Algorithmica (2006) 44: 103–115
DOI: 10.1007/s00453-005-1201-y

Algorithmica

Linear Time Algorithm for Approximating a Curve by a Single-Peaked Curve

Jinhee Chun,[1] Kunihiko Sadakane,[2] and Takeshi Tokuyama[1]

Abstract. Given a function $y = f(x)$ in one variable, we consider the problem of computing the single-peaked (unimodal) curve $y = \varphi(x)$ minimizing the L_2-distance between them. If the input function f is a histogram with $O(n)$ steps or a piecewise linear function with $O(n)$ linear pieces, we design algorithms for computing φ in linear time. We also give an algorithm to approximate f with a function consisting of the minimum number of unimodal pieces under the condition that each unimodal piece is within a fixed L_2-distance from the corresponding portion of f.

Key Words. Computational geometry, Optimization, Data analysis, Algorithms, Curve simplification.

1. Introduction. Given a function $y = f(x)$ defined on an interval $[0, 1]$, we consider the problem of approximating f by a unimodal function $y = \varphi(x)$.

Here, a function is *unimodal* if it has a unique maximal peak (the peak may be a flat interval). Equivalently, for any real number t, $\{x \in [0, 1] : \varphi(x) \geq t\}$ is either an interval or empty.

We assume that functions considered in this paper are bounded and Riemann integrable (e.g., a piecewise algebraic function). For two functions $g(x)$ and $h(x)$ defined on $[0, 1]$, we define their inner product as $g \cdot h = \int_0^1 g(x)h(x)\,dx$. The L_2-norm of a function g is $\|g\| = \sqrt{g \cdot g}$, and the L_2-distance between two functions g and h is $\|g - h\|$.

We consider the squared L_2-distance

$$\|f - \varphi\|^2 = \int_0^1 (f(x) - \varphi(x))^2\,dx$$

between f and φ. If the squared L_2-distance is minimized, we call φ the optimal unimodal approximation of f. See Figure 1 to get intuition. Without loss of generality, we assume f and φ are nonnegative functions, since we can vertically translate them without changing the distance between them.

The problem is a basic problem in statistics, and initially motivated from a problem in computer vision [2]. Moreover, it has an application to a data mining [8], [9] problem.

We give an $O(n)$ time algorithm for computing the optimal unimodal approximation

[1] GSIS, Tohoku University, Sendai, Japan. (jinhee,tokuyama)@dais.is.tohoku.ac.jp.
[2] Graduate School of System Information Science, Kyushu University, Hakozaki, Japan. sada@csce.kyushu-u.ac.jp.

Received January 27, 2004; revised October 17, 2004, and March 29, 2005. Communicated by N. Katoh.
Online publication December 30, 2005.

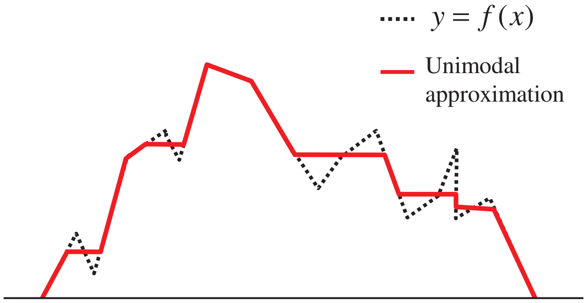

Fig. 1. An input polygonal function (dotted curve) and its optimal unimodal approximation.

if f is a piecewise linear function with n linear pieces. It is easy to see that our method also works for a piecewise algebraic function. The algorithm is designed by using three different forms of the objective function, and also convex hull algorithms. We remark that a different $O(n)$ time algorithm has been proposed by Stout [12] for the special case where the input is a histogram.[3]

A function $y = g(x)$ is called a piecewise unimodal approximation of $y = f(x)$ if g consists of k unimodal portions and the L_2-distance between each portion and the corresponding part of f is at most a given threshold ε. We can compute a piecewise unimodal approximation with the minimum number of maximal peaks in $O(n \log n)$ time by using the above-mentioned linear time algorithm.

The rest of the paper is organized as follows: We first give a geometric characterization of the (integral curve of) the optimal unimodal approximation in Section 2, which leads to the basic design of our algorithms. We next give an algorithm for a special case where the input is a histogram (see Figure 2) in Section 3, and then extend the idea for polygonal functions in Section 4. We give the algorithm for the piecewise unimodal approximation in Section 5, and finally we mention an application to data mining in Section 6.

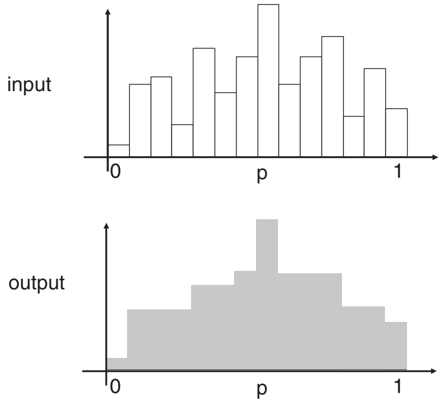

Fig. 2. A histogram and its unimodal approximation (p is the position of the peak).

[3] This was informed by a reviewer of this paper.

2. Geometric Characterization. We use the following notations: The integral functions of functions f and φ are denoted by $F(x) = \int_0^x f(z)\,dz$ and $\Phi(x) = \int_0^x \varphi(z)\,dz$, respectively. Although we do not assume continuity for f and φ, F and Φ are continuous functions.

The convex hull of a geometric object is the minimum convex region containing the object. The lower (resp. upper) hull of a planar geometric object is the lower (resp. upper) curve of the boundary curve of the convex hull (see [10] for a comprehensive description of these notions). By definition, a lower hull is a convex curve and an upper hull is a concave curve.

The following lemma is straightforward from the unimodality of φ:

LEMMA 2.1. *Suppose that φ takes its maximum value at $x = p$. Then $y = \Phi(x)$ is convex* (*resp. concave*) *in the range $x \in [0, p]$* (*resp.* $x \in [p, 1]$).

DEFINITION 1. For the curve $y = F(x)$ and a real value $0 \le p \le 1$, $\mathcal{L}(p)$ is the lower hull of the curve defined by $y = F(x)$ restricted to the range $x \in [0, p]$. Similarly, $\mathcal{U}(p)$ is the upper hull of the curve defined by $y = F(x)$ restricted to the range $x \in [p, 1]$.

Figure 3 illustrates $\mathcal{L}(p)$ and $\mathcal{U}(p)$ of the integral curve $y = F(x)$ of the input function of Figure 2.

Our algorithm for computing the optimal unimodal approximation is based on the following geometric theorem:

THEOREM 2.2. *Suppose that the optimal unimodal approximation φ of f attains its maximum value at $x = p$. Then the curve defined by $y = \Phi(x)$ coincides with $\mathcal{L}(p)$ and $\mathcal{U}(p)$ in the ranges $[0, p]$ and $[p, 1]$ of x, respectively.*

We devote the rest of this section to proving Theorem 2.2. A function f is a piecewise constant function (often called a *histogram*) if $[0, 1]$ is divided into n subintervals

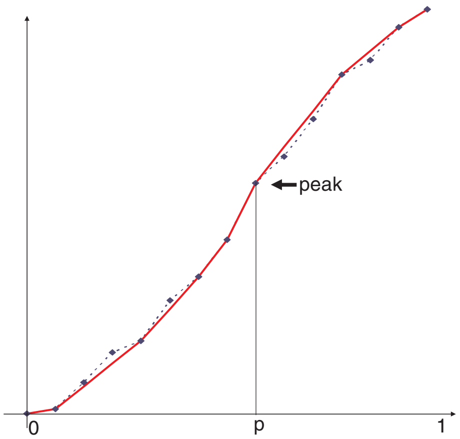

Fig. 3. The curve $y = F(x)$ (given by dotted lines) and $\mathcal{L}(p)$ and $\mathcal{U}(p)$ (solid curves).

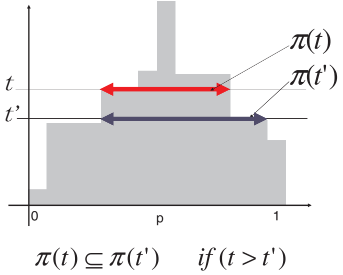

Fig. 4. The pyramid map π.

$I_1, I_2, \dots, I_n$, and f takes a constant value on each subinterval. We call these subintervals the *prime intervals*. We write $I_k = (\ell_k, r_k]$ by using its left and right endpoints, and assume that $\ell_1 = 0$, $\ell_k = r_{k-1}$ ($k = 2, 3, \dots, n$), and $r_n = 1$. Precisely speaking, $I_1 = [0, r_1]$ (instead of $(0, r_1]$). Without loss of generality, we assume that the value of f on I_i is different from that of I_{i+1} for each $i = 1, 2, \dots, n-1$.

The squared L_2-distance $\|f - \varphi\|^2$ is an integral of $(f(x) - \varphi(x))^2$, and the integral is defined as a limit of integrals of histograms. Thus, as shown later, it is essential to prove Theorem 2.2 for histograms. Therefore, we concentrate on the case where f is a histogram. The following lemma is obvious:

LEMMA 2.3. *An optimal unimodal approximation φ of a histogram f is a constant function in the interior of each prime interval.*

It suffices to consider φ that is a constant function on each prime interval, since the difference at the endpoints of prime intervals is irrelevant to our measurement.

Let $\mathcal{I}$ be the set of subintervals of $[0, 1]$. We call $\pi\colon t \in [0, \infty) \to \mathcal{I}$ a *pyramid map* (or *pyramid*) if $\pi(t) \subseteq \pi(t')$ for $t \geq t'$ (see Figure 4). We often call $\pi(t)$ the *horizontal section* of π at the height t.

We consider a unimodal function $y = \varphi(x)$ defined on $[0, 1]$, and define $\pi_\varphi(t) = \{x \in [0, 1]\colon \varphi(x) \geq t\}$ for a nonnegative real number t. Naturally, π_φ is a pyramid map. If φ is the optimal unimodal approximation that is constant on each prime interval, each image of π_φ is either empty or a union of intervals I_i ($i = 1, 2, \dots, n$), and $\varphi(x) = \inf\{t : x \in \pi_\varphi(t)\}$.

Each maximal interval on which φ is a constant function is called a *step* of φ. The value of φ in each step is called the height of the step. From Lemma 2.3, each step is a union of prime intervals.

For the input function f and a unimodal function φ, we consider

$$H(f, \varphi) = \|f\|^2 - \|f - \varphi\|^2.$$

Since $\|f\|$ is given, φ is an optimal unimodal approximation of f if and only if it maximizes $H(f, \varphi)$.

LEMMA 2.4.

$$H(f,\varphi) = 2\int_0^\infty (f(\pi_\varphi(t)) - t|\pi_\varphi(t)|)\,dt,$$

where $f(\pi_\varphi(t)) = \int_{x\in\pi_\varphi(t)} f(x)\,dx$ *and* $|\pi_\varphi(t)|$ *is the length of the interval* $\pi_\varphi(t)$.

PROOF. Let $\mathcal{S}$ be the set of all steps of φ, and let t_J be the height of a step $J \in \mathcal{S}$. It is observed that the horizontal section $\pi_\varphi(t)$ is the union of steps J satisfying $t_J \geq t$. Thus, for $x \in J$, $x \in \pi_\varphi(t)$ if and only if $0 \leq t \leq t_J$.

Let $V = \int_0^\infty (f(\pi_\varphi(t)) - t|\pi_\varphi(t)|)\,dt = \int_0^\infty \int_{x\in\pi_\varphi(t)} (f(x)-t)\,dx\,dt$. By exchanging the order of the double integral, we have

$$\begin{aligned} V &= \sum_{J\in\mathcal{S}} \left[\int_{x\in J} \left\{\int_0^{t_J} (f(x)-t)\,dt\right\} dx\right] \\ &= \sum_{J\in\mathcal{S}} \int_{x\in J} \left(t_J f(x) - \frac{t_J^2}{2}\right) dx. \end{aligned}$$

Thus,

$$\begin{aligned} 2V - \|f\|^2 &= \sum_{J\in\mathcal{S}} \int_{x\in J} 2t_J f(x) - t_J^2 - f(x)^2\,dx \\ &= -\int_0^1 (f(x)-\varphi(x))^2\,dx \\ &= -\|f-\varphi\|^2. \end{aligned}$$

Thus, $2V = \|f\|^2 - \|f-\varphi\|^2 = H(f,\varphi)$. □

Now, suppose that we know a prime interval I attaining the peak of the optimal unimodal function φ. We call I the peak interval. Thus, every nonempty $\pi_\varphi(t)$ contains I. Let p be any point (e.g., the right endpoint) of I. Let us consider the function $F(x) = \int_0^x f(z)\,dz$.

The minimum of $F(x) - tx$ in the range $x \leq p$ is attained at the x-coordinate value of the tangent point by a line with the slope t to the lower hull $\mathcal{L}(p)$ of the curve $y = F(x)$. Let $x_1(t)$ be the x-value of the tangent point. If t is a slope of an edge of the lower hull, we choose any x-value within the edge as $x_1(t)$. Note that $x_1(t)$ is the unique value minimizing $F(x) - tx$, if t is not a slope of an edge of the lower hull. Similarly, we define $x_2(t)$ to be the x-coordinate value of the tangent point by a line with the slope t to $\mathcal{U}(p)$. $F(x) - tx$ is maximized at $x = x_2(t)$ in the range $x \geq p$. Clearly, $x_1(t)$ is a nondecreasing function and $x_2(t)$ is a nonincreasing function with respect to t. Thus, $\pi(t) = (x_1(t), x_2(t)]$ defines a pyramid map π, and its corresponding unimodal function ψ is defined by $\psi(x) = \inf\{t\colon x \in \pi(t)\}$.

LEMMA 2.5. *The unimodal function* ψ *given above is an optimal approximation of* f. *Moreover,* ψ *is the unique optimal approximation that is constant on each prime interval.*

PROOF.

$$
\begin{aligned}
(*) \qquad H(f, \psi) &= \int_0^\infty (f(\pi(t)) - t|\pi(t)|)\,dt \\
&= \int_0^\infty (F(x_2(t)) - F(x_1(t)) - t(x_2(t) - x_1(t)))\,dt \\
&= \int_0^\infty (F(x_2(t)) - tx_2(t) - (F(x_1(t)) - tx_1(t)))\,dt \\
&= \int_0^\infty [F(x_2(t)) - tx_2(t)]\,dt - \int_0^\infty [F(x_1(t)) - tx_1(t)]\,dt.
\end{aligned}
$$

For each t, the integrand of the first term of $(*)$ is maximized and that of the second term is minimized under the condition that $x_1(t) \leq p \leq x_2(t)$, and hence $H(f, \psi)$ is maximized. Thus, ψ is an optimal unimodal approximation.

We next prove the uniqueness. Suppose that φ is any optimal approximation that is constant on each prime interval. Because of maximality, $H(f, \varphi) = H(f, \psi)$. Thus, the measure of the values of t such that $\pi_\varphi(t) \neq \pi(t)$ must be 0, since $\pi(t)$ is the unique interval maximizing $F(\pi(t)) - t|\pi(t)|$ if t is not a slope of an edge of $\mathcal{L}(p)$ or $\mathcal{U}(p)$. Thus, $\inf\{t\colon x \in \pi_\varphi(t)\} = \inf\{t\colon x \in \pi(t)\}$, and hence $\psi(x) = \varphi(x)$. □

Now, we are ready to prove Theorem 2.2. Consider $\Psi(x) = \int_0^x \psi(z)\,dz$ for the above ψ. Since ψ is piecewise constant, Ψ is piecewise linear and continuous. Consider the lower hull $\mathcal{L}(p)$. $\Psi(0) = F(0) = 0$ by definition, and the y value of $\mathcal{L}(p)$ at $x = 0$ is also 0, since the lower hull must contain the leftmost point of $y = F(x)$.

Now, consider any edge e of $\mathcal{L}(p)$. For an x-value x_0 of any point in e, $\psi(x_0) = \sup\{t\colon x_0 \in \pi(t)\} = \sup\{t : x_0 > x_1(t)\}$. Since $x_0 > x_1(t)$ for any t less than the slope of e, $\psi(x_0)$ equals the slope of e. This means that $\psi(x)$ equals the slope of $\mathcal{L}(p)$ for each x-value except the x-values of the vertices of $\mathcal{L}(p)$. Therefore, $\psi(x)$ equals the derivative of $\mathcal{L}(p)$ for $x \in [0, p]$ almost everywhere (i.e., except for a zero-measure subset of $[0, p]$). Thus, $y = \Psi(x)$ coincides with $\mathcal{L}(p)$ in the range $x \leq p$. Similarly, we can show that $y = \Psi(x)$ coincides with $\mathcal{U}(p)$ in the range $x \geq p$. Thus, we have completed the proof of Theorem 2.2 for histograms.

Finally, we prove that Theorem 2.2 holds for a general integrable function f. We consider a series f_i $(i = 0, 1, \ldots)$ of histograms defined by $f_i(0) = f(0)$ and $f_i(x) = f(s2^{-i})$ for $x \in ((s-1)2^{-i}, s2^{-i}]$ for $s = 1, 2, \ldots, 2^i$. Since f is integrable, $\int_{x\in I} f_i(x)\,dx$ uniformly converges to $\int_{x\in I} f(x)\,dx$ for intervals $I \subseteq [0, 1]$; that is, for any $\varepsilon > 0$ there exists N such that $|\int_{x\in I} f_i(x) - \int_{x\in I} f(x)\,dx| < \varepsilon$ for any $i > N$ and $I \subseteq [0, 1]$. In particular, the integral function F_i of f_i uniformly converges to F.

Thus, if we consider the optimal unimodal approximation φ_i of f_i, its integral function Φ_i uniformly converges to the integral function Φ of φ. On the other hand, Φ_i uniformly converges to $\mathcal{L}(p)$ for $x \in [0, p]$ and $\mathcal{U}(p)$ for $x \in [p, 1]$, where p is the limit of the right endpoint of the peak interval of φ_i. Thus, Theorem 2.2 holds for f.

3. Algorithm for Histograms.

Now, it is clear that we can compute the optimal unimodal approximation φ of a histogram f in polynomial time. For each $i = 1, 2, \ldots, n$,

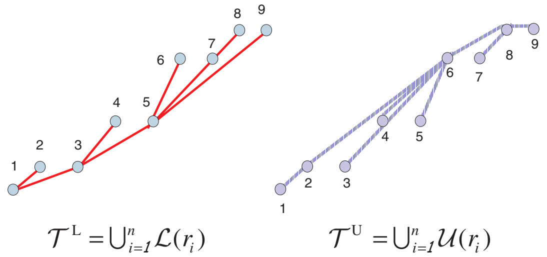
1
2
3
4
5
6
7
8
9
$\mathcal{T}^{\mathrm{L}} = \bigcup_{i=1}^{n} \mathcal{L}(r_i)$
1
2
3
4
5
6
7
8
9
$\mathcal{T}^{\mathrm{U}} = \bigcup_{i=1}^{n} \mathcal{U}(r_i)$

Fig. 5. Lower convex hull tree $\mathcal{T}^L$ and upper convex hull tree $\mathcal{T}^U$.

we consider the prime interval I_i as the peak interval, compute $\mathcal{L}(r_i)$ and $\mathcal{U}(r_i)$ for the right endpoint r_i of I_i, and compute $\psi_i(x)$ as its derivative (i.e., slope function). We call ψ_i the local optimal approximation with the peak index i. Then we select ψ_i minimizing $\|f - \psi_i\|^2$ as the optimal unimodal approximation.

We describe an $O(n)$ time algorithm based on this idea.

$\mathcal{L}(r_i)$ is the lower convex hull of the chain located to the left of r_i, and the union $\mathcal{T}^{\mathrm{L}} = \bigcup_{i=1}^{n} \mathcal{L}(r_i)$ forms a tree called the *lower convex hull tree*. Similarly, we define the tree $\mathcal{T}^{\mathrm{U}} = \bigcup_{i=1}^{n} \mathcal{U}(r_i)$. Figure 5 illustrates lower convex hull tree $\mathcal{T}^{\mathrm{L}}$ and upper convex hull tree $\mathcal{T}^{\mathrm{U}}$. Each vertex of these trees is specified by its coordinate values, and $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$ are regarded as rooted trees with roots $(0, 0)$ and $(1, F(1))$, respectively. Each edge e of the trees is a line segment defined by three parameters: the slope $t(e)$ and the x-values $x_1(e)$ and $x_2(e)$ of the left and right endpoints of e, respectively.

LEMMA 3.1. *$\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$ can be computed in linear time.*

PROOF. We only consider $\mathcal{T}^{\mathrm{L}}$, since $\mathcal{T}^{\mathrm{U}}$ is obtained analogously. Since $y = F(x)$ is piecewise linear, its trajectory is an x-monotone polygonal chain. Clearly, we can compute the polygonal chain in linear time. The lower convex hull of the polygonal chain is indeed the lower convex hull of the point set consisting of vertices of the chain.

We run a plane-sweep convex hull algorithm adding points of the point set one-by-one in the sorted order updating the convex hull (Beneath-and-Beyond method [6], a version of Graham's scan [10]). We are only interested in the lower convex hull. When a new point is inserted, a part of the previous lower convex hull is deleted in the convex hull algorithm; however, in order to construct the lower convex hull tree, we keep it as the "old branch," so that the union of all branches forms the lower convex hull tree. Similarly to the analysis of Graham's scan, this algorithm is linear time for a sorted set of points. See [7] for details. □

In order to compute the optimal unimodal approximation φ, we find the index i minimizing $\|f - \psi_i\|^2$ by using $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$ efficiently. For the purpose, we utilize another form of the objective function given below.

An interval J is called a *balancing interval* for a unimodal function $\psi(x)$ if $\int_{x \in J} (f(x) - \psi(x))\,dx = 0$.

LEMMA 3.2. *If ψ is a local optimal unimodal approximation of f, then each step J of ψ is a balancing interval.*

PROOF. Let t_0 be the height of the step, and let t_1 and t_2 be the heights of adjacent steps such that $t_1 < t_0 < t_2$. Since $\psi(x) = t_0$ for $x \in J$, the balancing condition means that $t_0 = \hat{f}_J$, where $\hat{f}_J = (\int_{x\in J} f(x)\,dx)/|J|$ is the average value of $f(x)$ over $x \in J$. The value $\int_{x\in J}(f(x)-t)^2\,dx$ must be minimized at $t = t_0$, since $\varphi(x) = t_0$ for $x \in J$:

$$\int_{x\in J}(f(x)-t)^2\,dx = \left(\int_{x\in J}(f(x)-t)\,dx\right)^2 + \int_{x\in J} f(x)^2\,dx - \left(\int_{x\in J} f(x)\right)^2.$$

The second and third terms are constants, and the first term takes its unique minimal value at $t = \hat{f}_J$. Thus, $t_0 = \hat{f}_J$. □

LEMMA 3.3. *A unimodal function φ gives the optimal unimodal approximation if and only if it maximizes $\|\varphi\|^2$ under the condition that each step of φ is a balancing interval.*

PROOF. For each step J of φ, it follows from the balancing condition that

$$(**)\qquad \begin{aligned}\int_{x\in J}(f(x)-\varphi(x))^2\,dx &= \int_{x\in J}(f(x)-\hat{f}_J)^2\,dx\\ &= -2\hat{f}_J\int_{x\in J} f(x)\,dx + \int_{x\in J}\hat{f}_J^2\,dx\\ &\quad + \int_{x\in J} f(x)^2\,dx\\ &= -2\hat{f}_J\int_{x\in J}(f(x)-\hat{f}_J)\,dx - \int_{x\in J}\hat{f}_J^2\,dx\\ &\quad + \int_{x\in J} f(x)^2\,dx.\end{aligned}$$

The first term of $(**)$ equals zero since $\hat{f}_J$ is the average value, and the second term equals $-\int_{x\in J}\varphi(x)^2\,dx$. Thus,

$$\int_{x\in J}(f(x)-\varphi(x))^2\,dx = -\int_{x\in J}\varphi(x)^2\,dx + \int_{x\in J} f(x)^2\,dx.$$

Thus, $\|f-\varphi\|^2 = \|f\|^2 - \|\varphi\|^2$, and φ minimizes $\|f-\varphi\|^2$ if and only if it maximizes $\|\varphi\|^2$. □

Thus, it suffices to find ψ_i with the minimum squared norm $\|\psi_i\|^2$. We construct a data structure such that we can compute $\|\psi_i\|^2$ for $i = 1, 2, \ldots, n$ in $O(n)$ time. Consequently, we can compute the optimal unimodal approximation in $O(n)$ time. The structure is constructed by giving weights to edges of $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$. For each edge e in $\mathcal{T}^{\mathrm{L}}$ or $\mathcal{T}^{\mathrm{U}}$, we define its weight $w(e) = (x_2(e) - x_1(e))t(e)^2$, where $t(e)$, $x_1(e)$, and $x_2(e)$ are the slopes and x-values of the left and right endpoints of e, respectively.

For each vertex v of $\mathcal{T}^{\mathrm{L}}$ (resp. $\mathcal{T}^{\mathrm{U}}$), let $W^{\mathrm{L}}(v)$ (resp. $W^{\mathrm{U}}(v)$) be the sum of the weights of edges on the unique path from v to the root of the tree.

LEMMA 3.4. *For each $i = 1, 2, \dots, n$, let v_i^{L} and v_i^{U} be the vertices of $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$ which have r_i as their x-values. Then $\|\psi_i\|^2 = W^{\mathrm{L}}(v_i^{\mathrm{L}}) + W^{\mathrm{U}}(v_i^{\mathrm{U}})$.*

PROOF. Straightforward from the definition of ψ_i. □

THEOREM 3.5. *The optimal unimodal approximation of a histogram f with n steps can be computed in $O(n)$ time.*

PROOF. The values $W^{\mathrm{L}}(v_i^{\mathrm{L}})$ and $W^{\mathrm{U}}(v_i^{\mathrm{U}})$ for all $i = 1, 2, \dots, n$ are computed in $O(n)$ time if we compute them in depth-first order. Thus, we have values $\|\psi_i\|^2$ for $i = 1, 2, \dots, n$, and compute the index i maximizing $\|\psi_i\|^2$ in $O(n)$ time. Once we know the index i, the optimal unimodal approximation is obtained as the slope function of the concatenation of corresponding paths of $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$. □

4. Approximating a Piecewise Linear Function. If we do not mind computational complexity, we can compute the optimal pyramid of a general integrable function f on [0, 1] by considering a series f_i ($i = 0, 1, \dots$) of histograms as discussed in the end of Section 2 to obtain φ from the optimal unimodal approximation φ_i of f_i.

However, if we consider the computational complexity, we should design an algorithm depending on the description of the function f. In this section we consider the case where f is a piecewise linear function consisting of n linear pieces (f is not necessarily continuous).

The interval [0, 1] is subdivided into prime intervals $I_k = (\ell_k, r_k]$ for $k = 1, 2, \dots, n$, and f is a linear function in each interval I_k.

From Theorem 2.2, once we know the x-value p of the peak of φ, then the curve $y = \Phi(x)$ is the lower hull $\mathcal{L}(p)$ (resp. upper hull $\mathcal{U}(p)$) of the curve $y = F(x)$ for $x \in [0, p]$ (resp. $x \in [p, 1]$).

Clearly, we can select the peak position p from $\{r_i\colon\ i = 0, 1, 2, \dots, n\}$, where $r_0 = 0$. Thus, we can basically apply the same algorithm as in the histogram case. The only difference is that the function $y = F(x)$ is piecewise quadratic, instead of piecewise linear. We observe that $\Phi(x) = F(x)$ in each nonlinear part of the curve $y = \Phi(x)$, since a nonlinear part of the boundary curve of the convex hull of an object should be a portion of the object.

Analogously we can construct trees $\mathcal{T}^{\mathrm{L}}$ and $\mathcal{T}^{\mathrm{U}}$ in linear time by modifying the plane-sweep algorithm for computing a convex hull, assuming that the bitangent line between two parabolas can be computed in constant time. Note that the computation of a bitangent line of two parabolas can be done if we can solve a quadratic equation numerically to a sufficient precision. A branching vertex of the trees does not need to correspond to an endpoint of a prime interval; precisely speaking, it can be an endpoint of a bitangent of parabolas.

There are two kinds of edges in the trees, linear edges and curved edges. A curved edge must follow the curve $y = F(x)$; thus, in the same range of x, $\psi(x) = f(x)$ for the corresponding unimodal approximation ψ. The weight $w(e)$ equals $(x_2(e) - x_1(e))t(e)^2$

for a linear edge e, while we define

$$w(e) = \int_{x_1(e)}^{x_2(e)} f(x)^2\,dx$$

for a curved edge e, where we recall that $x_1(e)$ and $x_2(e)$ are x values of the left and right endpoints of e, respectively. The computation of weights of all edges takes $O(n)$ time: We first compute $V(x(v)) = \int_0^{x(v)} f(x)^2\,dx$ for all the x-values $x(v)$ of vertices of trees, and then compute $w(e) = V(x_2(e)) - V(x_1(e))$ for all edges e.

Now, the rest of the algorithm is exactly the same as the case of histogram, and we have the following theorem:

THEOREM 4.1. *The optimal unimodal approximation φ of a piecewise linear function f with n linear pieces can be computed in $O(n)$ time.*

The same strategy works for an input function f that is a piecewise algebraic function of degree d, provided that we can compute all the bitangents of two algebraic curves of degrees $d + 1$ in $O(1)$ time.

5. Piecewise Unimodal Approximation of a Function. Although we have considered the problem where the output is unimodal, we often need to approximate a function with a function that has a small number of maximal peaks. A natural problem formulation is that we approximate f with a function ψ that has at most k peaks such that the squared L_2-distance $\|f - \psi\|^2$ is minimized. This can be done by using dynamic programming in which we use the unimodal approximation algorithm as a subroutine [5].

Here, we consider another problem in which we approximate f with a function g that has the minimum number of pieces under the condition that each piece of g has at most one maximal peak and the L_2-distance between the piece and the corresponding portion of f is within a given threshold ε.

Let $y = f(x)$ be a (piecewise linear) input function. The following greedy algorithm computes a piecewise unimodal approximation of a given function f with a small number of pieces:

1. Compute the largest index k such that there exists a unimodal approximation g_0 of f within the interval $[0, r_k]$.
2. Recursively compute the piecewise unimodal approximation g_1 of f within $[\ell_{k+1}, n]$ with an error ε.
3. Output the concatenation of g_0 and g_1.

The following lemma is easy to see:

LEMMA 5.1. *The greedy algorithm outputs a piecewise unimodal approximation with the minimum number of pieces.*

We can now apply an idea by Agarwal et al. [1].

THEOREM 5.2. *A piecewise unimodal approximation with the minimum number of pieces can be computed in $O(n \log n)$ time.*

PROOF. We apply a combination of the doubling strategy and binary search to compute k in step 1. We check the condition whether f has or does not have a unimodal approximation (with an error less than ε) in the range $[0, r_{2^i}]$ for each $i = 0, 1, \ldots$. Suppose j is the first violating value of i. Then k must satisfy $2^{j-1} \leq k < 2^j$. Now, we perform binary search in the range for k. Thus, it takes $O(2^j \log k) = O(k \log k)$ time for computing g_0. The rest of analysis is routine. □

6. Application to Data Mining. We give a data mining application of our algorithm. Consider a data set consisting of N data items, each of them containing a real-valued attribute $A()$ and an objective binary attribute $B()$. For example, suppose that A is "income" and B is "Ph.D.", and a data v has values $A(v) = 75{,}000$ and $B(v) = yes$. This means that the person associated with the data has a Ph.D. and his/her (annual) income is \$75,000.

An association rule with the form $A(v) \in J \rightarrow B(v) = yes$ is called an interval rule if J is an interval. We would like to find the best interval J such that the above rule separates data into "yes" and "no" with respect to the objective attribute B. For evaluating the interval rules, we define $\text{Support}(J) = |\{v : A(v) \in J\}|/N$, $\text{Hit}(J) = |\{v : A(v) \in J, B(v) = 1\}|/N$, $\text{Conf}(J) = \text{Hit}(J)/\text{Support}(J)$, and $\text{Ent}(J) = \text{Conf}(J) \log(\text{Conf}(J))^{-1} + (1 - \text{Conf}(J)) \log(1 - \text{Conf}(J))^{-1}$. Similarly, for the complement $\bar{J}$ of J, $\text{Support}(\bar{J})$, $\text{Hit}(\bar{J})$, $\text{Conf}(\bar{J})$, and $\text{Ent}(\bar{J})$ are defined.

The *entropy* (precisely speaking, entropy of data splitting into J and $\bar{J}$) defined below is a popular measurement for a data splitting [11]:

$$\text{Ent}(J; \bar{J}) = \text{Support}(J)\text{Ent}(J) + \text{Support}(\bar{J})\text{Ent}(\bar{J}).$$

The interval minimizing the entropy is called the *optimal entropy interval*. Given a data set, we divide the interval of values of A into n prime intervals (for simplicity, we assume n divides N) such that $\text{Support}(I) = 1/n$ for each prime interval I, and precompute $\text{Conf}(I)$ for each prime interval. In [8] an $O(n \log n)$ expected time algorithm is given for computing the optimal entropy interval that is obtained as a union of prime intervals.

We can define a histogram $f(x)$ by $f(x) = \text{Conf}(I)$ for $x \in I$, and consider its optimal unimodal approximation. We modify the criterion for the optimal entropy interval such that the interval minimizes the entropy of the data separation under the condition that it contains the peak position of the optimal pyramid π_φ. Then this modification does not practically affect the data separation, since the optimal entropy interval almost always contains the peak of π_φ when the rule gives a practically good data separation.

Because of the convexity of the entropy function (see [9] for details), the modified optimal entropy interval is obtained as a horizontal section $\pi_\varphi(t)$ of the optimal pyramid π_φ for a suitable height t. As we have shown, we can compute the optimal pyramid in $O(n)$ time and the pyramid has $O(n)$ different horizontal sections. Thus, we can compute the entropy values of all different horizontal slices in $O(n)$ time, and hence we can compute the optimal entropy interval in $O(n)$ time.

Moreover, users of a data mining system often feel that they need an intermediate representation between the interval rule and the original function Conf(), especially when we want to visualize the correlation between $A()$ and $B()$. The optimal unimodal approximation is a good intermediate visual representation to show the correlation. Furthermore, if we use the optimal piecewise approximation, we can consider a more flexible representation, and possibly mine a hidden precious rule that cannot be found by using interval rules.

7. Concluding Remarks. We often need to consider a different distance from the L_2-distance. The L_∞-distance looks easy to handle: If we want to decide whether we have a unimodal approximation whose L_∞-distance from the input is at most (a given) ε, we can answer it in linear time easily. The optimization version can be solved in $O(n \log n)$ time if we apply parametric searching: however, we do not know a linear time algorithm for the optimization problem. For the L_p-distance for a constant p, the authors recently obtained an $O(n \log^2 n)$ time algorithm extending the idea of this paper combined with range searching methods [5].

A two-dimensional version [4] and a higher-dimensional version [3] of this problem have been considered in application to data mining, where the complexity of the problem highly depends on the definition of multivariable unimodal functions. The above works only deal with single-peak problems. The problem of approximating a given surface by another surface with k peaks is a very attractive problem, and is related to the simplification of the Morse complex of a surface, although the authors do not know of any theoretical algorithm on it.

Acknowledgements. The unimodal approximation problem was proposed in the working-group discussion during the Dagstuhl workshop on Geometry, Morphology and Computational Imaging (April 2002) by the third author, inspired by a talk of Isabelle Bloch [2] in the workshop. The authors thank the organizers of the workshop and the working group, especially Tetsuo Asano, Reinhard Klette, Christian Ronse, and David Mount. The authors also thank Naoki Katoh, Akiyoshi Shioura, and Danny Chen for their valuable comments. Finally, the authors gratefully acknowledge anonymous referees who gave kind suggestions on the presentation of the paper.

References

[1] P. Agarwal, S. Hal-Peled, N. Mustafa, and Y. Wang, Near Linear Time Approximation Algorithms for Curve Simplification in Two and Three Dimensions, *Proceedings of the* 10*th European Symposium on Algorithms*, LNCS 2461 (2002), pp. 29–41.

[2] I. Bloch, Unifying Quantative, Semi-Quantitive and Qualitative Spatial Relation Knowledge Representations Using Mathematical Morphology, *Proceedings of the* 11*th Workshop on Theoretical Foundations of Computer Vision*: *Geometry*, *Morphology and Computational Imaging*, LNCS 2616 (2003), pp. 153–164.

[3] D. Chen, J. Chun, N. Katoh, and T. Tokuyama, Efficient Algorithms for Approximating a Multi-Dimensional Voxel Terrain by a Unimodal Terrain, *Proceedings of the* 10*th Computing and Combinatorics Conference* (*COCOON* 2004), LNCS 3106 (2004), pp. 238–248.

[4] J. Chun, K. Sadakane, and T. Tokuyama, Efficient Algorithms for Constructing a Pyramid From a Terrain, *Proceedings of Japan Conference on Discrete and Computational Geometry* 2002, LNCS 2866 (2002), pp. 108–117.

[5] J. Chun, K. Sadakane, T. Tokuyama, and M. Yuki, Peak-Reducing Fitting of a Curve under the L_p Metric, *Interdisciplinary Information Sciences* **11-2** (2005), pp.191–198.

[6] H. Edelsbrunner, *Algorithms in Combinatorial Geometry*, ETACS Monograph on Theoretical Computer Science 10, Springer-Verlag, Berlin, 1987.

[7] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, Mining Optimized Association Rules for Numeric Attributes, *Journal of Computer and System Sciences* **58** (1999), 1–12.

[8] T. Fukuda, Y. Morimoto, S. Morishita, and T. Tokuyama, Data Mining with Optimized Two-Dimensional Association Rules, *ACM Transactions on Database Systems* **26** (2001), 179–213.

[9] Y. Morimoto, T. Fukuda, S. Morishita, and T. Tokuyama, Implementation and Evaluation of Decision Trees with Range and Region Splitting, *Constraints* (1997), 402–427.

[10] F. P. Preparata and M. I. Shamos, *Computational Geometry – An Introduction*, Springer-Verlag, New York, 1988 (2nd edn.).

[11] J. R. Quinlan, *C*4.5: *Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.

[12] Q. F. Stout, Optimal Algorithms for Unimodal Regression, *Computing Science and Statics* **32** (2002), 348–355.