Concept Learning with simple geometric hypotheses

David P. Dobkin [1] [2] Dimitrios Gunopulos [1] [3]

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 three-dimensional convex k-hedra hypotheses. 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]).

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

[1]The research work of these authors was supported by NSF Grant CCR93-01254 and by DIMACS, an NSF Science and Technology Center.

[2]Department of Computer Science, Princeton University, 35 Olden St., Princeton, NJ 08540, USA, e-mail: dpd@cs.princeton.edu

[3]Department of Computer Science, Princeton University, 35 Olden St., Princeton, NJ 08540, USA, e-mail: dg@cs.princeton.edu

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 several halfspaces. Hypotheses are thus piecewise linear and convex, and offer intuitive solutions that are more accurate compared to hyperplanes.

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:

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

Usually χ is a mapping to $\{-1, +1\}$, and is called a *coloring* of S (Fig. 1). To simplify things and avoid the absolute values, we also define two more functions:

$$\Delta_{max}(S, \chi, \mathcal{R}) = \max_{R \in \mathcal{R}_2} \Delta(R) \quad \text{and} \quad \Delta_{min}(S, \chi, \mathcal{R}) = \min_{R \in \mathcal{R}_2} \Delta(R)$$

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 weight 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.

In addition we show that 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]). Combined our results then show efficient agnostic PAC-learning with the sets of k-gons and k-hedra.

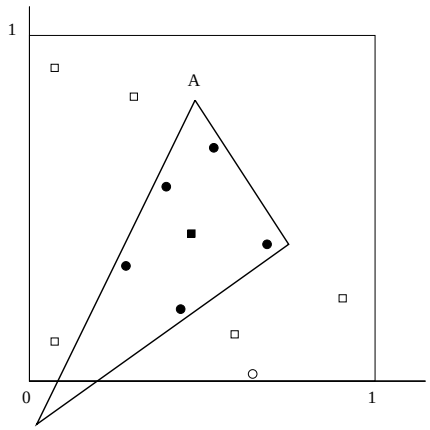

Figure 1: This triangle maximizes the bichromatic discrepancy of this point set for all convex sets (circles are mapped to +1, and squares are mapped to -1, the filled points are the ones in the interior).

2. The Discrepancy and the VC-dimension of k-gons

We begin this section with a closer examination of the set system $(S, \mathcal{H}_k(S))$. Let us first give 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.

Set systems with finite VC-dimension occur naturally in geometry and in learning theory. Lemma 1.a has been also shown by Welzl ([W]).

Lemma 1.a: The VC-dimension of the set system $([0,1]^2, \mathcal{H}_k([0,1]^2))$ is $2k+1$.

1.b: 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: *a:* 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. 2.a); 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. 2.b). It is then easy to see that k halfspaces are sufficient to separate any subset of T, and so T is shattered.

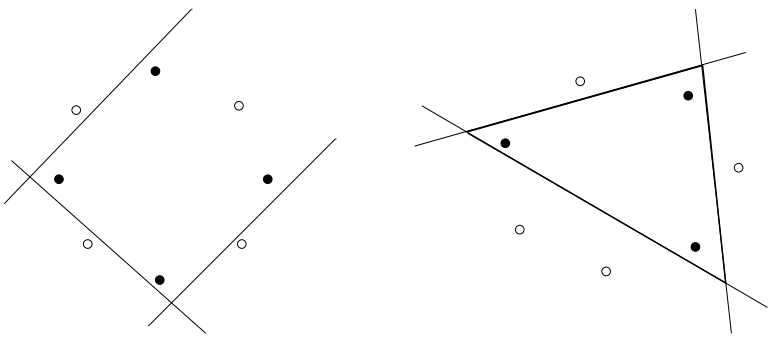

Figure 2: Three halfspaces cannot shatter 8 points but they can shatter 7.

b: 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 $O(n^{2k+1})$ sets of points that we have to consider in order to find $\Delta_{max}(S, \chi, \mathcal{H}_k)$. The following two lemmata show how to find them.

Lemma 2: 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. 3) 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. □

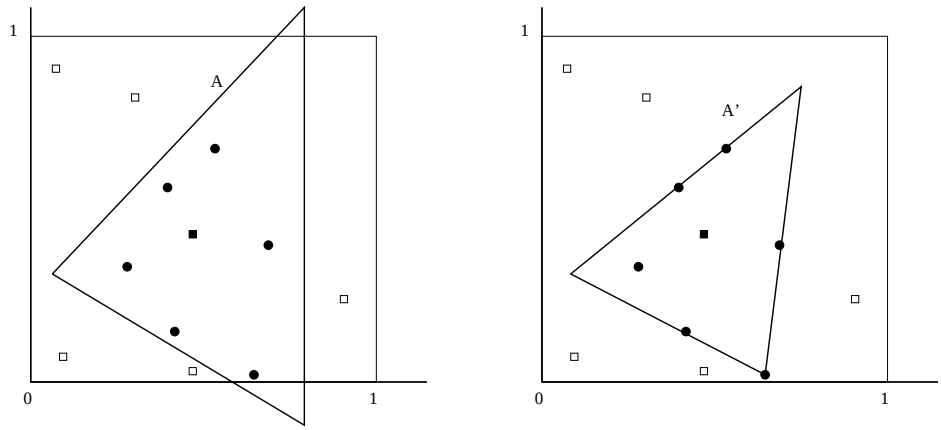

Figure 3: The discrepancy of A' is equal to that of A.

Lemma 3: The bichromatic discrepancy $\Delta_{max}(S, \chi, \mathcal{H}_k(S))$ of a set S ($S \subset [0,1]^2$, $|S| = n$) and a weight function χ ($\chi : S \rightarrow \Re$) for the set of convex k-gons can be computed in $O(n^{\alpha+1/2} \log n)$, where $\alpha =$ VC-dim $(S, \mathcal{H}_k(S))$.

Proof: Assume S and χ are given. First we compute $\alpha =$ VC-dim $((S, \mathcal{H}_k(S)))$ in $O(n^3)$ time. We consider all subsets of S with at most α points. For each, we find its convex hull. Chazelle and Welzl ([CW]) show how to construct an algorithm that computes the weight of points inside a k-gon in $O(n^{1/2} \log n)$ time, provided a spanning path of low stabbing number ($O(n^{1/2})$ is known. Matousek ([Mat]) shows that we can construct such a spanning path in $O(n^{1+\delta})$ time (δ depends on the O constant). We use this algorithm to find the number of points of S in the interior of the convex hull. We do two separate queries for positive and negative points. If $\alpha \leq k$ then obviously this subset can be in the interior of a k-gon, but if $\alpha > k$ we have to determine if there exists a k-gon that contains exactly these α points and the points in the interior of their convex hull. To do so, we find the $O(\alpha^2)$ halfspaces that the $O(\alpha)$ points in the convex hull define, and check, in $(O(\alpha^{2k} n^{1/2} \log n))$ time, if any of the $O(\alpha^{2k})$ k-gons thus formed contains exactly the same number of points with the convex hull of the α points. 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 (Fig. 4).

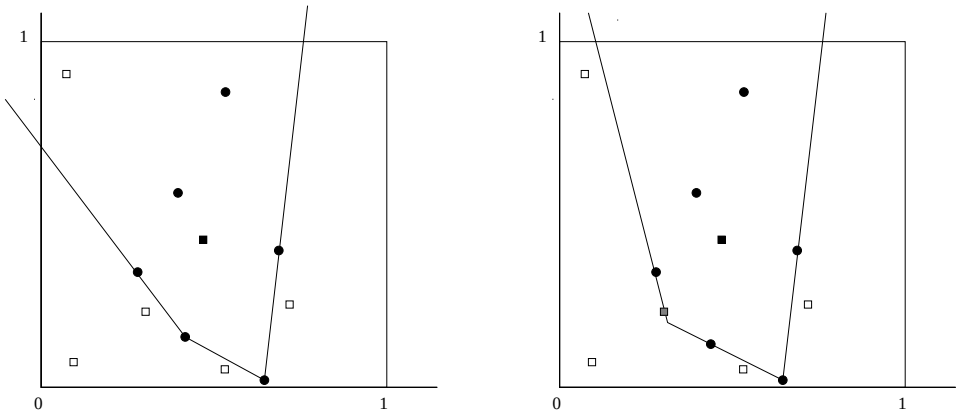

Figure 4: The subset on the left cannot be separated by 3 halfspaces that pass through its points. But the subset on the right has at least equal discrepancy.

Since for every $F \in \mathcal{H}_k$, the set $F \cap S$ has a convex hull of at most α, we have from that for any k-gon, we have thus considered one with at least equal discrepancy. It follows that this algorithm computes $\Delta_{max}(S, \chi, \mathcal{H}_k(S))$ in $O(n^{\alpha+1/2} \log n)$ time. □

It is clear that this algorithm is not optimal when $\alpha = 2k+1$. Indeed we can enumerate all k-gons that have each edge passing through a point in $O(n^{2k})$ time, and compute each discrepancy in linear time. Lemma 2 shows that this $O(n^{\alpha})$ algorithm gives the correct solution.

Here we show how to improve this simple algorithm. The following lemma is shown in [DG].

Lemma 4: The bichromatic discrepancy $\Delta_{max}(S, \chi, \mathcal{H}_2(S))$ of a set S ($S \subset [0,1]^2$, $|S| = n$) and a weight function χ ($\chi : S \rightarrow \Re$) for the set of wedges (intersections of 2 halfspaces) can be computed in $O(n^3 \log n)$, where

The algorithm of lemma 4 can be used to find the k-gon that maximizes the discrepancy. 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 maximizes the discrepancy.

Theorem 1: The bichromatic discrepancy $\Delta_{max}(S, \chi, \mathcal{H}_{m-k}(F))$ of a set S ($S \subset [0,1]^2$, $|S| = n$) and a weight function χ ($\chi : S \rightarrow \Re$) for the set of unions of m convex k-gons can be

computed in $O(n^{\min(m\alpha+1/2,2mk-1)}\log n)$, where $\alpha =$ VC-dim $((S,\mathcal{H}_k(S)))$ and m,k are constants.

Proof: For $m=1$, lemmata 3 and 4 provide the proof. The algorithms are simply extended to handle a constant number of k-gons. □

3. The discrepancy of k-hedra

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

Lemma 5: 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 $(S,\mathcal{F}_k(S))$ with $S\subset[0,1]^3$ cannot be larger than the size of the maximum convex subset $S_c\subset S$ either.

It is interesting to observe that the VC dimension of tetrahedra is at most 16 and at least 14 ($14\leq$ VC-dim $([0,1]^3,\mathcal{F}_4([0,1]^3))\leq 16$). 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. 5). □

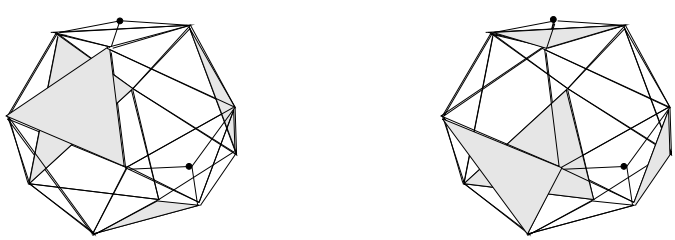

Figure 5: VC-dim$([0,1]^3,\mathcal{F}_4([0,1]^3))\geq 14$.

Let us consider first the $k = 1$ case. The following lemma is based on ideas given in [DE] and a dynamic algorithm given in [DGM] that computes the bichromatic discrepancy in one dimension.

Lemma 6: 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 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: The bichromatic discrepancy $\Delta_{max}(S, \chi, \mathcal{H}_k(F))$ of a set S ($S \subset [0,1]^3$, $|S| = n$) and a weight function χ ($\chi : S \to \Re$) for the set of convex k-hedra can be computed in $O(n^{3k-1} \log n)$.

Proof: Extending lemma 2 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 6 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 compute the maximum for every set of points that lies in the interior of the intersection of $k-2$ halfspaces. The running time is then $O(n^{3(k-2)}\ n^5 \log n) = O(n^{3k-1} \log n)$. □.

4. 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.

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 forn. 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 (§5), 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 $c(d \log(d/e) + \log 1/\delta)/\epsilon^2$ independent draws from S is an e-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 7: 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 8: Let $S \subset [0,1]$, and take the set system $(S, \mathcal{H}_k(S))$. Let $\pi^*_{\mathcal{H}}(m)$ be the dual shatter function of $(S, \mathcal{H}_k(S))$. Then $(S, \mathcal{H}_k(S)) = O(m^2)$.

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 $(S, \mathcal{H}_k(S)) = O(m^2)$. □

Using this construction we can compute, for any $\epsilon > 0$, ϵ-approximations for S_+ and S_- with $O(\epsilon^{-4/3}(\log(1/\epsilon)^{4/3})$ points each. Lemma 8 shows that we can use these point sets to find the discrepancy of the original point set within a $2\epsilon n$ additive term.

5. 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).

[CW] B. Chazelle and E. Welzl, Quasi-optimal Range Searching in Spaces of Finite VC-Dimension. *Disc. Comput. Geom.*, 4, 467-489 (1989).

[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 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, Princeton U., 1994.

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

[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.

[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.

[Mat] J. Matusek, Efficient Partition Trees, *7th Ann. ACM Symp. on Comput. Geom.* 1-9 (1991).

[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.