Concept Learning with geometric hypotheses

David P. Dobkin * and **Dimitrios Gunopulos** †

Abstract

We present a general approach to solving the minimizing disagreement problem for geometric hypotheses with finite VC-dimension. These results also imply efficient agnostic-PAC learning of these hypotheses classes. In particular we give an $O(n^{\min(\alpha+1/2,2k-1)}\log n)$ algorithm that solves the m.d.p. for two-dimensional convex k-gon hypotheses (where α is the VC dimension of the implied set system, k is constant), and an $O(n^{3k-1}\log n)$ algorithm for convex k-hedra hypotheses in three dimensions. We extend these results to handle unions of k-gons and give an approach to approximation algorithms.

1 The Minimizing Disagreement Problem

An important goal of applied machine learning is to provide tools for the design and analysis of learning algorithms that provide satisfactory solutions for real-world learning problems. There are a lot of experimental results regarding the performance of various heuristic learning algorithms on a number of "benchmark"-datasets for real world classification problems ([BN], [Ho], [Ma], [WK90], [WGT], [WK91]).

*Department of Computer Science, Princeton University, 35 Olden St., Princeton, NJ 08540, USA, e-mail: dpd@cs.princeton.edu. The research work of this author was supported by NSF Grant CCR93-01254 and by DIMACS, an NSF Science and Technology Center.

†Department of Computer Science, Princeton University, 35 Olden St., Princeton, NJ 08540, USA, e-mail: dg@cs.princeton.edu. The research work of this author was supported by NSF Grant CCR93-01254 and by DIMACS, an NSF Science and Technology Center.

Concept learning is a typical machine learning problem. The objective here is to find a classifier for a concept. We are given a labeled training sequence (a point set of examples, each labeled either as a positive example of the concept or a negative example), and we want to find a hypothesis that can be used as a predictor for other query points ([WK91]). An important optimization problem thus arises, the *minimizing disagreement problem*: Given a labeled training sequence T, and a set of hypotheses $\mathcal{H}$, find a hypothesis $H \in \mathcal{H}$ that minimizes the error for the input set. For a given hypothesis H, the error of H ($Error_T(H)$), is the number of points in T that H classifies incorrectly ([Ma]).

In practice many times the set of hypotheses $\mathcal{H}$ is the set of hyperplanes ([WK91]). The dimension of the space the input set is drawn from is the number of the attributes of the concept, and a hyperplane is the simpler separator between positive and negative examples. Such linear separators are widely used in practice because they are relatively fast to compute, give intuitive solutions, and degrade well ([WK91]). Their simplicity however often limits their accuracy. There is empirical evidence that more complex models can work particularly well in practice. Weiss and Kulikowski ([WK91]) report that for some real life data sets, hypotheses defined as the disjunction or conjunction of simple rules of the form $x_i \leq c_1$ or $x_j = c_2$ outperform neural nets and decision trees. [DGM] present an algorithm to find the optimal hypothesis, when the hypothesis set is restricted to axis aligned boxes.

In this work we examine an extention of the linear separator model. We consider hypotheses that are each defined as the intersection of a constant number of halfspaces. Hypotheses are thus piecewise linear and convex, and are still simple enough to offer intuitive solutions that are more accurate compared to hyperplanes. In a similar approach, Fisher ([F93], [F95]) considers general convex hypotheses.

In addition, the VC dimensions of the geometric hypothesis sets that we consider are finite. As Haussler has shown ([Ha]), if a given hypothesis class $\mathcal{H}$ has finite VC-dimension then a polynomial algorithm that solves the minimizing disagreement problem for $\mathcal{H}$ is a sufficient condition for efficient agnostic PAC-learning

with $\mathcal{H}$ (see also [KSS] and [V]). So our results show efficient (polynomial-time) agnostic PAC-learning with hypothesis sets of k-gons or k-hedra.

To formalize the problem, we turn to the concept of the bichromatic discrepancy of set systems ([C]): A set system is a pair $(S, \mathcal{R})$, where S is a set of points, and $\mathcal{R}$ is a set of subsets (we will call them ranges) of S. Let $(S, \mathcal{R})$ be a set system and let $\chi : S \rightarrow \Re$ be a mapping. For a set $R \in \mathcal{R}$, let $\Delta(R) = \sum_{x \in R} \chi(x)$ be the *bichromatic discrepancy* of Y. We define the *maximum bichromatic discrepancy* of χ on $(S, \mathcal{R})$ by:

$$\Delta_{max}(S, \chi, \mathcal{R}) = \max_{R \in \mathcal{R}} \Delta(R)$$

Usually χ is a mapping to $\{-1, +1\}$, and is called a *coloring* of S (Fig. 1).

Let $S \subset [0,1]^2$ and $\mathcal{H}_k$ be the set of all planar k-gons, where a (potentially open) k-gon is defined as the intersection of k halfspaces. k is a (small) constant. The set $\mathcal{H}_k(S)$ is the set of subsets of S that are defined in the natural way, for $A \subset S, A \in \mathcal{H}_k(S)$ iff there exists a k-gon W such that $S \cap W = A$.

The minimizing disagreement problem and the problem of computing the maximum discrepancy are equivalent ([DGM]). To solve the minimizing disagreement problem for the set of k-gon hypotheses and for a training sequence T, we have to find $\Delta_{max}(S, \chi, \mathcal{H}_k(S))$, the k-gon that maximizes Δ for the set of points S (comprising the points of the examples in T) and the mapping $\chi : S \rightarrow \{-1, +1\}$ which is defined by the labels of the examples in T. To deal with multisets we allow arbitrary weights.

In the next section we present an efficient algorithm that, for a given set $S \in [0,1]^d$ ($d = 2$ or 3, and for the rest of this paper, we assume that $|S| = n$) and a given function $\chi : S \rightarrow \Re$, computes $\Delta_{max}(S, \chi, \mathcal{H}_k(S))$, and therefore solves the equivalent minimizing disagreement problem. To simplify the presentation, we assume that the input points are in general position, that is, there are no three collinear points. If this assumption does not hold, we can use a perturbation technique on the input point coordinates. We note here that we use the RAM model of computation ([HU]).

2 The maximum bichromatic discrepancy of k-gons

First we show that we can find a k-gon that mazimizes the discrepancy even if we consider a finite number of k-gons.

Lemma 1 *For given $S \in [0,1]^2$ and χ, there exists a polygon A with at most k edges such that $\Delta(A) = \Delta_{max}(S, \chi, \mathcal{H}_k(S))$ and each edge of A passes through two points of S.*

Proof Take a maximal k-gon A, and first translate each edge parallel to itself until it touches a point. Then, rotate each edge around the attached points until a second point is encountered (Fig. 2) thus obtaining A'. This operation does not increase the number of edges of A'. For a point p on the boundary of A', we consider it to be in the interior of A' if $\chi(p) > 0$, otherwise we assume it is in the exterior. This assumption increases the discrepancy, and can always be realized by a k-gon if no 3 points are collinear. So during this transformation the discrepancy of A' can only increase, or $\Delta(A') \geq \Delta(A)$. Since A maximizes the discrepancy, $\Delta(A')$ has to be maximum too. □

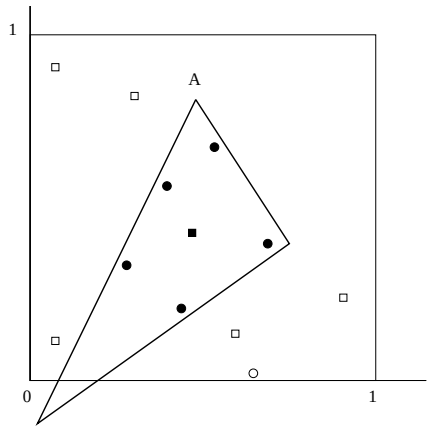

Figure 1: This triangle maximizes the bichromatic discrepancy of this point set for all convex sets (circles are mapped to +1, and squares are mapped to -1, the filled points are the ones in the interior).

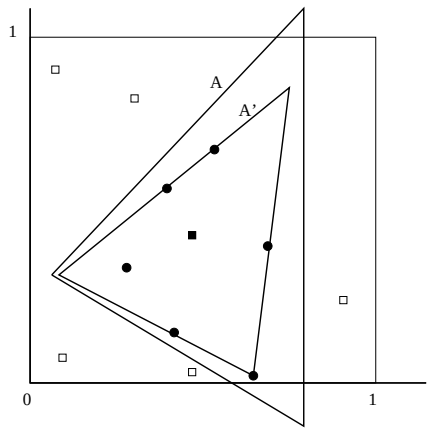

Figure 2: The discrepancy of A' is equal to that of A.

Lemma 1 shows that we have to consider only k-gons with edges defined by pairs of points in S.

Let us consider the special case of halfspaces ($\mathcal{H}_1$). A dynamic algorithm to find the halfspace that maximizes the discrepancy will be useful in the general case. The problem of computing the numerical discrepancy [1] of halfspaces was addressed by Dobkin and Eppstein ([DE]),

[1] In the numerical discrepancy model, a point set is compared to a continuous function (e.g. the area function, [DGM]).

and a dynamic algorithm for the numerical discrepancy is given. It is easy to show the following lemma if we follow their ideas and use the dynamic one-dimensional algorithm (given in [DGM]) that computes the bichromatic discrepancy of a point set in $O(\log n)$ time per insertion or deletion.

Lemma 2 *We can insert and delete points in a set $S \subset [0,1]^2$ with an arbitrary coloring $\chi : S \to \{-1, +1\}$, and recompute the maximum halfspace bichromatic discrepancy in $O(n \log n)$ time, and using $O(n^2)$ space.*

This dynamic algorithm can now be used to find the k-gon that maximizes the discrepancy.

Lemma 3 *We can compute the maximum bichromatic discrepancy $\Delta_{max}(S, \chi, \mathcal{H}_k(F))$ of a set S (where $S \subset [0,1]^2$ and $|S| = n$) and a weight function χ (where $\chi : S \to \Re$) for the set of convex k-gons and find a convex k-gon that maximizes the bichromatic discrepancy in $O(n^{2k-1} \log n)$ time.*

Proof Let us fix $k-2$ halfspaces. Let S' be the subset of S in the intersection of those halfspaces. If we find the wedge that maximizes the discrepancy for S' for all possible choices of $k-2$ halfspaces, we find the k-gon that maxmimizes the discrepancy. For a given point $p \in S'$, we sort the points of $S' \setminus \{p\}$ counter-clockwise around p. We then find the set of points $S'_p = \{q | q \in S \text{ and } q = (q.x, q.y) \text{ and } q.y > p.y\}$. This is the set of the points of S' that lie in the top halfspace that is defined by the horizontal line that passes from p. Using the dynamic algorithm, we find the halfspace that maximizes the discrepancy for S'_p. We then sweep the horizontal line counter-clockwise, appropriately inserting or deleting the points we encounter, and using the dynamic date structure to find the new maximizing halfspace. Thus, in $O(nn \log n) = O(n^2 \log n)$ time we find the wedge that passes from p and maximizes the discrepancy (Fig. 3). We repeat the process for all points, and thus find the maximizing wedge. This algorithm finds a k-gon that maximizes the discrepancy in $O(n^{2(k-2)} n^3 \log n) = O(n^{2k-1} \log n)$ time, and $O(n^2)$ space. □

Next, let's consider the set system $(S, \mathcal{H}_k(S))$. We begin with the definition of the VC-dimension of set systems and related concepts. Let $(S, \mathcal{R})$ be a set system. For a set $Y \subset S$, we call the set system $(Y, \{U \mid (U = R \cap Y) \wedge (R \in \mathcal{R})\})$ the subspace induced by Y. We say that Y is shattered by $\mathcal{R}$ if, in the subspace induced by Y, every possible subset of Y is a range (in other words, if $|\{U \mid (U = R \cap Y) \wedge (R \in \mathcal{R})\}| = 2^{|Y|}$). The Vapnik-Chervonenkis dimension, or VC-dimension of the set system $(S, \mathcal{R})$ is the maximum size of a shattered subset of S ([VC]). The primary shatter function of a set system $(S, \mathcal{R})$ is defined as follows: $\pi_{\mathcal{R}}(m) = \max_{A \subset S, |A|=m} |\{R \cap A | R \in \mathcal{R}\}|$. A result by [VC] gives a bound on the shatter function of a set system with VC-dimension α: $\pi_{\mathcal{R}}(m) = O(m^\alpha)$. Intuitively the VC-dimension of a set system shows how complicated it is to describe.

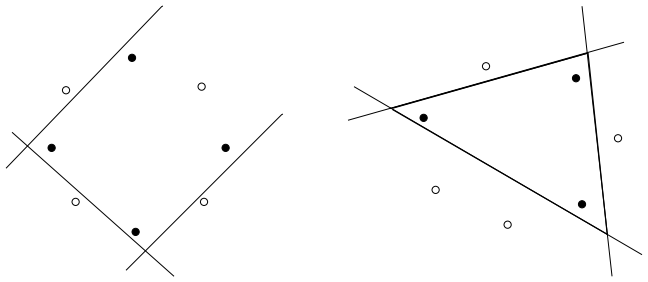

Figure 4: Three halfspaces cannot shatter 8 points but they can shatter 7.

Set systems with finite VC-dimension occur naturally in geometry and in learning theory. The first part of Lemma 2 has been also shown by Welzl ([W]).

Lemma 4 *The VC-dimension of the set system*

$$([0,1]^2, \mathcal{H}_k([0,1]^2))$$

is $2k+1$. The VC-dimension of the set system $(S, \mathcal{H}_k(S))$ is equal to the minimum of the size of the maximum convex subset of S and $2k+1$, and can be computed in $O(|S|^3)$ time.

Proof Let us first show that no subset of $[0,1]^2$, with more than $2k+1$ points can be shattered. Take any T, $T \subset [0,1]^2$, $|T| = 2k+2$. Find the convex hull of T. If a point of T is not in the convex hull, then, after we triangulate the convex hull this point is in the interior of a triangle. Any convex polygon that includes the corners of the triangle has to include the interior point as well; therefore T is not shattered. If all the points of T are on the convex hull, then we can partition them in two independent sets of $k+1$ points each. A convex polygon that includes exactly one of the two sets cannot have less than $k+1$ edges (Fig. 4); again T is not shattered. Now let T be a convex set with $2k+1$ points. Because the set is convex, a single halfspace can separate any subset T' of consecutive points and $T \setminus T'$ (Fig. 4). It is then easy to see that k halfspaces are sufficient to separate any subset of T, and so T is shattered.

From the first part we know that a maximal shattered subset has to be convex, and have cardinality at most $2k+1$. [EG] give an algorithm that computes the maximum convex subset of a two dimensional point set S in $O(|S|^3)$ time. □

The bound on the shatter function shows that there are at most $O(n^\alpha)$ sets of points that we have to consider in order to find $\Delta_{max}(S, \chi, \mathcal{H}_k)$ and we will take advantage of that. Before we prove the main theorem, we need the following lemma.

Lemma 5 *Assume that a k-gon A that maximizes the discrepancy has at least $\alpha - 1$ points from S on its boundary. Also assume that $\alpha \leq 2k$. Then the discrepancy is*

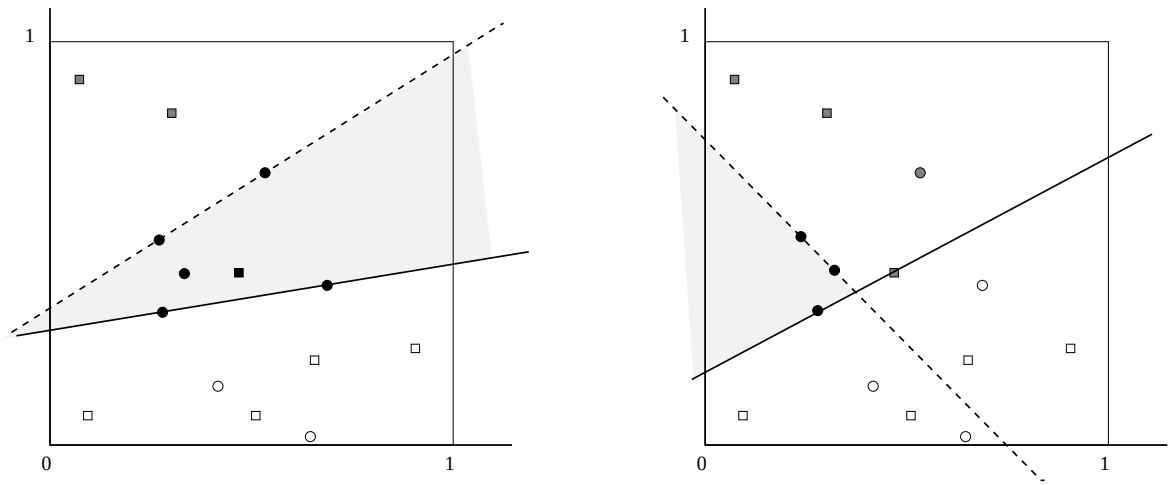

Figure 3: We use the dynamic algorithm for the halfspace discrepancy to find the new maximum wedge (circles are mapped to +1, and squares are mapped to -1, the filled points are the ones in the interior; the dotted line is in each case the maximum halfspace).

also maximized by a convex polygon A' with at most k edges, each passing through two points of S, and two of which don't have points of S on their endpoints.

Proof Let A be an k-gon that maximizes the discrepancy with al least $\alpha - 1$ points on its boundary. Find the convex hull $CH(T)$ of the points $T = \{p \in (A \cap S)\}$. Then $\alpha \geq |CH(T)| \geq \alpha - 1$.

If $|CH(T)| = \alpha$, we can pair consecutive points along the convex hull and we obtain a polygon A' with $\lceil \alpha/2 \rceil$ edges (since $(\alpha \leq 2k$, $\lceil \alpha/2 \rceil \leq k)$ such that no points of S are on A's vertices . In addition $\Delta(A) = \Delta(A')$, because if there exists a point in A' but not in A, then this point and the α points in $CH(T)$ would form a convex hull of $\alpha + 1$ points, larger that the maximum convex hull.

If $|CH(T)| = \alpha - 1$ and α is even, then by pairing consecutive points along the convex hull, we can obtain $\alpha - 1$ separate $\lceil |CH(T)|/2 \rceil$-gons that include all the points in A, and at most two of their edges have a point of S in one of the endpoints. Each such $\lceil |CH(T)|/2 \rceil$-gon differs from $CH(T)$ by the addition of $\lfloor |CH(T)|/2 \rfloor$ triangles. Since the maximum convex hull has size α, only one of these triangles can contain any points, for each $\lceil |CH(T)|/2 \rceil$-gon. It is easy to see that for every non consecutive pair of triangles, there exists a $\lceil |CH(T)|/2 \rceil$-gon that contains both. It follows that at most two of the triangles can contain points, and these two have to be consecutive. Therefore the $\lceil |CH(T)|/2 \rceil$-gon that doesn't include these two satisfies the condition of the lemma (Fig. 5). If α is odd, then again only two triangles can contain additional points, but now they don't have to be consecutive. So a polygon that only includes T has to have $(\alpha + 1)/2$ points. But if α is odd then $(\alpha + 1)/2 \leq k$ and so such a polygon satisfies the conditions of the lemma. □

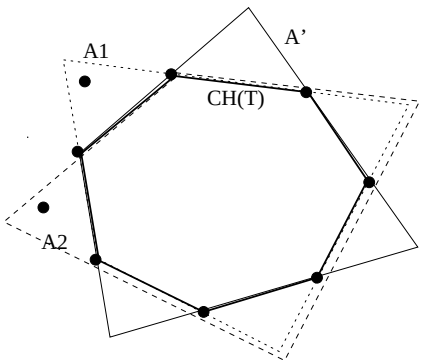

Figure 5: A' maximises the discrepancy. A_1, A_2 show that A' cannot contain any points not in $CH(T)$.

Theorem 1 *We can compute the maximum bichromatic discrepancy $\Delta_{max}(S, \chi, \mathcal{H}_{m-k}(F))$ of a set S (where $S \subset [0,1]^2$ and $|S| = n$) and a weight function χ (where $\chi : S \rightarrow \Re$) for the set of unions of m convex k-gons and find m convex k-gons that maximize the bichromatic discrepancy in $O(n^{\min(m\alpha-1, 2mk-1)} \log n)$, where $\alpha =$ VC-dim $((S, \mathcal{H}_k(S)))$ and m, k are constants.*

Proof We prove the theorem for $m = 1$, but the algorithms are simply extended to handle a constant number of k-gons.

Assume S and χ are given. First we compute $\alpha =$ VC-dim $((S, \mathcal{H}_k(S)))$ in $O(n^3)$ time. If $\alpha = 2k + 1$, we use the algorithm of lemma 3.

Assume $\alpha \leq 2k$. The k-gon that maximizes the discrepancy can either have less that $\alpha - 1$ points in its boundary or more.

For the first case, we consider any subset T of S of most $\alpha - 2$ points. We we find the $O(|T|^2)$ halfspaces that the $|T|$ points define, and compute the discrepancy of

each of the $O(|T|^{2k})$ k-gons thus formed in $O(n|T|^{2k})$ time. To increase the discrepancy, we assume that from the points that lie on the boundary of any k-gon, only the positive ones are in the interior and contribute to its discrepancy. From lemma 4, α is bounded by the size of the largest convex subset of S and $2k+1$. If $\alpha \leq 2k$, then for every k-gon the points in its interior have a convex hull of at most α points. Lemma 1 shows that there exists an $F \in \mathcal{H}_k$ that maximizes the discrepancy and each edge passes through two points. These points must be in the convex hull of $F \cap S$, and therefore F is considered by the algorithm.

For the second case, we use every subset of $\alpha - 4$ or $\alpha - 5$ points to define $k-2$ halfspaces, and find the points of S in the intersection of these halfspaces. Then we run the $O(n^3 \log n)$ algorithm of lemma 3 to find the maximum wedge for these points. We do this for all possible $(\alpha-4)^{2k-4}$ polygons with at most $k-2$ edges that $(\alpha-4)$ points can define, and for all choices of $\alpha-4$ or $\alpha-5$ points. Lemma 5 shows that if there exists a maximum k-gon with α or $\alpha-1$ points on its boundary, this algorithm finds one.

In each case this algorithm finds at least a k-gon that maximizes the bichromatic discrepancy, and therefore computes $\Delta_{max}(S, \chi, \mathcal{H}_k(S))$ in $O(n^{\min(\alpha-1, 2k-1)} \log n)$ time. □

3 The maximum bichromatic discrepancy of k-hedra

In this section we consider the three dimensional problem. The general approach we used for the two dimensional problem partially fails because there is no fast way to compute the exact VC-dimension of the set system $(S, \mathcal{H}_k(S))$. The following lemma gives an upper bound.

Lemma 6 *The VC-dimension of the set system*

$$([0,1]^3, \mathcal{H}_k([0,1]^3))$$

is at most $4k$.

Proof First we observe that a convex set of $4k+1$ points in three dimensions has always an independent set of vertices of size $k+1$. Indeed, a three-dimensional convex set can be represented as a planar graph, and any planar graph can be four-colored ([W]). However k hyperplanes cannot include exactly the remaining $3k$ points because each one can separate only one of the $k+1$ points of the independent set. Therefore a convex set of $4k+1$ points cannot be shattered. In addition a set of points in non-convex position cannot be shattered either because no intersection of hyperplanes can include the points on the convex hull and exclude the points in the interior. So the VC-dimension cannot exceed $4k$. The VC-dimension of $(S, \mathcal{H}_k(S))$ with $S \subset [0,1]^3$ cannot be larger than the size of the maximum convex subset S_c of S either.

It is interesting to observe that the VC dimension of tetrahedra is at most 16 and at least 14. The 12 vertices of a regular icosahedron can be shattered because there exist 4 independent triangular faces, and a halfspace can separate any subset of the 3 points of each such face. Two more points can be positioned in so that the choice on the independent set of faces determines if they are on the inside or the outside of the tetrahedron (Fig. 6). □

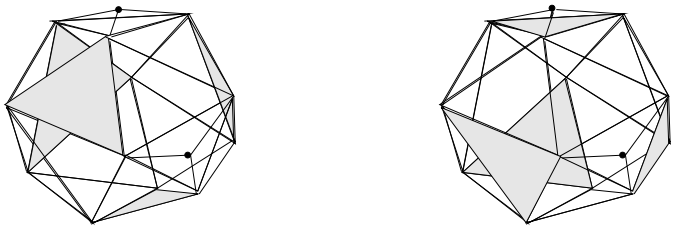

Figure 6: VC-dim$([0,1]^3, \mathcal{H}_4([0,1]^3)) \geq 14$.

Lemma 2 implies an $O(n^{4k+1})$ algorithm. In the rest of the section we develop a faster one. Let us consider first the $k=1$ case. The following lemma uses the dynamic algorithm of Lemma 2.

Lemma 7 *We can insert and delete points in a set* $S \subset [0,1]^3$ *with an arbitrary coloring* χ *and recompute the maximum halfspace bichromatic discrepancy and find a halpspace that maximizes it in* $O(n^2 \log n)$ *time, and using* $O(n^3)$ *space.*

Proof Clearly there exists a halfspace with maximum discrepancy that passes through 3 points of S. For every pair of points p, q, we order the points in $S \setminus \{p, q\}$ in counter clock wise order around the line pq. Then finding the halfspace that maximizes the discrepancy among all halfspaces that pass through p, q is an one dimensional problem, for which an $O(\log n)$ dynamic algorithm with $O(n)$ space requirements is given in [DGM]. If we build such a structure for every pair of points, after the insertion of a point in S we have to update $O(n^2)$ structures and create $O(n)$ new structures. We can therefore compute the new maximum in $O(n^2 \log n)$ time. □

Theorem 2 *We can compute the maximum bichromatic discrepancy* $\Delta_{max}(S, \chi, \mathcal{H}_k(F))$ *of a set* S *(where* $S \subset [0,1]^3$ *and* $|S| = n$*) and a weight function* χ *(where* $\chi : S \rightarrow \Re$*) for the set of convex* k*-hedra and find a convex* k*-hedron that maximizes the bichromatic discrepancy in* $O(n^{3k-1} \log n)$ *time.*

Proof Extending lemma 1 in three dimensions, we observe that there exists a convex polyhedron with at most k faces that maximizes the discrepancy, such that three vertices lie on each face.

To compute the maximum intersection of two halfspaces, for every pair of points p, q first we order the points in $S \setminus \{p, q\}$ in counter clock wise order around the line pq, and then we sweep a hyperplane around pq. After a new point is inserted or deleted, we use the dynamic halfspace algorithm of Lemma 7 to compute the

new maximum discrepancy. It follows that we can compute the intersection of two hyperplanes that maximizes the discrepancy in $O(n^2\ n\ n^2 \log n)$.

We perform this operation for every set of points that lies in the interior of the intersection of $k-2$ halfspaces. The total time is $O(n^{3(k-2)}\ n^5 \log n) = O(n^{3k-1} \log n)$. □

4 The maximum bichromatic discrepancy of stripes

In this section we examine a problem that is similar, but allows a very different approach. Let us give some notation first. A stripe can be defined as the intersection of two halfspaces with parallel supporting lines. Let's assume that l_1 and l_2 are parallel lines. We name the stripe they form $s(l_1, l_2)$ (Fig. 7). For any constant k, let a k-stripe be the union of k parallel planar stripes and let $\mathcal{S}_k$ be the set of planar k-stripes. Holte ([Ho]) has shown that the class of hypotheses that consist of the union of a constant number of horizontal or vertical stripes performs very well in some practical problems. In our model we allow arbitrary orientation.

In the following discussion for simplicity we consider $\mathcal{S}_1$. The following lemma, similar to lemma 1, shows that to find $\Delta_{max}(S, \chi, \mathcal{S}_1(S))$ we have to consider only a finite number of stripes.

Lemma 8 *There exists a stripe $s(l_1, l_2)$ that maximizes Δ such that l_1 and l_2 are parallel to one of the $n(n-1)/2$ lines defined by all pairs of points of S.*

A simple algorithm to compute $\Delta_{max}(S, \chi, \mathcal{S}_k(S))$ is then to find, for each of the $O(n^2)$ possible pairs of points, the stripe which is parallel to the line and maximizes Δ. Our algorithm does exactly that.

To do so we use a dual space transformation. The basic idea is to create a mapping μ between points in the primal space and non-vertical lines in the dual space. A point (a, b) in the primal space is mapped to the line $\mu(a, b) = (y = a\ x + b)$ in the dual space. The dual of a line $y = a\ x + b$ is then the point $\mu(y = a\ x + b) = (-a, b)$ in dual space. An important property of this transformation is that if the point (a_1, b_1) is above the line $y = a_2 x + b_2$ in primal space (that is $(b_1 > a_2 a_1 + b_2)$, then the line it is mapped to $(y = a_1\ x + b_1)$ is above the point $(-a_2, b_2)$ (because $b2 < -a_1\ a_2 + b_1$).

In the dual space, the n points of S are mapped to n lines (Fig. 7). The resulting arrangement has $O(n^2)$ vertices and regions, and it can be constructed in $O(n^2)$ time. The following lemma shows that a non-vertical stripe in the primal space maps to a vertical line segment in the dual space.

Lemma 9 *Let $s(l_1, l_2)$ be a stripe in primal space, such that l_1 is not vertical. Then the line segment $\mu(l_1)\mu(l_2)$ is vertical, and, for every $p \in S$, $p \in s(l_1, l_2)$ if and only if $\mu(l_1)\mu(l_2)$ intersects $\mu(p)$.*

Proof Since l_1 and l_2 are parallel and not vertical, there exist a, b, c such that l_1 can be expressed as $y = a\ x + b$ and l_2 can be expressed as $y = a\ x + c$ (WLOG also assume that $b > c$). Then $\mu(l_1) = (a, b)$ and $\mu(l_2) = (a, c)$, so $\mu(l_1)\mu(l_2)$ is vertical. Take a point $p_i = (x_i, y_i) \in S \cap s(l_1, l_2)$. Since l_1 is above l_2, p_i must be below l_1 (that is $y_i < a\ x_i + b$) and above l_2 ($y_i > a\ x_i + c$). Then in the dual space, the line $\mu(p)$ must pass below the point $\mu(l_1)$ and above the point $\mu(l_2)$, it must therefore cross the line segment $\mu(l_1)\mu(l_2)$. Conversely, a line that crosses $\mu(l_1)\mu(l_2)$ in the dual space corresponds to a point that lies below l_1 and above l_2 in the primal space, so it must be in the interior of $s(l_1, l_2)$. □

Let us take a vertical crosssection of the dual arrangement at some x-coordinate a. A line segment on this crosssection defines a stripe parallel to the line $y = (-a)\ x$. The vertical crosssection intersects the dual lines, and we get an one-dimensional set S' of $m = n(n-1)/2$ points. From lemma 9, the line segment that maximizes the discrepancy for this set S' and the original χ defines the stripe that maximizes the discrepancy among all stripes parallel to $y = (-a)\ x$.

Lemma 9 defines the only $n(n-1)/2$ x-coordinates that we have to consider. They are the lines defined by pairs of points of S, and in the dual space they map to the points of the arrangement. To solve the $O(n^2)$ problems efficiently, we sort the m dual points on their x-coordinate, and we sweep the arrangement with a vertical line. Using the dynamic one dimensional algorithm from [DGM] (lemma 2), sweeping the arrangement takes $O(m^2 \log m)$ time.

Theorem 3 *We can compute the maximum k parallel stripe bichromatic discrepancy of a set $S \subset [0, 1]^2$ with an arbitrary coloring $\chi : S \rightarrow \Re$, and find k parallel stipes that maximize the discrepancy, in $O(k^2 n^2 \log n)$ time, and in $O(n^2)$ space.* [2]

Proof The discussion above proves the theorem for $k = 1$. To extend the algorithm to k-stripes we note that, using the dual transformation, a k-stripe transforms to the union of k intervals. [DGM] extend their dynamic algorithm to dynamically compute the maximum union of k intervals in $O(k^2 \log m)$ (where m is the number of one dimensional points) per update, and this gives a running time of $O(k^2 n^2 \log n)$. □

5 Approximation algorithms

The algorithms we give in the previous chapters have running times that make them impractical for even small input sets. Here we provide a general approach to approximation algorithms to the same problems. Our approach is based on the properties of set ranges with bounded VC-dimension.

[2] We note here that [dB] gives a more complicated algorithm to compute the numerical discrepancy of 1-stripes in $O(n^2 2^{\alpha(n)} \log n)$.

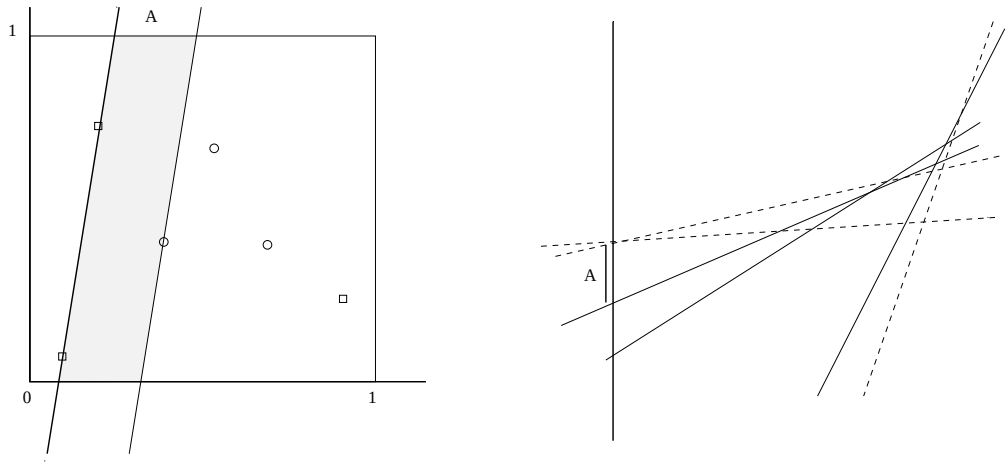

Figure 7: A stripe and its dual.

Let us begin with a technique that was given by [C]. He considers the problem of computing the halfspace that maximizes the numerical discrepancy of a two dimensional point set. His technique easily extends to the bichromatic discrepancy, and we briefly explain it here.

Given a point set S ($|S| = n$), we find the n lines in the dual form. Then, in $O(nr)$ time we compute an $(1/r)$-cutting of the n lines. This gives us a partition of the dual arrangement of into $O(r^2)$ triangles (bounded or unbounded), each of which is cut by at most (n/r) lines. Now, from the properties of the dual transformation (§4), a line in primal space maps in a point in the dual space, and the number of points below a line in the primal space is equal to the number of lines passing below the corresponding point in the dual. Therefore the discrepancy of any halfspace is not more than (n/r) different from the discrepancy of the halfspaces that correspond to the corners of the triangle (of the cutting) that it falls in. Sweeping the partitioning, we compute the maximum discrepancy of these $O(r^2)$ points, in $O(nr)$ time. Assuming the discrepancy is Δ, we start with $r = 2$, doubling r until the discrepancy we find is larger than n/r. The algorithm takes $O(nr) = O(n^2/\Delta)$ time.

The previous technique performs a sampling in the dual space, something that becomes increasingly difficult to do when the primary objects become more complex, moving from halfspaces to wedges and convex k-gons. In the following technique, also suggested by Maass ([Ma2]), we sample in the primal space. For the moment we don't allow arbitrary weight functions.

Let us first introduce the concept of an ϵ-approximation ([VC]). A subset $A \subset S$ is an ϵ-approximation for a set system $(S, \mathcal{R})$ provided that $||A \cap R|/|A| - |R|/|S|| \leq \epsilon$ for every $R \in \mathcal{R}$. A result by Vapnic and Chervonenkis ([VC]) shows that if $(X, \mathcal{R})$ is a set system with finite VC-dimension d, $S \subset X$ is a finite set and ϵ and δ are real numbers, $0 < \epsilon, \delta \leq 1$, then a random sample V os S formed by at least $c(d \log(d/\epsilon) + \log 1/\delta)/\epsilon^2$ independent draws from S is an ϵ-approximation of S for $\mathcal{R}$ with probability at least $1 - \delta$.

This theorem allows us to compute e-approximations to the set of the positive and the negative points, and then run the previous algorithms in the smaller input sets. The following lemma gives a bound on the error.

Lemma 10 *Assume $S \subset [0,1]^2$, $\chi : S \to \{-1, +1\}$. Let $S_+ = \{p | p \in S, \chi(p) = +1\}$ and $S_- = \{p | p \in S, \chi(p) = -1\}$. Let V_+ be an e-approximation of S_+ for $\mathcal{H}_k$ and let V_- be an e-approximation of S_- for $\mathcal{H}_k$. Let $V = V_+ \cup V_-$ and $\chi' : V \to \{-1, +1\}$, such that $\chi'(p) = +1$ if $p \in V_+$, $\chi'(p) = -1$ otherwise. Then:*

$$|\Delta_{max}(V, \chi', \mathcal{H}_k)/|V| - \Delta_{max}(S, \chi, \mathcal{H}_k)/|S|| \leq 2\epsilon$$

Proof For a given range $A \in \mathcal{H}_k$, we have:

$$|\Delta_V(A)/|V| - \Delta_S(A)/|V|| =$$
$$|(|V_+ \cap A| - |V_- \cap A|)/|V| -$$
$$(|S_+ \cap A| - |S_- \cap A|)/|S|| =$$
$$|(|V_+ \cap A|/|V| - |S_+ \cap A|/|S|) -$$
$$(|V_- \cap A|/|V - |S_- \cap A|/|S|)| \leq$$
$$||V_+ \cap A|/|V| - |S_+ \cap A|/|S|| +$$
$$||V_- \cap A|/|V| - |S_- \cap A|/|S|| \leq 2\epsilon$$

It follows that the two maxima cannot be more that 2ϵ different. □

This method has a serious drawback however because the constants can be very large. In fact, to guarantee a .01-approximation for the set of triangles (with VC-dimension 7), the set V must include approximately 10 million points. To achieve a better result we have to use a result about the dual shatter function of a set system. Recall (§2) that the primary shatter function of a set system $(S, \mathcal{R})$ is $\pi_{\mathcal{R}}(m) = \max_{A \subset S, |A|=m} |\{R \cap A | R \in \mathcal{R}\}|$. The dual shatter function $\pi^*_{\mathcal{R}}(m)$ is the primary shatter function of the dual set system arising by exchanging the roles of points and ranges. So $\pi^*_{\mathcal{R}}(m)$ is the maximum number of equivalence classes that the

points of S can be partitioned into, by m ranges in $\mathcal{R}$. Matousek, Welzl and Wernish ([MWW]) show that if the dual shatter function of a set system $(S, \mathcal{R})$ (with $|S| = n$) is bounded, $\pi^*_{\mathcal{R}}(m) \leq Cm^d$ for all $m \leq n$, then for every $1/\epsilon \leq n$ an ϵ-approximation for $(S, \mathcal{R})$ of size $O(r^{2/(d+1)-2}(\log 1/r)^{2-2/(d+1)})$ can be computed in polynomial time. To use this theorem we have to bound the dual shatter function of the set system $(S, \mathcal{H}_k(S))$ and we do so with the following lemma.

Lemma 11 *Let $S \subset [0,1]$, and take the set system $(S, \mathcal{H}_k(S))$. Let $\pi^*_{\mathcal{H}_k}(m)$ be the dual shatter function of $(S, \mathcal{H}_k(S))$. Then $\pi^*_{\mathcal{H}_k}(m) = O(m^2)$.*

Proof We have to bound the number of equivalent classes that m k-gons partition S. Clearly if instead of m k-gons we consider km halfspaces, the number of equivalent classes can only increase. In two dimensions km halfspaces define $O((km)^2) = O(m^2)$ (since k is a constant) different regions, and therefore the number of equivalent classes is $O(m^2)$. It follows then that $\pi^*_{\mathcal{H}_k}(m) = O(m^2)$. □

Using [MWW]'s construdtion and lemma 10, we obtain the following theorem.

Theorem 4 *Given a set S of points ($S \in [0,1]^2$, $|S| = n$), $\chi : S \to \{-1, +1\}$, and $\epsilon > 0$, we can find an ϵ-approximation with $O(\epsilon^{-4/3}(\log(1/\epsilon)^{4/3})$ points and so compute the maximum bichromatic discrepancy for convex k-gons within a ϵn additive term.*

6 Future directions and open problems

The algorithms we present for computing the maximum discrepancy have direct applications in machine learning. Their running time, when simple hypotheses such as triangles and tetrahedra are considered, allows them to be useful with the large data sets common in many real-life problems. A very interesting open problem is giving practical algorithms for higher dimensions. The algorithms we present extend to higher dimensions, but the running times become much worse.

These running times bring the problems of showing if these algorithms are optimal, and of finding good approximations in reasonable time bounds. A possible approach would be to combine sampling on the input set and sampling on the set of ranges, to reduce the running time. Such a probabilistic algorithm seems possible, but difficult to analyze.

References

[AL] D. Angluin and P. Laird, Learning from noisy examples. *Machine Learning*, 2 (1988), 343-370.

[BN] W. Buntine and T. Niblett, A further comparison of splitting rules for decision-tree induction. *Machine Learning*, 8 (1992), 75-82.

[C] B. Chazelle, Geometric Discrepancy Revisited. *34th IEEE Symp. Foundat. Computer Science* (1993).

[dB] M. deBerg, *Computing Half-Plane and Strip Discrepancy of Planar Point Sets,* to appear (1994).

[DE] D. Dobkin and D. Eppstein, Computing the Discrepancy. *9th Ann. Symp. on Comput. Geom.* (1993), 47-52.

[DGM] D. Dobkin, D. Gunopulos and W. Maass, Computing the maximum Bichromatic Discrepancy, with applications in Computer Graphics and Machine Learning. *JCSS*, to appear.

[DG] D. Dobkin and D. Gunopulos, The maximum discrepancy of simple geometric ranges, TR-480-94, Princeton U., 1994.

[EG] H. Edelsbrunner and L. J. Guibas, Topologically Sweeping an Arrangement. *JCSS*, 38, 165-194 (1989).

[F93] P. Fisher, Learning unions of convex polygons. *Proc. of EURO-COLT '93*, 1993.

[F95] P. Fisher, More or Less Efficient Agnostic Learning of Convex Polygons. *these procedings (COLT-95)*, to appear.

[Ha] D. Haussler, Decision theoretic generations of the PAC-model for neural nets and other applications. *Inf. and Comp.*, 100 (1992), 78-150.

[Ho] R.C. Holte, Very simple classification rules perform well on most commonly used datasets. *Machine Learning*, 11 (1993), 63-91.

[HU] J.E. Hopcroft and J.D. Ullman, *Introduction to Automata Theory, Languages and Computation.* Addison-Wesley (1979).

[KSS] M. Kearns, R.E. Schapire and L.M. Sellie, Toward efficient agnostic learning. *5th ACM Workshop on Computational Learning Theory* (1992) 341-352.

[Ma] W. Maass, Efficient Agnostic PAC-Learning with Simple Hypotheses. *7th Ann. ACM Conference on Computational Learning Theory* (1994), 67-75.

[Ma2] W. Maass, private communication.

[MWW] J. Matousek, E. Welzl and L. Wernish, Discrepancy and ϵ-approximations for bounded VC-dimension. *25th ACM Symp. on Theory of Computing* (1993), 424-430.

[V] L.G. Valiant, A theory of the learnable. *Comm. of the ACM* 27 (1984), 1134-1142.

[VC] V.N. Vapnik and A.Ya. Chervonenkis, On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Applic.* 16 (1971), 264-280.

[WGT] S.M. Weiss, R. Galen and P.V. Tadepalli, Maximizing the predictive value of production rules. *Art. Int.* 45 (1990), 47-71.

[WK90] S.M. Weiss and I. Kapouleas, An empirical comparison of pattern recognition, neural nets, and machine learning classification methods. *11th Int. Joint Conf. on Art. Int.* (1990), Morgan Kauffmann, 781-787.

[WK91] S.M. Weiss and C.A. Kulikowski, *Computer Systems that Learn* (1991), Morgan Kauffmann.

[W] E. Welzl, personal communication.