(x) - f(y)| \leq d(x,y) \text{ for all } x, y\}$. We also define $L_c = \{f \in L \mid |\int_x f(x)dx| \leq c\}$. Notice that L_c is a compact set.

We wish to predict the average $\overline{f} = \int_x f(x)dx$ of all the function values. When $\mathcal{M}$ is finite of size n, then the average is of course $\overline{f} = \frac{1}{n} \cdot \sum_x f(x)$ instead. The algorithm first gets to choose a single point x according to a (polynomial-time computable) density function $\mathbf{p}$; it then learns the value $f(x)$, and may post-process it with a *prediction function* $g(x, f(x))$ to produce its estimate of the average $\overline{f}$. The goal is to minimize the expected estimation error of the average, under the assumption that f is chosen adversarially from L with knowledge of the algorithm, but not its random choices. Formally, the goal is to minimize $\hat{E}(\mathbf{p}, g) = \sup_{f\in L}(\int_x p_x \cdot |\overline{f} - g(x, f(x))|dx)$. If $\mathcal{M}$ is finite, then $\mathbf{p}$ will be a probability distribution instead of a density, and the error now can be written as $\hat{E}(\mathbf{p}, g) = \sup_{f\in L}(\sum_x p_x \cdot |\overline{f} - g(x, f(x))|)$.

Formally, we consider an algorithm to be the pair $(\mathbf{p}, g)$ of the distribution and prediction function. Let $\mathcal{A}$ denote the set of all such pairs, and $\mathcal{D}$ the set of all *deterministic* algorithms, i.e., algorithms for which $\mathbf{p}$ has all its density on a single point. Our analysis will make heavy use of Yao's Minimax Principle [8]. To state it, we define $\mathcal{L}$ to be the set of all probability distributions over L. We also define the estimation error $\Delta(f, A) = \int_x p_x \cdot |\overline{f} - g(x, f(x))|dx$, where A corresponds to the pair $(\mathbf{p}, g)$.

Theorem 1 (Yao's Minimax Principle [8])

$$\sup_{q\in\mathcal{L}} \inf_{A\in\mathcal{D}} \mathrm{E}_{f\sim q}\left[\Delta(f, A)\right] = \inf_{A\in\mathcal{A}} \sup_{f\in L} \Delta(f, A).$$

The next theorem shows that without loss of generality, we can focus on algorithms whose post-processing is just to output the observed value, i.e., algorithms $(\mathbf{p}, \mathrm{id})$ with $\mathrm{id}(x, y) = y$, for all x, y.

Theorem 2. *Let $A^* = (\mathbf{p}^*, g^*)$ be the optimum randomized algorithm. Then, for every $\epsilon > 0$, there is a randomized algorithm $A = (\mathbf{p}, \mathrm{id})$ with $\hat{E}(A) \leq \hat{E}(A^*) + \epsilon$.*

3 Discrete Metric Spaces

In this section, we focus on finite metric spaces, consisting of n points. Thus, instead of integrals and densities, we will be considering sums and probability distributions. The result from Theorem 2 holds in this case as well; hence w.l.o.g., we assume that all algorithms simply output the value they observe. The problem of finding the best probability distribution for a single sample can be expressed as a linear program, with variables p_x for the sampling probabilities at each of the n points x, and a variable Z for the estimation error.

$$\begin{array}{lll} \text{Minimize} & Z & \\ \text{subject to} & \text{(i)} \ \ \sum_x p_x = 1 & \\ & \text{(ii)} \ \ \sum_x p_x \cdot |\overline{f} - f(x)| \le Z & \text{for all } f \in L \\ & \text{(iii)} \ 0 \le p_x \le 1 & \text{for all points } x \end{array} \tag{1}$$

Since this LP (which we refer to as the "exact LP") has infinitely many constraints, our approach is to replace the set L in the second constraint with a set Q_δ. We will choose Q_δ carefully to ensure that it "approximates" L well, and such that the resulting LP (which we refer to as the "discretized LP") can be solved efficiently.

To define the notion of approximation formally, let o be a 1-median of the metric space, i.e., a point minimizing $\sum_x d(o,x)$. Let $m = \frac{1}{n}\sum_x d(o,x)$ be the average distance of all points from o. Since w.l.o.g. $\max_{x,y \in \mathcal{M}} d(x,y) = 1$, at least one point has distance at least $\frac{1}{2}$ from o, and thus $m \ge \frac{1}{2n}$. The median value m forms a lower bound for randomized algorithms in the following sense.

Lemma 1. *The worst-case expected error for any randomized algorithm is at least $\frac{1}{4 \cdot 6^\beta} \cdot m$, where β is the doubling dimension of the metric space.*

Proof. Consider any randomized algorithm with probability distribution $\mathbf{p}$. Let $R = \{x \mid \frac{m}{2} \le d(x,o) \le \frac{3m}{2}\}$ be the ring of points at distance between $\frac{m}{2}$ and $\frac{3m}{2}$ from o. We distinguish two cases:

1. If $\sum_{x \in R} p_x \le \frac{1}{2}$, consider the Lipschitz-continuous function $f(x) = d(x,o)$. Then, $\overline{f} = m$. With probability at least $\frac{1}{2}$, the algorithm samples a point outside R, and thus outputs a value outside the interval $[\frac{m}{2}, \frac{3m}{2}]$, which incurs error at least $\frac{m}{2}$. Thus, the expected error is at least $\frac{m}{4}$.
2. If $\sum_{x \in R} p_x > \frac{1}{2}$, then consider a collection of balls $B_1, \ldots, B_k$ of diameter $\frac{m}{2}$ covering all points in R. Because R is contained in a ball of diameter $3m$, the doubling constraint implies that $k \le 6^\beta$ balls are sufficient. At least one of these balls — say, B_1 — has $\sum_{x \in B} p_x \ge \frac{1}{2k}$. Fix an arbitrary point $y \in B_1$, and define the Lipschitz-continuous function f as $f(x) = d(x,y)$. Because o was a 1-median, we get that $\overline{f} \ge m$. With probability at least $\frac{1}{2k}$, the algorithm will choose a point inside B_1 and output a value of at most $\frac{m}{2}$, thus incurring an error of at least $\frac{m}{2}$. Hence, the expected error is at least $\frac{1}{2k} \cdot \frac{m}{2} \ge \frac{1}{4 \cdot 6^\beta} \cdot m$. ■

We now formalize our notion for a set of functions Q_δ to be a good approximation.

Definition 1 (δ-approximating function classes). *For any sampling distribution $\mathbf{p}$, let $E_L(\mathbf{p}) = \max_{f \in L} \Delta(f, \mathbf{p})$ and $E_Q(\mathbf{p}) = \max_{f \in Q_\delta} \Delta(f, \mathbf{p})$ be the maximum error of sampling according to $\mathbf{p}$ against a worst-case function from L and Q_δ, respectively, where $\Delta(f, \mathbf{p}) = \sum_x p_x \cdot |\overline{f} - f(x)|$. The class Q_δ δ-*approximates *L if*

1. *For each $f \in L$, there is a function $f' \in Q_\delta$ such that for all distributions $\mathbf{p}$, we have $|\Delta(f', \mathbf{p}) - \Delta(f, \mathbf{p})| \le \frac{\delta}{2} \cdot E_L(\mathbf{p})$.*

2. *For each $f \in Q_\delta$, there is a function $f' \in L$ such that for all distributions $\mathbf{p}$, we have $|\Delta(f', \mathbf{p}) - \Delta(f, \mathbf{p})| \le \frac{\delta}{2} \cdot E_L(\mathbf{p})$.*

Theorem 3. *Assume that for every δ, Q_δ is a class of functions δ-approximating L, such that the following problem can be solved in polynomial time (for fixed δ): Given $\mathbf{p}$, find a function $f \in Q_\delta$ maximizing $\Delta(f, \mathbf{p})$.*

Then, solving the discretized LP gives a PTAS for the problem of finding a sampling distribution that minimizes the worst-case expected error.

Proof. First, an algorithm to find a function f maximizing $\sum_x p_x \cdot |\overline{f} - f(x)|$ gives a separation oracle for the discretized LP. Thus, using the Ellipsoid Method (e.g., [5]), an optimal solution to the discretized LP can be found in polynomial time, for any fixed δ.

Let $\mathbf{p}$, $\mathbf{q}$ be optimal solutions to the exact and discretized LPs, respectively. Let $f_1 \in L$ maximize $\sum_x q_x \cdot |\overline{f} - f(x)|$ over $f \in L$, and $f_2 \in Q_\delta$ maximize $\sum_x p_x \cdot |\overline{f} - f(x)|$ over $f \in Q_\delta$. Thus, $\Delta(f_1, \mathbf{q}) = E_L(\mathbf{q})$ and $\Delta(f_2, \mathbf{p}) = E_Q(\mathbf{p})$.

Now, applying Definition 1 to $f_1 \in L$ gives us a function $f_1' \in Q_\delta$ such that $|\Delta(f_1', \mathbf{q}) - E_L(\mathbf{q})| \le \frac{\delta}{2} E_L(\mathbf{q})$. Since $E_Q(\mathbf{q}) \ge \Delta(f_1', \mathbf{q})$, we obtain that $E_Q(\mathbf{q}) \ge E_L(\mathbf{q})(1 - \frac{\delta}{2})$.

Similarly, applying Definition 1 to $f_2 \in Q_\delta$, gives us a function $f_2' \in L$ with $|\Delta(f_2', \mathbf{p}) - E_Q(\mathbf{p})| \le \frac{\delta}{2} E_L(\mathbf{p})$. Since $E_L(\mathbf{p}) \ge \Delta(f_2', \mathbf{p})$, we obtain that $E_L(\mathbf{p}) \ge E_Q(\mathbf{p}) - \frac{\delta}{2} E_L(\mathbf{p})$, or $E_L(\mathbf{p}) \ge \frac{E_Q(\mathbf{p})}{1+\frac{\delta}{2}}$. Also, by optimality of $\mathbf{q}$ in Q_δ, $E_Q(\mathbf{q}) \le E_Q(\mathbf{p})$. Thus, $E_L(\mathbf{q}) \le \frac{E_Q(\mathbf{q})}{1-\frac{\delta}{2}} \le \frac{E_Q(\mathbf{p})}{1-\frac{\delta}{2}} \le \frac{E_L(\mathbf{p})(1+\frac{\delta}{2})}{1-\frac{\delta}{2}} \le E_L(\mathbf{p})(1 + 2\delta)$. ■

3.1 A PTAS for Arbitrary Metric Spaces

We first observe that since the error for any translation of a function f is the same as for f, we can assume w.l.o.g. that $f(o) = 0$ for all functions f considered in this section. Thus, we implicitly restrict L to functions with $f(o) = 0$.

We next describe a set Q_δ of functions which δ-approximate L. Roughly, we will discretize function values to different multiples of γ, and consider distance scales that are different multiples of γ. We later set $\gamma = \frac{\delta}{48 \cdot 6^\beta + 6}$. We show in Lemma 2 that Q_δ has size $n^{\log(2/\gamma)(2/\gamma)^\beta} = n^{O(1)}$ for constant δ; this immediately implies that the discretized LP can be solved in time $O(\text{poly}(n) \cdot n^{\log(2/\gamma)(2/\gamma)^\beta})$ (using exhaustive search for the separation oracle), and we obtain a PTAS for finding the optimum distribution.

We let $k = \log_2 \frac{1}{2m}$, and define a sequence of k rings of exponentially decreasing diameter around o, dividing the space into $k+1$ regions $R_1, \ldots, R_{k+1}$. Specifically, we let $R_{k+1} = \{x \mid d(x, o) \le 2m\}$, and $R_i = \{x \mid 2^{-i} < d(x, o) \le 2^{-(i-1)}\}$ for $i = 1, \ldots, k$. Since $m \ge \frac{1}{2n}$, we have that $k \le \log n$.

Since the metric space has doubling dimension β, each region R_i can be covered with at most $(2/\gamma)^\beta$ balls of diameter $2\gamma \cdot 2^{-i}$. Let $B_{i,j}$ denote the j^{th} ball from the cover of R_i. W.l.o.g., each $B_{i,j}$ is non-empty and contained in R_i (otherwise, consider its intersection with R_i instead). We call $B_{i,j}$ the j^{th} *grid* *ball* for region i. Thus, the grid balls cover all points, and there are at most $(2/\gamma)^\beta \cdot \log n$ grid balls.

For each grid ball $B_{i,j}$, let $o_{i,j} \in B_{i,j}$ be an arbitrary, but fixed, *representative* of $B_{i,j}$. The exception is that for the grid ball containing o, o must be chosen as the representative. We now define the class Q_δ of functions f as follows:

1. For each i, j, $f(o_{i,j})$ is a multiple of $\gamma \cdot 2^{-i}$.
2. For all $(i,j), (i',j')$, the function values satisfy the *relaxed Lipschitz-condition* $|f(o_{i,j}) - f(o_{i',j'})| \le d(o_{i,j}, o_{i',j'}) + \gamma \cdot (2^{-i} + 2^{-i'})$.
3. All points in $B_{i,j}$ have the same function value, i.e., $f(x) = f(o_{i,j})$ for all $x \in B_{i,j}$.

Lemma 2. *The size of Q_δ is at most $n^{\log(2/\gamma)(2/\gamma)^\beta}$.*

We need to prove that Q_δ approximates L well, by verifying that for each function $f \in L$, there is a "close" function in Q_δ, and vice versa. We first show that for any function satisfying the relaxed Lipschitz condition, we can change the function values slightly and obtain a Lipschitz continuous function. In Lemma 4, we then apply this result specifically to functions in Q_δ. Finally, in Lemma 5 (whose proof is deferred to the full version, due to space constraints), we show the converse approximation direction.

Lemma 3. *For each $x \in \mathcal{M}$, let s_x be some non-negative number. Assume that f satisfies the "relaxed Lipschitz condition" $|f(x) - f(y)| \le d(x,y) + s_x + s_y$ for all x, y. Then, there is a Lipschitz continuous function $f' \in L$ such that $|f(x) - f'(x)| \le s_x$ for all x.*

Proof. We describe an algorithm which runs in iterations ℓ, and sets the value of one point x per iteration. S_ℓ denotes the set of x such that $f'(x)$ has been set. We maintain the following two invariants after the ℓ^{th} iteration: 1) f' satisfies the Lipschitz condition for all pairs of points in S_ℓ, and $|f'(x) - f(x)| \le s_x$ for all $x \in S_\ell$, and 2) For every function f'' satisfying the previous condition, $f'(x) \le f''(x)$ for all $x \in S_\ell$.

Initially, this clearly holds for $S_0 = \emptyset$. And clearly, if it holds after iteration n, the function f' satisfies the claim of the lemma.

In iteration ℓ, for each $x \notin S_{\ell-1}$, let $t_x = \max_{y \in S_{\ell-1}}(f'(y) - d(x,y))$. We show below that for all x, we have $t_x \le f(x) + s_x$. Let $x \notin S_{\ell-1}$ be a point maximizing $\max(f(x) - s_x, t_x)$, and set $f'(x) = \max(f(x) - s_x, t_x)$. It is easy to verify that this definition satisfies both parts of the invariant.

It remains to show that $t_x \le f(x) + s_x$ for all points $x \notin S_{\ell-1}$. Assume that $t_x > f(x) + s_x$ for some point x. Let x_1 be the point in $S_{\ell-1}$ for which $t_x = f'(x_1) - d(x, x_1)$. By definition, $f'(x_1) = f(x_1) - s_{x_1}$ or there is an x_2 such that $f'(x_1) = t_{x_1} = f'(x_2) - d(x_1, x_2)$. Thus, we obtain a chain $x_1, \ldots, x_r$ with $f'(x_i) = f'(x_{i+1}) - d(x_i, x_{i+1})$ for all $i < r$, and $f'(x_r) = f(x_r) - s_{x_r}$. Rearranging as $f'(x_{i+1}) - f'(x_i) = d(x_i, x_{i+1})$, and adding all these equalities for $i = 1, \ldots, r$ gives us that $f(x_r) - f'(x_1) = s_{x_r} + \sum_{i=1}^{r-1} d(x_i, x_{i+1})$. By assumption, we have $f'(x_1) - d(x, x_1) = t_x > f(x) + s_x$. Substituting the previous equality, rearranging, and applying the triangle inequality gives us that $f(x_r) - f(x_1) > s_x + s_{x_r} + d(x, x_1) + \sum_{i=1}^{r-1} d(x_i, x_{i+1}) \geq s_x + s_{x_r} + d(x, x_r)$, which contradicts the relaxed Lipschitz condition for the pair x_1, x_r. ∎

Lemma 4. *Let $f \in Q_\delta$. There exists an $f' \in L$ such that for all distributions $\mathbf{p}$, we have $|\Delta(f, \mathbf{p}) - \Delta(f', \mathbf{p})| \leq \frac{\delta}{2} \cdot E_L(\mathbf{p})$.*

Proof. Because f is in Q_δ, it must satisfy the relaxed Lipschitz condition $|f(o_{i,j}) - f(o_{i',j'})| \leq d(o_{i,j}, o_{i',j'}) + \gamma \cdot (2^{-i} + 2^{-i'})$ for all $(i, j), (i', j')$. Thus, applying Lemma 3 with $s_{o_{i,j}} = \gamma \cdot 2^{-i}$ gives us function values $f'(o_{i,j})$ for all i, j, satisfying the Lipschitz condition, as well as $f'(o_{i,j}) - f(o_{i,j}) \leq \gamma \cdot 2^{-i}$. For any other point x, let $L_{\max}(x, f) = \min_{i,j}(f'(o_{i,j}) + d(x, o_{i,j}))$ and $L_{\min}(x, f) = \max_{i,j}(f'(o_{i,j}) - d(x, o_{i,j}))$, and set $f'(x) = \frac{1}{2} \cdot (L_{\max}(x, f) + L_{\min}(x, f))$. It is easy to see that $L_{\min}(x, f) \leq L_{\max}(x, f)$ for all x, and that this definition gives a Lipschitz continuous function f'. For a point $x \in B_{i,j}$, triangle inequality, the above construction, and the fact that $B_{i,j}$ has diameter $2\gamma \cdot 2^{-i}$ imply that $|f'(x) - f(x)| \leq |f'(x) - f'(o_{i,j})| + |f'(o_{i,j}) - f(o_{i,j})| + |f(o_{i,j}) - f(x)| \leq 2\gamma \cdot 2^{-i} + \gamma \cdot 2^{-i} + 0 = 3\gamma \cdot 2^{-i}$.

For each point x, let $\iota(x)$ be the index of the region i such that $x \in R_i$. Now, using the triangle inequality and Lemma 1, we can bound $|\overline{f'} - \overline{f}| \leq \frac{1}{n} \cdot \sum_x |f'(x) - f(x)| \leq \frac{1}{n} \cdot \sum_x 3\gamma \cdot 2^{-\iota(x)} \leq \frac{1}{n} \cdot (\sum_{x \notin R_{k+1}} 3\gamma \cdot d(x, o) + \sum_{x \in R_{k+1}} 3\gamma \cdot m) \leq \frac{1}{n} \cdot (3\gamma nm + 3\gamma nm) \leq 24 \cdot 6^\beta \cdot \gamma \cdot E_L(\mathbf{p})$.

Similarly, we can bound $\sum_x p_x \cdot |f'(x) - f(x)| \leq 3\gamma \cdot (\sum_{x \notin R_{k+1}} p_x \cdot d(x, o) + \sum_{x \in R_{k+1}} p_x m) \leq 3\gamma \cdot (m + \sum_{x \notin R_{k+1}} p_x \cdot d(x, o))$.

Let f'' be defined as $f''(x) = d(x, o)$. Clearly, $f'' \in L$, $\overline{f''} = m$, and the estimation error for $\mathbf{p}$ when the input is f'' is $\Delta(f'', \mathbf{p}) = \sum_x p_x \cdot |f''(x) - m| \geq \sum_{x \notin R_{k+1}} p_x \cdot |d(x, o) - m| \geq (\sum_{x \notin R_{k+1}} p_x \cdot d(x, o)) - m$.

Combining these observations, and using Lemma 1 and the fact that $\Delta(f'', \mathbf{p}) \leq E_L(\mathbf{p})$, we get $\sum_x p_x \cdot |f'(x) - f(x)| \leq 6\gamma \cdot m + 3\gamma \Delta(f'', \mathbf{p}) \leq (8 \cdot 6^\beta + 1) \cdot 3\gamma \cdot E_L(\mathbf{p})$.

Now, by using that $|\Delta(f, \mathbf{p}) - \Delta(f', \mathbf{p})| \leq |\overline{f'} - \overline{f}| + \sum_x p_x \cdot |f'(x) - f(x)|$, and setting $\gamma = \frac{\delta}{48 \cdot 6^\beta + 6}$, we obtain the desired bound. ∎

Lemma 5. *Let $f \in L$. There exists an $f' \in Q_\delta$ such that for all distributions $\mathbf{p}$, we have $|\Delta(f, \mathbf{p}) - \Delta(f', \mathbf{p})| \leq \frac{\delta}{2} \cdot E_L(\mathbf{p})$.*

If the metric consists of a discrete point set on the line, then the PTAS can be improved to an FPTAS, as discussed in the full version of the paper.

4 Sampling in the Interval $[0, 1]$

In this section, we focus on what is probably the most basic version of the problem: the metric space is the interval $[0, 1]$. It is easy to see (and follows from a more general result in [4]) that the best deterministic algorithm samples the function at $\frac{1}{2}$ and outputs the value read. The worst-case error of this algorithm is $\frac{1}{4}$. We prove that randomization can lead to the following improvement.

Theorem 4. *An optimal distribution that minimizes the worst-case expected estimation error is to sample uniformly from the interval* $[2-\sqrt{3}, \sqrt{3}-1]$*. This sampling gives a worst-case error of* $1-\frac{\sqrt{3}}{2} \approx 0.134$*.*

In this section, we restrict our analysis w.l.o.g. to functions $f \in L_0$, i.e., we assume that $\int_0^1 f(x)dx = 0$. Then, the expected error of a distribution $\mathbf{p}$ against input f is $\Delta(f, \mathbf{p}) = \int_0^1 p_x|f(x)|dx$. We say that f is a *worst-case function for* $\mathbf{p}$ if it maximizes $\Delta(f, \mathbf{p})$; because L_0 is compact, this notion is well-defined.

The key part of the proof of Theorem 4 is to characterize worst-case functions for distributions $\mathbf{p}$ that are uniform over an interval $[c, 1-c]$ for some $c \leq \frac{1}{2}$.

Theorem 5. *If* $\mathbf{p}$ *is uniform over* $[c, 1-c]$*, then there exists a worst-case function for* $\mathbf{p}$ *of the form* $f(x) = \frac{1}{2} + b^2 - b - |b-x|$*, for a parameter* b*.*

All of Section 4.1 is devoted to the proof of Theorem 5. Here, we show how to use Theorem 5 to prove the upper bound from Theorem 4.

Let $c = 2-\sqrt{3}$, so that the algorithm samples uniformly from $[c, 1-c]$. Using Theorem 5, there exists a worst-case function for this distribution of the form $f(x) = \frac{1}{2} + b^2 - b - |b-x|$. We distinguish two cases:

1. If $b \leq c$, then $\Delta(f, \mathbf{p}) = \frac{1}{1-2c} \cdot \int_c^{1-c} |\frac{1}{2} + b^2 - b - |b-x||dx = \frac{1}{1-2c} \cdot (\frac{1}{2}(b^2 + \frac{1}{2} - c)^2 + \frac{1}{2}(1-c-b^2)^2) = \frac{1}{1-2c} \cdot (b^4 + (\frac{1}{2}-c)^2)$.
2. If $b \geq c$, then $\Delta(f, \mathbf{p}) = \frac{1}{1-2c} \cdot \int_c^{1-c} |\frac{1}{2} + b^2 - b - |b-x||dx = \frac{1}{1-2c} \cdot (2bf(b) + 2cb + f(b)^2 - c - b^2) = \frac{1}{1-2c}(b^4 - b^2 + 2cb + \frac{1}{4} - c) = \frac{1}{1-2c}(b^4 - (b-c)^2 + (\frac{1}{2}-c)^2)$.

The first formula is increasing in b, and thus maximized at $b = c$; at $b = c$, the value equals that of the second formula, so the maximization must happen for $b \geq c$. A derivative test shows that it is maximized for $b = \frac{\sqrt{3}-1}{2}$, giving an error of $1 - \frac{\sqrt{3}}{2}$.

Next, we prove optimality of the uniform distribution over $[2-\sqrt{3}, \sqrt{3}-1]$, by providing a lower bound on all randomized sampling distributions. Again, by Theorem 2, we focus only on algorithms which output the value $f(x)$ after sampling at x, by incurring an error $\epsilon > 0$ that can be made arbitrarily small. Our proof is based on Yao's Minimax principle: we explicitly prescribe a distribution q over L_0 such that for any deterministic algorithm using the identity function, the expected estimation error is at least $1-\frac{\sqrt{3}}{2}$. Since a deterministic algorithm is characterized completely by its sampling location x, this is equivalent to showing that $\mathrm{E}_{f\sim q}[|f(x)|] \geq 1 - \frac{\sqrt{3}}{2}$ for all x.

We let $b = \frac{\sqrt{3}-1}{2}$, and define two functions f, f' as $f(x) = \frac{1}{2} + b^2 - b - |x-b|$ and $f'(x) = f(1-x)$. The distribution q is then simply to choose each of f and f' with probability $\frac{1}{2}$. Fix a sampling location x; by symmetry, we can restrict ourselves to $x \leq \frac{1}{2}$. Because $\overline{f} = \overline{f'} = 0$, the expected estimation error is $\frac{1}{2}(|f(x)| + |f'(x)|) = \frac{1}{2}(|\frac{1}{2} + b^2 - b - |x-b|| + |\frac{1}{2} + b^2 - b - |1-x-b||) =$

$$\begin{cases} \frac{1}{2} - b, & \text{if } x \leq b \\ \frac{1}{2} - x, & \text{if } b \leq x \leq \frac{1}{2} - b^2 \\ b^2, & \text{if } \frac{1}{2} - b^2 \leq x \leq \frac{1}{2}. \end{cases}$$

This function is non-increasing in x, and thus minimized at $x = \frac{1}{2}$, where its value is $b^2 = 1 - \frac{\sqrt{3}}{2}$. Thus, even at the best sampling location $x = \frac{1}{2}$, the error cannot be less than $1 - \frac{\sqrt{3}}{2}$. This completes the proof of Theorem 4. ■

The proof of Theorem 4 has an interesting alternative interpretation. For a (finite) multiset $S \subset L_0$ of Lipschitz continuous functions f with $\int_x f(x)dx = 0$, we say that S is *δ-close* if there exist x, y such that $\frac{1}{n} \cdot \sum_{f \in S} |f(x) - y| \leq \delta$. In other words, the average distance of the functions from a carefully chosen reference point is at most δ. Then, the proof of Theorem 4 implies:

Theorem 6. *Every set $S \subseteq L_0$ is $(1 - \frac{\sqrt{3}}{2})$-close, and this is tight.*

4.1 Characterization of Worst-Case Functions

We begin with the following lemma which guarantees that there exists a worst case function f with a finite number of points x such that $f(x) = 0$.

Lemma 6. *W.l.o.g., there are a finite number of points x such that $f(x) = 0$.*

We focus on points $x \in (c, 1-c)$ with $f(x) = 0$. Let $c \leq z_1 \leq \ldots \leq z_k \leq 1 - c$ be all such points. For ease of notation, we write $z_0 = c$ and $z_{k+1} = 1 - c$. By continuity, $f(x)$ has the same sign for all $x \in (z_i, z_{i+1})$, for $i = 0, \ldots, k$. Next, we show that w.l.o.g., f is as large as possible over areas of the same sign.

Lemma 7. *Assume w.l.o.g. that $f(x) \geq 0$ for all $x \in [z_i, z_j]$, with $j > i$. Then, w.l.o.g., f maximizes the area over $[z_i, z_j]$ subject to the Lipschitz constraint and the function values at z_i and z_j. More formally, w.l.o.g., f satisfies,*

1. *If $1 \leq i < j \leq k$, then $f(x) = \min(x - z_i, z_j - x)$ for all $x \in [z_i, z_j]$.*
2. *If $i = 0$, then $f(x) = \min(f(c) + (x - c), z_1 - x)$ for all $x \in [c, z_1]$, and if $i = k$, then $f(x) = \min(f(1-c) + (1-c) - x, x - z_k)$ for all $x \in [z_k, 1-c]$.*

Proof. We prove the first part here; the second is analogous and proved in the full version. Define a function f' as $f'(x) = \min(x - z_i, z_j - x)$ for $x \in [z_i, z_j]$, and $f'(x) = f(x)$ otherwise. Let $f'' = f' - \overline{f'}$, so that f'' is renormalized to have integral 0. Since $f'(x) \geq f(x)$ for all x, and $\overline{f} = 0$, we have that $\overline{f'} \geq 0$. Then

$$\begin{aligned}
&\textstyle\int_c^{1-c} |f''(x)| - |f(x)|dx \\
&= \textstyle\int_{z_i}^{z_j} |f''(x)| - |f(x)|dx + \int_c^{z_i} |f(x) - \overline{f'}| - |f(x)|dx + \int_{z_j}^{1-c} |f(x) - \overline{f'}| - |f(x)|dx \\
&\geq \textstyle\int_{z_i}^{z_j} (|f'(x) - \overline{f'}| - |f'(x)|) + (|f'(x)| - |f(x)|)dx - (1 - 2c - (z_j - z_i))\overline{f'} \\
&\geq \textstyle\int_{z_i}^{z_j} |f'(x)| - |f(x)|dx - \int_{z_i}^{z_j} \overline{f'}dx - (1 - 2c - (z_j - z_i))\overline{f'} \\
&= \textstyle\int_{z_i}^{z_j} f'(x) - f(x)dx - (1 - 2c)\overline{f'} = 2c \cdot \overline{f'} \geq 0.
\end{aligned}$$

Thus, the estimation error of f'' is at least as large as the one for f, so w.l.o.g., f satisfies the statement of the lemma. ■

Lemma 8. *W.l.o.g., there are at most two points $x \in (c, 1-c)$ where $f(x) = 0$.*

Proof. Assume that $f(z_1) = f(z_2) = f(z_3) = 0$. Mirror the function on the interval $[z_1, z_3]$, i.e., define $f'(x) = f(z_3 - x)$ if $x \in [z_1, z_3]$, and $f'(x) = f(x)$

otherwise. Clearly, f' is Lipschitz continuous and has the same average and same expected estimation error as f. However, the signs of f' on the intervals $[c, z_1]$ and $[z_1, z_1 + z_3 - z_2]$ are now the same; similarly for the intervals $[z_1 + z_3 - z_2, z_3]$ and $[z_3, 1 - c]$. Thus, applying Lemma 7, we can further reduce the number of x with $f(x) = 0$, without decreasing the estimation error. ■

Hence, the worst-case function f must have at most two points $z \in (c, 1 - c)$ with $f(z) = 0$. We distinguish three cases accordingly:

1. If there is no point $z \in (c, 1 - c)$ with $f(z) = 0$, then $f(c)$ and $f(1 - c)$ have the same signs. Then, the expected error is maximized when $\int_0^c f(x)dx$ and $\int_{1-c}^1 f(x)dx$ are as positive as possible, subject to the Lipschitz condition and the constraint that $\int_0^1 f(x)dx = 0$. Otherwise, we could increase the value of $\int_0^c f(x)dx$ and $\int_{1-c}^1 f(x)dx$, and then lower the function to restore the integral to 0. By doing this, the expected estimation error cannot decrease. Thus, by Lemma 7, f is of the form $f(x) = |x - b| + f(b)$, where $b = \text{argmin}_{x\in(c,1-c)} f(x)$.
2. If there is exactly one point $z \in (c, 1 - c)$ with $f(z) = 0$, then $f(c)$ and $f(1 - c)$ have opposite signs. W.l.o.g., assume that $f(c) > 0 > f(1 - c)$ and that $z \le \frac{1}{2}$ (otherwise, we consider $f'(x) = f(1 - x)$ instead). The expected error is maximized when $f(c)$ is as large as possible, and $\int_z^{1-c} f(x)dx$ is as negative as possible, subject to the Lipschitz condition and the constraint that $\int_0^1 f(x)dx = 0$. Since $z \le \frac{1}{2}$ and the integral of the function $f'(x) = z - x$ is thus negative, by starting from f', then raising the function in the interval $[1 - c, 1]$ and, if necessary, increasing $f'(1 - c)$, it is always possible to ensure that $f(x) = z - x$ for all $x \in [0, z]$. Then, $\int_z^{1-c} f(x)dx$ is as negative as possible if $f(x) = -(x - z)$ for $x \le b$ (for some value b), and $f(x) = -(b - z) + (x - b) = z + x - 2b$ for $x \ge b$. Thus, f overall is of the form $f(x) = |x - b| - (b - z)$.
3. If there are two points $z_1 < z_2 \in (c, 1 - c)$ with $f(z_1) = f(z_2) = 0$, then again, it can be shown that w.l.o.g $f(x) = |x - \frac{z_1+z_2}{2}| - \frac{z_2-z_1}{2}$. Due to space constraints, the formal proof is deferred to the full version of the paper.

In all three cases, we have thus shown that w.l.o.g., $f(x) = |x - b| - t$, for some values b, t. Finally, the normalization $\int_0^1 f(x)dx = 0$ implies that $t = \frac{1}{2} + b^2 - b$, completing the proof of Theorem 5.

5 Future Work

Our work is a first step toward obtaining optimal (as opposed to asymptotically optimal) randomized algorithms for choosing k sample locations to estimate an aggregate quantity of a function f. The most obvious extension is to extend our results to the case of estimating the average using k samples. It would be interesting whether approximation guarantees for the k-median problem (the deterministic counterpart) can be exceeded using a randomized strategy.

Also, our precise characterization of the optimal sampling distribution for functions on the $[0, 1]$ interval should be extended to higher-dimensional continuous metric spaces. Another natural direction is to consider other aggregation goals, such as predicting the function's maximum, minimum, or median. For predicting the maximum from k deterministic samples, a 2-approximation algorithm was given in [4], which is is best possible unless P=NP. However, it is not clear if equally good approximations can be achieved for the randomized case. For the median, even the deterministic case is open.

On a technical note, it would be interesting whether finding the best sampling distribution for the single sample case is NP-hard. While we presented a PTAS in this paper, no hardness result is currently known.

Acknowledgments. We would like to thank David Eppstein, Bobby Kleinberg, Alex Slivkins and several anonymous referees for helpful feedback.

References

1. Arya, V., Garg, N., Khandekar, R., Meyerson, A., Munagala, K., Pandit, V.: Local search heuristics for k-median and facility location problems. In: Proc. ACM Symposium on Theory of Computing (2001)
2. Bakhvalov, N.S.: On approximate calculation of integrals. Vestnik MGU, Ser. Mat. Mekh. Astron. Fiz. Khim 4, 3–18 (1959)
3. Baran, I., Demaine, E., Katz, D.: Optimally adaptive integration of univariate lipschitz functions. Algorithmica 50(2), 255–278 (2008)
4. Das, A., Kempe, D.: Sensor selection for minimizing worst-case prediction error. In: Proc. ACM/IEEE International Conference on Information Processing in Sensor Networks (2008)
5. Grötschel, M., Lovász, L., Schrijver, A.: The ellipsoid method and its consequences in combinatorial optimization. Combinatorica 1, 169–197 (1981)
6. Gupta, A., Krauthgamer, R., Lee, J.R.: Bounded geometries, fractals, and low-distortion embeddings. In: Proc. IEEE Symposium on Foundations of Computer Science (2003)
7. Mathe, P.: The optimal error of monte carlo integration. Journal of Complexity 11(4), 394–415 (1995)
8. Motwani, R., Raghavan, P.: Randomized Algorithms. Cambridge University Press, Cambridge (1990)
9. Novak, E.: Stochastic properties of quadrature formulas. Numer. Math. 53(5), 609–620 (1988)
10. Traub, J.F., Wasilkowski, G.W., Woźniakowski, H.: Information-Based Complexity. Academic Press, New York (1988)
11. Traub, J.F., Werschulz, A.G.: Complexity and Information. Cambridge University Press, Cambridge (1998)
12. Wozniakowski, H.: Average case complexity of linear multivariate problems part 1: Theory. Journal of Complexity 8(4), 337–372 (1992)
13. Wozniakowski, H.: Average case complexity of linear multivariate problems part 2: Applications. Journal of Complexity 8(4), 373–392 (1992)