Approximating Points by Piecewise Linear Functions*

Danny Z. Chen† Haitao Wang†‡

1 Introduction

Approximating a set of points by a functional curve or surface in the d-D space is a fundamental topic in mathematics and computational geometry. It finds applications in many areas. Different error metrics, constraints, and objective functions give rise to a large number of variations of the problem. For each variation, based on the optimization criteria, two problem versions, *min-#* and *min-ϵ*, are often considered in the literature. In the *min-#* version, the objective is to minimize the complexity of the approximating function f for a given error tolerance under a certain error metric; it is motivated by the desire to obtain an approximating function with the smallest possible complexity while maintaining a certain level of approximation accuracy. In the *min-ϵ* version, the goal is to minimize the error tolerance for a given complexity of f; it is motivated by the desire to achieve the best possible approximation with a specified degree of data compression.

We study a number of variations of the point approximation problem in 2-D and 3-D. For some of the problems, we present the first known results; for others, we improve the previous algorithms in the running time.

2 Statements of Problems

Let $P = \{p_1, p_2, \ldots, p_n\}$ be the input point set, with $p_i = (x_i, y_i, z_i)$ (in the 2-D case, every $z_i = 0$). The *vertical distance* between any point $p_i \in P$ and an approximating functional curve (or surface) f is defined as $d(p_i, f) = |y_i - f(x_i)|$ in 2-D and $|z_i - f(x_i, y_i)|$ in 3-D. The *uniform metric* of error, also known as the L_∞ or *Chebychev* metric, is defined to be $e(P, f) = \max_{1 \le i \le n} d(p_i, f)$. All problems in this paper use the uniform error metric. The *complexity* of f is the total number of line segments in 2-D (or faces in 3-D) of f. Formally, we define min-# and min-ϵ as follows.

min-#: Given an error tolerance $\epsilon \ge 0$, find an approximating function f under the specified constraints such that $e(P, f) \le \epsilon$ and the complexity of f is minimized.

min-ϵ: Given an integer $k > 0$, find an approximating function f under the specified constraints such that the complexity of f is no bigger than k and the error $e(P, f)$ is minimized.

Depending on different constraints on f, we consider the following variations of the problem.

Planar point approximation by a step function Given P in 2-D, the sought f is a step function, which can be represented by a rectilinear curve (see Fig. 1). The problem is motivated by query optimizations and histogram constructions in database management systems. The histogram corresponding to our step function is called the *maximum error histogram* and has been studied in the database area [2, 3]. In the paper, we use **SF** to denote this problem.

Planar point approximation by a piecewise linear function Given P in 2-D, f is piecewise linear and any two consecutive line segments of f need not be jointed (see Fig. 2). This problem is often used in regression analysis. Denote this problem by **PF**.

Weighted version Each point $p_i \in P$ has a weight $u_i \ge 0$ and $d(p_i, f)$ is defined to be $u_i \cdot |y_i - f(x_i)|$. The weighed version is motivated by applications with data of non-uniform significance. Denote the weighted versions of **SF** and **PF** by **WSF** and **WPF**, respectively.

---

*This research was supported in part by NSF under Grant CCF-0515203.

†Department of Computer Science and Engineering, University of Notre Dame, Notre Dame, IN 46556, USA. E-mail: {dchen, hwang6}@nd.edu.

‡Corresponding author.

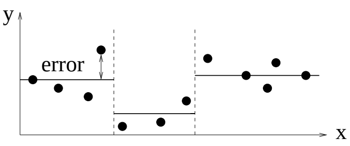

Figure 1: A step function

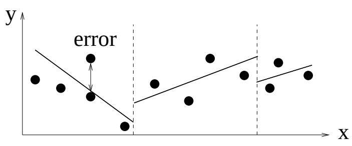

Figure 2: A piecewise linear function

Violation version When approximating P with f, at most g points of P are allowed to violate the error tolerance. Formally, if P' is the set of the violation points and $|P'| \leq g$, then $e(P, f) = \max_{i \in P \setminus P'} d(p_i, f)$. This problem is motivated by applications in, e.g., statistics, machine learning, data mining, databases, where outliers must be reduced. Denote the violation versions of **SF**, **PF**, **WSF** and **WPF** by **VSF**, **VPF**, **VWSF** and **VWPF**, respectively.

3-D version A step function in 3-D can be represented by a rectilinear surface consisting of rectangular faces parallel to the xy-plane, such that any line parallel to the z-axis intersects at most one such face. Denote the 3-D versions of **SF** and **WSF** by **SF3** and **WSF3**, respectively.

We consider both *min-#* and *min-ϵ* algorithms for all the above problems. To simplify the exposition, we assume all points are in non-degenerate positions. Namely, no two points in P have the same x-coordinate in the 2-D problems and no two points have the same x-coordinate *and* the same y-coordinate in the 3-D problems. In all 2-D problems, the input points, $P = \{p_1, p_2, \ldots, p_n\}$, are already sorted in increasing x-coordinate.

3 Previous Work and Our Contributions

In this abstract, we summarize the previously best-known results and our new results in Table 1 shown below. Some problems already have their optimal solutions and we mention them here just for completeness. In the following table, $\phi(i, n) = \underbrace{\log \cdots \log}_{i \text{ times}} n$ and $f(g, n) = \min\{i \geq 1 | \phi(i, n) \leq g\}$.

	min-#		*min-ϵ*	
	Previous	Ours	Previous	Ours
SF	$O(n)$[4]		$O(n)$[1]	
WSF	$O(n)$[3]		$O(n \log^4 n)$[1], $O(n \log n + k^2 \log^6 n)$[2]	$O(\min\{n \log^2 n, n \log n + T \log^2 n\}$
PF	$O(n)$[5]			$O(\min\{n \log n, n + T \log n \log\log n\})$
WPF	$O(n)$[5]			$O(\min\{n \log^3 n, n \log n + T \log^3 n\})$
VSF	$O(ng^2)$[1]		$O(ng^2 \log g)$[1]	$O(ng^2 f(g, n))$
VWSF		$O(ng^2)$		$O(n^2 + ng^2 \log n)$
VPF		$O(ng^4 \log^2 n)$		$O(ng \cdot \min\{kW, W \log n + g^3 \log^3 n\})$
VWPF		$O(ng^4 \log^2 n)$		$O(ng \cdot \min\{kW, W \log n + g^3 \log^3 n\})$
SF3		2-Approx		NP-Hard
WSF3		2-Approx		NP-Hard

Table 1: Our result summary ($T = k^2 \log^2 \frac{n}{k}$, $W = n \log g + g^3 n^\delta$, and $f(g, n) = O(1)$ when $g > \phi(O(1), n)$)

References

[1] H. Fournier and A. Vigneron. Fitting a step function to a point set. To appear in *the 16th Annual European Symposium on Algorithms (ESA)*, Karlsruhe, Germany, September 15–17, 2008.

[2] S. Guha and K. Shim. A note on linear time algorithms for maximum error histograms. *IEEE Transactions on Knowledge and Data Engineering*, 19(7):993–997, 2007.

[3] P. Karras, D. Sacharidis, and N. Mamoulis. Exploiting duality in summarization with deterministic guarantees. In *Proc. of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 380–389, 2007.

[4] J. Díaz-Bá nez and J. Mesa. Fitting rectilinear polygonal curves to a set of points in the plane. *European Journal of Operational Research*, 130:214–222, 2001.

[5] J. O'Rourke. An on-line algorithm for fitting straight lines between data ranges. *Communications of the ACM*, 24:574–578, 1981.