Journal of Machine Learning Research 2 (2002) 335-358 Submitted 6/01; Published 2/02

Machine Learning with Data Dependent Hypothesis Classes

Adam Cannon CANNON@CS.COLUMBIA.EDU
Department of Computer Science
Columbia University
New York, NY 10027, USA

J. Mark Ettinger ETTINGER@LANL.GOV
Nonproliferation and International Security Group, NIS-8
Los Alamos National Laboratory
Los Alamos, NM 87545, USA

Don Hush DHUSH@LANL.GOV
Clint Scovel JCS@LANL.GOV
Modeling, Algorithms, and Informatics Group, CCS-3
Los Alamos National Laboratory
Los Alamos, NM 87545, USA

Editor: Peter Bartlett

Abstract

We extend the VC theory of statistical learning to data dependent spaces of classifiers. This theory can be viewed as a decomposition of classifier design into two components; the first component is a restriction to a data dependent hypothesis class and the second is empirical risk minimization within that class. We define a measure of complexity for data dependent hypothesis classes and provide data dependent versions of bounds on error deviance and estimation error. We also provide a structural risk minimization procedure over data dependent hierarchies and prove consistency. We use this theory to provide a framework for studying the trade-offs between performance and computational complexity in classifier design. As a consequence we obtain a new family of classifiers with dimension independent performance bounds and efficient learning procedures.

Keywords: Computational Learning Theory, Empirical Process Theory, Classification, Shatter Coefficient, Structural Risk Minimization

1. Introduction

Vapnik motivated his development of support vector machines as a kind of structural risk minimization. However, the corresponding class sequence is data dependent and so Vapnik's theory of structural risk minimization does not apply. To resolve this issue, Shawe-Taylor et al. (1998) have developed a theory of structural risk minimization over data dependent hierarchies which gives performance guarantees for support vector machines (see also Shawe-Taylor & Cristianini, 2000). In addition, the work on data dependent complexity regularization of Buescher and Kumar (1996), the posterior bounds of Freund (1998), the conditional bounds of Devroye (1988), and the work on data dependent penalties of Koltchinskii et al. (2000), Koltchinskii (2001), Boucheron et al. (2000), and Bartlett et al. (2000) have a similar goal.

Devroye (1988) developed performance bounds for data dependent hypothesis classes in a similar spirit to those presented here. However Devroye's approach provides conditional bounds whereas the approach taken here more closely resembles the VC framework developed by Vapnik. Recently Gat (1999) proved a performance bound for the perceptron that extended the VC theorem to the case of data dependent classes of classifiers. This bound uses a counting complexity instead of a shattering complexity for the data dependent class. In this paper we show that Gat's result can be extended in terms of a shatter coefficient for data dependent classes. We build on this result to construct a new framework for learning with data dependent hypothesis classes. Although the framework of Shawe-Taylor et al. may be more general, its exact relationship to the framework developed here is not clear. At a minimum the two frameworks appear to differ in the ease with which they can be applied to particular learning paradigms. For example, our framework has not yet provided performance guarantees for support vector machines. However, it has facilitated the discovery of new families of classifiers that possess dimension independent performance guarantees for empirical risk minimization.

We contrast our framework with the VC framework. In the VC framework the generalization error can be decomposed into two components, the approximation error A which quantifies the lack of optimality introduced by our choice of hypothesis class, and the estimation error E which quantifies the lack of optimality of empirical error minimization due to finite sample size. For classes with finite VC dimension the celebrated VC theorem provides a distribution independent bound on E that goes to zero as the number of training samples n goes to infinity. Control on the approximation error is provided through a structural risk minimization (SRM) learning procedure that is proved to be consistent for an infinite sequence of classes with finite VC dimension.

In our framework classifier design is decomposed into two components; the first component is the restriction to the data dependent hypothesis class and the second is empirical risk minimization within that class. We define data dependent versions of approximation error A_D and estimation E_D error in the obvious way. We provide a VC-like theorem that gives bounds on E_D in terms of a shatter coefficient for data dependent classes. Based on Vapnik's construction we also provide a structural risk minimization procedure over data dependent hierarchies and prove consistency. When the data dependent hypothesis class is obtained by restricting a traditional class through a data dependency rule the data dependent approximation error A_D splits into

$$A_D = A + E_{DD}$$

where A is the approximation error for the traditional class and E_{DD} is the *data dependency error*. Consequently the analysis of approximation error is similar to that for traditional classifiers, except for the data dependency error term. We have just begun the study of this random variable. For example we show that when the data dependency is symmetric in its dependence on the n-sample the infinite sample limit of this random variable is a constant. At present we do not know general conditions on the data dependency that allow this constant to be computed, and in particular when it is zero. However, for the structural risk minimization over multi-sphere classifiers described in Section 5.2 this constant is zero. In this case we have also shown that the infinite sample limit of the data dependent approximation error is zero.

We also wish to address the computational requirements of the learning procedure. Despite the performance guarantees provided by the VC theory, empirical error minimization is computationally intractable for nearly all nontrivial hypothesis classes. One of the strengths of our framework is that it allows us to explore trade-offs between generalization error and computational requirements through the choice of data dependency. For example, it allows us to modify a traditional classifier to obtain a family of classifiers through various data dependencies, and to quantify the performance and computational requirements over this family with a greater degree of flexibility than we have seen before. Among the families considered here are classifiers with dimension independent performance bounds that may benefit from kernel mappings to high dimensions like those used in support vector machines.

This paper is organized as follows. We prove uniform convergence of empirical processes over data dependent classes as an extension of Gat's theorem. Specific data dependent classes are introduced and their computational and structural complexity analyzed. A structural risk minimization procedure is described and a consistency theorem proven. The structural risk minimization procedure is demonstrated on a specific family of spherical classifiers and conditions for consistency with the Bayes error are provided. Finally we describe a framework for analyzing the trade-offs between computation and performance as a function of the data dependency and apply it to the family of spherical classifiers.

2. Uniform convergence of empirical processes over data dependent classes

Let Z denote a metric space with its Borel σ-algebra and a Borel probability measure μ. We let Z_n denote the n-fold product of Z with itself with product σ-algebra and let $\mathcal{P}_{Z_n} = \mu^n$ denote the n-fold product of the measure μ. We denote n independent samples $\{z_n(i), i = 1, .., n\}$ as $z_n = (z_n(1), .., z_n(n)) \in Z_n$. In this paper we consider functions $f : Z \to \{0, 1\}$. We define $\mathcal{F}_n$ to be a class of such functions, where members of this class are determined through application of a data dependency rule to n-samples. For example, in Section 4.2 we consider functions that dichotomize $\Re^d$ with a sphere where the data dependency rule forces the sphere to be centered at one of the n data points. Therefore, $\mathcal{F}_n$ is a class of binary functions on (Z_n, Z). We define a data dependent class $\mathcal{F} = \{\mathcal{F}_n\}$ to be a collection of classes $\mathcal{F}_n$.

We now assume some structure concerning the classes' data dependency and the classes' description as a parametric family. We denote the class $\mathcal{F}$ restricted to the n-sample z_n by $\mathcal{F}_{z_n}$. We assume we can describe this class in parametric form through some parameter space Y_n, not necessarily a product space, such that each function is $f_{y_n,z_n}(z)$ for some choice of $y_n \in Y_n$ and each choice of y_n gives a function in the class. Then the whole data dependent class on n samples can be described by a single function $\Xi_n(y_n, z_n, z) = f_{y_n,z_n}(z)$. Recall that a Polish space is a complete separable metric space and a Suslin space is a Borel measurable image of a Polish space. We say that the data dependent class $\mathcal{F}_n$ is image admissible Suslin if there exists a Suslin parameter space Y_n and a z_n parameterized family of maps $T_{z_n} : Y_n \to \mathcal{F}_{z_n}$ such that the evaluation function

$$\Xi_n; Y_n \times Z_n \times Z \to \{0, 1\}$$

defined by $\Xi_n(y_n, z_n, z) = (T_{z_n}(y_n))(z)$ is jointly measurable in (y_n, z_n, z) with the product σ-algebra of the Borel sets of Y_n and Z_{n+1}. The inclusion of measurability considerations for the results in this section goes through in much the same way as for the VC theorem as presented by Dudley (1999). On the other hand in the proof of Theorem 15 in Section 5.1.2 measurability is crucial and so we consider it explicitly then.

There is a one to one mapping between functions $f : Z \to \{0, 1\}$ and their indicator sets $I(f) = \{z : f(z) = 1\}$. Throughout this paper we make this identification regularly without comment. In particular we identify $\mu(f) = \mu(I(f))$.

Definition 1 *For $n \leq m$ define $\mathcal{N}_n(z_m, \mathcal{F})$ to be the number of distinct dichotomies of the m points z_m generated as the functions vary over the union of the data dependent classes determined by all subsets of z_m containing n points. That is, let $I_f = \{z : f(z) = 1\}$ denote the set where the function is equal to one. Then $\mathcal{N}_n(z_m, \mathcal{F})$ is the number of different sets in*

$$\{\{z_m(1), ..., z_m(m)\} \cap I_f : f \in \mathcal{F}_{w_n}, w_n \subset z_m\}$$

where $w_n \subset z_m$ means that $\{w_n(i), i = 1, .., n\} \subseteq \{z_m(j), j = 1, .., m\}$. The shatter coefficient for a data dependent class $\mathcal{F}$ is defined as

$$S_{n/m}(\mathcal{F}) = \sup_{z_m} \mathcal{N}_n(z_m, \mathcal{F}).$$

We begin by citing Gat's (1999) observation that Vapnik's basic lemma regarding ghost samples still applies for data dependent classes. Let z_n denote the n-sample and z_{n_2} the ghost sample. Denote $z_m = (z_n, z_{n_2})$ with $m = n + n_2$. We write the empirical means

$$\hat{\mu}(f) = \hat{\mu}_1(f) = \frac{1}{n} \sum_{i=1}^{n} f(z_n(i))$$

and

$$\hat{\mu}_2(f) = \frac{1}{m-n} \sum_{i=1}^{m-n} f(z_{n_2}(i)).$$

Lemma 2 *(Gat)*

$$\mathcal{P}_{Z_n}\left(\sup_{f \in \mathcal{F}_{z_n}} \left|\mu(f) - \hat{\mu}_1(f)\right| > \epsilon\right) \leq 2\mathcal{P}_{Z_m}\left(\sup_{f \in \mathcal{F}_{z_n}} \left|\hat{\mu}_2(f) - \hat{\mu}_1(f)\right| > \epsilon - \frac{1}{m-n}\right)$$

Proof Just follow Vapnik's proof (Vapnik, 1998, page 132) and observe that the relevant n is n_2 and the data dependency makes no difference. ■

Now we state and prove our main theorem for data dependent classes.

Theorem 3 *For any $m > n$,*

$$\mathcal{P}_{Z_n}\left(\sup_{f \in \mathcal{F}_{z_n}} \left|\mu(f) - \hat{\mu}_1(f)\right| > \epsilon\right) \leq 2S_{n/m}(\mathcal{F})e^{2\epsilon}e^{-(\frac{1}{n}+\frac{1}{m-n})^{-1}\epsilon^2}.$$

Proof Gat (1999) proved the version of this theorem that counted the number of functions in $\{\mathcal{F}_{w_n}, w_n \subseteq z_m\}$ when $m = 2n$. We show this proof can generate bounds in terms of the shatter coefficient for the data dependent class. Consider

$$\mathcal{P}_{Z_m}\left(\sup_{f\in\mathcal{F}_{z_n}} \left|\hat{\mu}_2(f) - \hat{\mu}_1(f)\right| > \acute{\epsilon}\right)$$

where $\acute{\epsilon} = \epsilon - \frac{1}{m-n}$. Let Σ denote the set of permutations of the integers $(1, 2, ..., m)$ and use the same designation for the set of permutations induced on the m sample z_m. Since $\mathcal{P}_{Z_m}$ is invariant under Σ, for any permutation σ

$$\mathcal{P}_{Z_m}\left(\sup_{f\in\mathcal{F}_{z_n}} \left|\hat{\mu}_2(f) - \hat{\mu}_1(f)\right| > \acute{\epsilon}\right) = \mathcal{P}_{Z_m}\left(\sup_{f\in\mathcal{F}_{\sigma(z_m)_1}} \left|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)\right| > \acute{\epsilon}\right)$$

where $\sigma(z_m) = (\sigma(z_m)_1, \sigma(z_m)_2)$ and $\sigma(z_m)_1$ is of length n.

Place the uniform probability distribution on Σ. It then follows that

$$\mathcal{P}_{Z_m}\left(\sup_{f\in\mathcal{F}_{z_n}} \left|\hat{\mu}_2(f) - \hat{\mu}_1(f)\right| > \acute{\epsilon}\right) = \mathcal{P}_{Z_m,\Sigma}\left(\sup_{f\in\mathcal{F}_{\sigma(z_m)_1}} \left|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)\right| > \acute{\epsilon}\right)$$

$$= \int \left(\mathcal{P}_{\Sigma|Z_m}\Big(\sup_{f\in\mathcal{F}_{\sigma(z_m)_1}} \left|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)\right| > \acute{\epsilon} \;\Big|\; Z_m\Big)\right) d\mathcal{P}_{Z_m} \tag{1}$$

but for fixed Z_m

$$\mathcal{P}_{\Sigma|Z_m}\left(\sup_{f\in\mathcal{F}_{\sigma(z_m)_1}} \left|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)\right| > \acute{\epsilon} \;\Big|\; Z_m\right) \leq$$

$$S_{n/m}(\mathcal{F}) \sup_{f\in\bigcup_{\sigma\in\Sigma}\mathcal{F}_{\sigma(z_m)_1}} \mathcal{P}_{\Sigma|Z_m}\left(\left|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)\right| > \acute{\epsilon} \;\Big|\; Z_m\right). \tag{2}$$

To bound

$$\mathcal{P}_{\Sigma|Z_m}\left(\left|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)\right| > \acute{\epsilon} \;\Big|\; Z_m\right)$$

we map to the hypergeometric distribution in the following fashion. For f fixed, let l be the number of indices such that $f(z_m(i)) = 1$. For fixed σ let k denote the number of indices $i = 1, .., n$ such that $f(\sigma(z_m)_1(i)) = 1$. For each combination of k indices chosen from l and $n-k$ chosen from $m-l$, there are $n!(m-n)!$ permutations. The number of combinations which obtain k indices from the l and $n-k$ from the $m-l$ is $\binom{l}{k}\binom{m-l}{n-k}$ so the number of corresponding permutations is $\binom{l}{k}\binom{m-l}{n-k}n!(m-n)!$. Consequently, the fraction of permutations with k ones in the first n indices and the remaining $l-k$ ones in the remaining $m-n$ indices is

$$\frac{\binom{l}{k}\binom{m-l}{n-k}n!(m-n)!}{m!} = \frac{\binom{l}{k}\binom{m-l}{n-k}}{\binom{m}{n}}$$

and so is distributed like the hypergeometric random variable $K(n,m,l)$ of how many defectives are selected when n items are randomly selected without replacement from a population of m items of which l are defective. For each of these permutations

$$\hat{\mu}_{\sigma(z_m)_1}(f) - \hat{\mu}_{\sigma(z_m)_2}(f) = \frac{k}{n} - \frac{l-k}{m-n} = \frac{m}{n(m-n)}\Big(k - \frac{ln}{m}\Big).$$

Consequently,

$$\mathcal{P}_{\Sigma|Z_m}\Big(|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)| > \acute{\epsilon} \;\Big|\; Z_m\Big) = \mathcal{P}_{K(n,m,l)}\left(\Big|k - \frac{ln}{m}\Big| > \frac{n(m-n)}{m}\acute{\epsilon}\right).$$

Applying Serfling's result (Serfling, 1974) on concentration of sampling without replacement to the hypergeometric distribution

$$\mathcal{P}_{K(n,m,l)}\left(\Big|\frac{k}{n} - \frac{l}{m}\Big| > t\right) \leq e^{-\frac{2nm}{m-n+1}t^2}$$

with $t = \frac{m-n}{m}\acute{\epsilon}$ we obtain

$$\mathcal{P}_{\Sigma|Z_m}\Big(|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)| > \acute{\epsilon} \;\Big|\; Z_m\Big) \leq e^{-2\frac{n}{m}\frac{(m-n)^2}{m-n+1}\acute{\epsilon}^2}.$$

We use the crude lower bound $\frac{1}{m-n+1} \geq \frac{1}{2}\frac{1}{m-n}$, losing almost a factor of 2 in exponent, to obtain the simpler formula

$$\mathcal{P}_{\Sigma|Z_m}\Big(|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)| > \acute{\epsilon} \;\Big|\; Z_m\Big) \leq e^{-(\frac{1}{n}+\frac{1}{m-n})^{-1}\acute{\epsilon}^2}.$$

Consequently, from Equation 2 we obtain

$$\mathcal{P}_{\Sigma|Z_m}\left(\sup_{f\in\mathcal{F}_{\sigma(z_m)_1}} \big|\hat{\mu}_{\sigma(z_m)_2}(f) - \hat{\mu}_{\sigma(z_m)_1}(f)\big| > \acute{\epsilon}\right) \leq S_{n/m}(\mathcal{F})e^{-(\frac{1}{n}+\frac{1}{m-n})^{-1}\acute{\epsilon}^2} \tag{3}$$

Substituting $\acute{\epsilon} = \epsilon - \frac{1}{m-n}$ we bound

$$-\Big(\frac{1}{n} + \frac{1}{m-n}\Big)^{-1}\acute{\epsilon}^2 \leq 2\epsilon - \Big(\frac{1}{n} + \frac{1}{m-n}\Big)^{-1}\epsilon^2$$

so that combining with the bound from Lemma 2 and Equation 1, we obtain

$$\mathcal{P}_{Z_n}\left(\sup_{f\in\mathcal{F}_{z_n}} \big|\mu(f) - \hat{\mu}(f)\big| > \epsilon\right) \leq 2S_{n/m}(\mathcal{F})e^{2\epsilon}e^{-(\frac{1}{n}+\frac{1}{m-n})^{-1}\epsilon^2}$$

and the proof is finished.

■

Before we proceed, let us mention that the data dependent version of Vapnik and Chervonenkis' lemma inducing bounds on estimation error from bounds on error deviance (as stated by Devroye et al., 1996) also holds. The proof is the same as the data independent version.

Lemma 4 *Let $\hat{f}(z_n) = \arg\min_{f\in\mathcal{F}_{z_n}} \hat{\mu}(f)$ denote the empirical mean minimizer in the data dependent class $\mathcal{F}_{z_n}$ and let $\mu^*(z_n) = \inf_{f\in\mathcal{F}_{z_n}} \mu(f)$ be the data dependent optimum. Then*

$$\mu\big(\hat{f}(z_n)\big) - \mu^*\big(z_n\big) \leq 2\sup_{f\in\mathcal{F}_{z_n}} \big|\hat{\mu}(f) - \mu(f)\big|$$

3. Classification

We now consider classification. Let $Z = (X, Y)$ with $Y = \{0, 1\}$ where X is a metric space with Borel σ-algebra inducing a σ-algebra on Z in the obvious way. We assume additionally that X is Polish so that Z is also Polish and so Suslin. Again we consider independent samples so all n-sample probability measures will be the product measures. Let the Borel probability measure be decomposed as

$$\mu(\acute{f}) = \mu_1\big(\acute{f}(\cdot, 1)\big)p(1) + \mu_0\big(\acute{f}(\cdot, 0)\big)p(0)$$

for two Borel probability measures μ_1 and μ_0 where we use the notation $p(1) = p(y = 1)$ and $p(0) = p(y = 0)$ for the probability on $Y = \{0, 1\}$. We require that each function correspond to a classifier hypothesis. That is, each function $\acute{f} : Z \to \{0, 1\}$ satisfies

$$\acute{f}(x, y) = I_{\{f(x)=y\}}$$

for some classifier $f : X \to \{0, 1\}$. Equivalently $\acute{f}$ must satisfy

$$\acute{f}(x, 1) + \acute{f}(x, 0) = 1$$

and so is determined by some classifier $f : X \to \{0, 1\}$ by $f(x) = \acute{f}(x, 1)$. For any class $\mathcal{C}$ of classifiers $f : X \to \{0, 1\}$ let $\acute{\mathcal{C}}$ be the class of functions defined by $\acute{f}(x, 1) = f(x)$ and $\acute{f}(x, 0) = 1 - f(x)$. The decomposition of the measure combined with the constraints for classifier hypotheses determines the formula for generalization error

$$e(f) = \mu(1 - \acute{f}) = \big(1 - \mu_1(f)\big)p(1) + \mu_0(f)p(0). \tag{4}$$

Now we define

$$e^*(z_n) = \inf_{f \in \mathcal{F}_{z_n}} e(f)$$

to be the optimal generalization error in the data dependent class. Also define

$$e^*(z_\infty) = \limsup_{n \to \infty} e^*(z_n)$$

to be its infinite sample limit. Let $\mathcal{B}$ denote the Borel measurable sets. Define

$$e_{\mathcal{B}} = \inf_{f \in \mathcal{B}} e(f)$$

to be the Bayes error.

4. Data Dependent Classes for Classification

In this section we introduce and analyze some specific data dependent hypothesis classes for classification. The first is a class we call *simple linear classifiers*, and the second is a family of classes that we refer to collectively as *multi-sphere classifiers* (see Cannon and Cowen, 2000; Marchette and Priebe, 2001; Priebe, DeVinney, and Marchette, 2001 for related work). In each case we derive bounds on their shatter coefficients and establish hardness results for the computational complexity of empirical risk minimization. Throughout this section we let $Z = (X, Y)$ with X Polish so that Z is also Polish and so Suslin. Also, our data dependent classes $\acute{\mathcal{C}}$ are defined through function classes $\mathcal{C}$ of functions $X \to \{0, 1\}$ as described in the previous section. With this definition the shatter coefficient for $\acute{\mathcal{C}}$ is the same as for $\mathcal{C}$.

4.1 Simple linear classifiers

Here $X = \Re^d$ with metric determined by the usual inner product. Let the data dependent class $\mathcal{C}_{x_n}$ be the subset of linear classifiers whose orientations are determined by the pairwise differences between samples, that is

$$\mathcal{C}_{x_n} = \{f,\ 1-f : f(x) = H\left((x_n(i) - x_n(j)) \cdot x + b\right), i, j \le n, b \in \Re\}$$

where H is the heaviside function. The shatter coefficient is determined as follows. Recall that in the determination of the shatter coefficient we need to include a ghost sample.

Consider linear classifiers whose orientations are determined by a single sample pair $(x(i), x(j))$. The class of sets $\{x : (x(i) - x(j)) \cdot x + b < 0\}$ is ordered by subset relation as we vary the threshold b and so has VC dimension equal to one (Devroye et al., 1996), and so the number of subsets of the $2n$ points is at most $2(2n+1)$. For any sample of size $2n$ we find at most $\binom{2n}{2} = n(2n-1)$ unique sample pairs, so we can bound $S_{2n/2n}(\mathcal{C})$ by $8n^3 - 2n$. Since each function in $\mathcal{C}_{x_n}$ is in $\mathcal{C}_{x_{2n}}$, $\quad S_{n/2n}(\mathcal{C}) \le S_{2n/2n}(\mathcal{C}) \le 8n^3 - 2n$. Note that for this simple class the shatter coefficient is independent of dimension. This class also admits tractable learning algorithms in that empirical risk minimization over $\mathcal{C}_{x_n}$ can be solved in polynomial time. To see this consider the run time of the brute force algorithm which examines all $O(n^2)$ sample pairs and determines the optimal threshold for each. Since it takes $O(nd + n\log n)$ to determine the threshold for each pair the overall run time is $O(n^3(d + \log n))$.

4.2 Multi-sphere classifiers

For the computational considerations we let $X = \Re^d$ with metric determined by the usual inner product. Let z_{n_1} be the ordered subset of z_n with $y = 1$ where the ordering is induced from z_n, and the random variable n_1 denotes its size. We define the *single sphere* data dependent class as

$$\mathcal{C}_{z_n} = \{f : f(x) = H\left(r - \|x - x_{n_1}(i)\|\right), i \le n_1, r \in \Re_+\}$$

That is, x is assigned the label 1 if it falls in a closed ball $B(x_{n_1}(i), r) \subseteq X$ of radius r centered at one of the data samples with label $y = 1$, and 0 otherwise. Determination of the shatter coefficient parallels the procedure above. The class of sets $\{B(x, r), r \ge 0\}$, is ordered by subset relation and so has VC dimension equal to one, and so the number of subsets of the $2n-1$ remaining points (which include the ghost sample) is at most $2n$. Now, for any sample of size $2n$ we have at most $2n$ unique centers, so we can bound $S_{2n/2n}(\mathcal{C})$ by $(2n)^2$. Since each function in $\mathcal{C}_{z_n}$ is in $\mathcal{C}_{z_{2n}}$, $\quad S_{n/2n}(\mathcal{C}) \le S_{2n/2n}(\mathcal{C}) \le (2n)^2$. Note that once again the shatter coefficient is independent of dimension. In addition, empirical risk minimization over $\mathcal{C}_{z_n}$ can be solved in polynomial time since the brute force algorithm, which examines all $2n$ centers and determines the optimal radius for each, has an overall run time of $O(n^2(d + \log n))$.

Now consider an extension of $\mathcal{C}_{z_n}$ to multiple spheres. We define

$$\mathcal{C}^q_{z_n} = \left\{ f : f = \bigcup_{i=1}^{q} B\left(x_q(i), r_i\right),\ x_q \subseteq x_{n_1} \right\}$$

which is the set of functions corresponding to the union of q closed balls of q possibly different radii centered at q of the data points x_{n_1}, where $q \leq n$. To bound the shatter coefficient for this class recall that the number of dichotomies induced by a single ball placed at one of the data points is at most $2n$. Then we rewrite

$$\mathcal{C}^q_{z_n} = \bigcup_{x_q \subseteq x_{n_1}} \bigsqcup_{i=1}^{q} B\left(x_q(i), r_i\right)$$

where $\bigcup$ denotes the union of classes of sets and $\bigsqcup$ denotes the operation of forming unions of sets to form new classes. As shown by Devroye et al. (1996, page 219), the shatter coefficient of $A_1 \bigcup A_2$ is bounded by the sum of the shatter coefficients of A_1 and A_2 and the shatter coefficient of $A_1 \bigsqcup A_2$ is bounded by the product of the shatter coefficients of A_1 and A_2. Consequently, for q balls placed at q points the shattering of $2n$ points is bounded by $(2n)^q$ and for the $\binom{2n}{q}$ choices of subsets of size q from $2n$ points the shatter coefficient for q balls on $2n$ points is at most $\binom{2n}{q}(2n)^q \leq (2n)^{2q}$. Therefore $S_{n/2n}(\mathcal{C}^q) \leq (2n)^{2q}$. Note that the shatter coefficient remains independent of dimension.

We now discuss a family of multi-sphere classifiers which are obtained by varying the data dependency. Specifically we consider variants of the q-sphere class where we vary our specification of the q centers and their radii. The classes we consider include all combinations where

1. the q centers may be any subset of x_{n_1}, or
2. the q centers are a pre-determined subset of x_{n_1},

and

1. a different radius is chosen for each center, or
2. a single radius (same for all centers) must be chosen, or
3. the radii are pre-determined (i.e. fixed ahead of time).

The simplest case is where both the index set of center points and the radii are pre-determined. This class has only one function and so its shatter coefficient is 1 and its computational requirements are trivial.

The next level of sophistication is when $q = 1$. We have already considered the case where both the center sample index and its radius are variable. The tables below give bounds on the shatter coefficient and computational requirements for empirical risk minimization for all combinations with $q = 1$.

Shatter Coefficient Bounds $q = 1$		
	Radius Dependency	
Center Index Dependency	Fixed	Variable
Fixed	1	$2n$
Variable	$2n$	$(2n)^2$

Computation Bounds $q = 1$		
	Radius Dependency	
Center Index Dependency	Fixed	Variable
Fixed	$O(1)$	$O(dn \log n)$
Variable	$O(dn^2)$	$O(n^2(d + \log n))$

The next variant we consider is one where the $q > 1$ center sample indices are fixed and we allow a single variable radius (the same for each ball). It is easy to show that the shatter coefficient for this variation is bounded by $2n$, and the empirical error can be minimized with a brute force algorithm that runs in $O(nq(d + \log n))$ time.

All remaining variants allow either the $q > 1$ center sample indices to vary or the $q > 1$ radii to vary. The table below provides a summary of shatter coefficient bounds for these variants.

Shatter Coefficient Bounds $q > 1$		
	Radius Dependency	
Center Index Dependency	Fixed	Variable
Fixed	1	$(2n)^q$
Variable	$(2n)^q$	$(2n)^{2q}$

We now show that any variant that requires $q > 1$ to be determined algorithmically is NP-Hard. The formal definitions for these problems are stated as prize collecting versions of the class cover problem (Cannon & Cowen, 200).

Definition 5 (PRIZE COLLECTING CLASS COVER: PC-CC) *Given an n-sample $x_n = (x_{n_0}, x_{n_1})$, $x_n(i) \in X$, a distance function $d : X \times X \to \Re^+$, and a positive integer $q \le n_1$. Determine a subset $x_q \subseteq x_{n_1}$ of size q and a set of radii r_q, $r_q(i) \in \Re^+$, one for each sample in x_q that minimizes the error*

$$e(x_q) = |x_{n_1} \cap \bar{x}_c| + |x_{n_0} \cap x_c|$$

where x_c is the set of points covered by the q balls,

$$x_c = \{x_n(i) : \exists x_{n_1}(j) \in x_q \text{ such that } d(x_n(i), x_{n_1}(j)) \le r_{n_1}(j)\}$$

and $\bar{x}_c = x_n \setminus x_c$.

Definition 6 (SINGLE RADIUS CLASS COVER: SRCC) *This problem is the same as PC-CLASS COVER except that every sample in x_q is forced to use the same radius, so only a single radius value must be determined.*

Definition 7 (FIXED RADIUS CLASS COVER: FRCC) *This problem is the same as PC-CLASS COVER except that the radii are fixed, i.e. a fixed set of radii r_{n_1}, one for each sample in x_{n_1}, is provided as part of the problem instance, and we must determine only x_q.*

Definition 8 (FIXED CENTER CLASS COVER: FCCC) *This problem is the same as PC-CLASS COVER except that the centers are fixed, i.e. x_q is provided as part of the problem instance, and we must determine only r_q.*

The last variant presented involves fixing $q > 1$ centers and algorithmically determining the corresponding radii. We do not present a hardness proof for this version but conjecture that even this simple variant is hard for $q > 1$. The following theorem gives hardness results for the first three variants.

Theorem 9 *PC-CC, SRCC and FRCC are NP-Hard.*

Proof Our proofs are by reduction from K-CENTER (Hochbaum, 1997, Section 9.4.1).

Definition 10 (K-CENTER: KC) *Given a weighted graph $G = (V, E)$ with $|V| = n$, and a positive integer $k \leq n$. Let $\{w_{i,j}\}$ represent the set of shortest path distances between vertices v_i and v_j in V. Find a subset of vertices S of size k so that the greatest distance among all vertices in V to a nearest vertex in S is minimized. More formally,*

$$\min_{S \subset V, |S|=k} \{ \max_{v_i \in V \setminus S} \min_{v_j \in S} w_{i,j} \}.$$

We begin with the reduction to PC-CC, and then show how the same reduction can be used for both SRCC and FRCC as well.

Notice that the value of the objective function will always be one of $\binom{n}{2}$ possible shortest path distances. We will generate at most $\binom{n}{2}$ instances of PC-CC. First, order the $\binom{n}{2}$ distances $w_{i,j}$. Let $d_{(i)}$ be the i^{th} smallest distance and let ϵ be half of the smallest nontrivial pairwise difference between the distances,

$$\epsilon = \min_{d_{(i)} \neq d_{(j)}} \frac{1}{2} |d_{(i)} - d_{(j)}|$$

Now we generate a copy of the weighted graph $G(V, E)$ and color every vertex blue. Add n red vertices, each with an edge to a corresponding blue vertex that has weight $d_{(\binom{n}{2})} + \epsilon$. Note that each blue vertex now has a red vertex whose shortest path distance is $d_{(\binom{n}{2})} + \epsilon$ and no red vertices closer. The shortest path distances for all vertices in this modified graph satisfy the properties of a metric and can thus be embedded as points in a Euclidean space of finite dimension in polynomial time (Young, 1938). These points serve as inputs to the PC-CC problem where we make a one to one correspondence between the blue (red) vertices and the points in x_{n_1} (x_{n_0}). Now set $q = k$ and solve the PC-CC problem. Observe that this version of PC-CC is trivially solved with zero error.

We repeat the process now using $d_{(\binom{n}{2}-1)}$. The red vertices are now "one step" closer. This time we have no guarantee of existence of a zero error solution. If the error is zero, then we repeat the process bringing the red nodes another step closer. We continue until we reach a PC-CC solution with nonzero error. As soon as this occurs the blue vertices corresponding to the points x_q of the most recent zero error solution provide to an optimal solution to the original K-CENTER instance. Indeed if this were not the case then any superior solution to the K-CENTER instance would lead to a zero error solution for the most current version of PC-CC. If it turns out that all $\binom{n}{2}$ versions of PC-CC have zero error solutions then at least $n - k$ distances must be the same and the latest (final) solution is optimal for the k-center problem.

Hence if we could solve PC-CC in polynomial time, $T_{pc-cc}(n_0+n_1)$, then we could solve K-CENTER on n vertices in $O(n^2 \cdot T_s(n) \cdot T_e(n) \cdot T_{pc-cc}(n))$ time, where $T_s(n)$ is the time required to compute the shortest path distances for the modified graph, and $T_e(n)$ is the time required to embed the modified graph into a Euclidean space.

The same reduction can be used to prove hardness for SRCC by noting that at each step, as we move the red vertices closer, the set of optimal solutions (e.g. points and their radii) always includes an equal radius solution. For example at the step where $d_{(i)}$ is used to determine the new edge weights for the modified graph, if there is a zero error solution then there is a zero error solution where the radii for all q blue points are set to $d_{(i)}$ (at this radius each of the q blue points covers as many other blue points as possible without introducing errors). So the reduction works when we substitute SRCC for PC-CC at each step.

The reduction also works when we substitute FRCC for PC-CC. This follows directly from the observation that if there is a zero error solution, then there is a zero error solution whose radii are all $d_{(i)}$. This means that we can simply set all the radii in r_{n_1} to $d_{(i)}$ and solve FRCC (instead of PC-CC).

■

In summary, we have provided bounds on both the structural and computational complexity for a family of spherical classifiers. However, we have not provided a procedure for selecting the best member from this family. This is the topic of the next section.

5. Structural Risk Minimization

In classification it often not clear what hypothesis space to choose. Vapnik (1998) addressed this problem by considering a sequence of hypothesis spaces large enough to contain a good classifier. The structural risk minimization procedure (Vapnik, 1998) is a method designed to find this good classifier even though classifier complexity of the sequence is large. We now present a structural risk minimization procedure and prove convergence for a sequence of data dependent classifier spaces. Although utilized in the construction of classifiers, we present the analysis in terms of empirical process theory. We consider a sequence $\mathcal{F} = \{\mathcal{F}^q, q = 1, ...\}$ of data dependent classes $\mathcal{F}^q$. The symbol $\mathcal{F}$ used here is the same as we used for the data dependent class before. However, now the data dependent classes in the sequence are denoted $\mathcal{F}^q$ and $\mathcal{F}$ is the sequence so there should be no confusion. We say that $\mathcal{F}$ is image admissible Suslin if each $\mathcal{F}^q$ is. For simplicity of presentation we set $m = 2n$. Let $S_{q,n}$ denote bounds on the shatter coefficients $S_{n/2n}(\mathcal{F}^q)$. Our method is similar to that presented by Vapnik (1998) (also see Devroye, Györfi, and Lugosi, 1996). Given a training set z_n we select a function $\tilde{f}^q_{z_n}$ from every class $\mathcal{F}^q_{z_n}$ which minimizes empirical mean over the class. From these we select the function $f^*_{z_n}$ that minimizes over q the complexity penalized function:

$$\tilde{\mu}(\tilde{f}^q_{z_n}) = \hat{\mu}(\tilde{f}^q_{z_n}) + r(q, n)$$

where $\hat{\mu}$ is the empirical mean and $r(q,n)$ is the penalty term

$$r(q,n) = \sqrt{2\frac{\log en}{n\log n}\log(S_{q,n})}$$

Let

$$\mu_q^*(z_n) = \inf_{f\in\mathcal{F}_{z_n}^q} \mu(f)$$

define the function $\mu_{q,n}^*$ and

$$\mu^*(z_\infty) = \inf_q \limsup_{n\to\infty} \mu_q^*(z_n)$$

define the function μ_∞^*. If we let the intermediate

$$\mu_q^*(z_\infty) = \limsup_{n\to\infty} \mu_q^*(z_n)$$

define the function μ_q^*, then

$$\mu^*(z_\infty) = \inf_q \mu_q^*(z_\infty).$$

$\mu_{q,n}^*$ can be extended to a function with the same name on Z_∞ in the obvious way. In general we make this extension without notice. When we apply this theory to classification we make use of Equation 4 for the generalization error

$$e(f) = \mu(1-\acute{f}) = p(1) + \mu_0(f)p(0) - \mu_1(f)p(1)$$

and the corresponding definitions for $e_{q,n}^*$,e_∞^*, e_q^*.

Let $\mathcal{SE}$ denote the class of subexponential functions on the positive integers. That is $g\in\mathcal{SE}$ if and only if

$$\sum_{n>0} g(n)e^{-n\alpha} < \infty$$

for all $\alpha > 0$. This class includes the polynomials, polynomials with logarithms and even such functions as $e^{\frac{n}{\log n+1}}$. We use the bound from Theorem 3 with $m = 2n$

$$\mathcal{P}_{Z_n}\Big(\sup_{f\in\mathcal{F}_{z_n}^q} \big|\mu(f) - \hat{\mu}(f)\big| > \epsilon\Big) \le 2S_{q,n}e^{2\epsilon}e^{-\frac{n\epsilon^2}{2}}.$$

The proof of the the following structural risk minimization theorem closely resembles a theorem presented by Devroye et al. (1996, Theorem 18.2). However, because of the data dependencies, the conditions of the theorem are somewhat different than usual. Let *wp1* denote the terminology “with probability 1”.

Theorem 11 *Suppose that $\mathcal{F}$ is image admissible Suslin and there exists bounds $S_{q,n}$ for the shatter coefficients such that $\sum_q S_{q,n}^{-\frac{1}{\log n}} \in \mathcal{SE}$ and $S_{q,n}\in\mathcal{SE}$ for every q. Let $f_{z_n}^*$ be determined by the structural risk minimization procedure. Then $\limsup_{n\to\infty}\mu(f_{z_n}^*) \le \mu_\infty^*$ wp1.*

5.1 Structural assumptions on $\mathcal{F}$

This structural risk theorem was very general and assumed only that $\mathcal{F}$ was image admissible Suslin, but assumed nothing about the relationship between $\mathcal{F}^{q_1}_{n_1}$ and $\mathcal{F}^{q_2}_{n_2}$. Although stronger limit theorems can be made under structural assumptions about $\mathcal{F}$ we analyze this question at present only in terms of μ^*_∞.

5.1.1 z-increasing $\mathcal{F}$

Definition 12 *We say that $\mathcal{F}$ is z-increasing if $\mathcal{F}^q_{z_n} \subseteq \mathcal{F}^q_{z_{\acute{n}}}$ when $z_n \subseteq z_{\acute{n}}$ for each q.*

If $\mathcal{F}$ is z-increasing, then $\limsup_{n\to\infty} \mu^*_q(z_n) = \inf_n \mu^*_q(z_n)$ and so

$$\mu^*(z_\infty) = \inf_q \inf_n \mu^*_q(z_n) = \inf_q \inf_n \inf_{f\in\mathcal{F}^q_{z_n}} \mu(f)$$

but since generally $\inf_w \inf_v = \inf_{w,v}$ we obtain that

$$\mu^*(z_\infty) = \inf_{f\in\mathcal{F}} \mu(f)$$

and since $\mu(f^*_{z_n}) \geq \inf_{f\in\mathcal{F}} \mu(f) = \mu^*(z_\infty)$ we obtain the following much stronger theorem.

Theorem 13 *Suppose the data dependent sequence $\mathcal{F}$ is z-increasing and image admissible Suslin and there exists bounds $S_{q,n}$ for the shatter coefficients such that $\sum_q S_{q,n}^{-\frac{1}{\log n}} \in \mathcal{SE}$ and $S_{q,n} \in \mathcal{SE}$ for every q. Let $f^*_{z_n}$ be determined by the structural risk minimization procedure. Then $\mu(f^*_{z_n}) \underset{n\to\infty}{\longrightarrow} \inf_{f\in\mathcal{F}} \mu(f)$ wp1.*

An interesting thing also happens to the shatter coefficients.

Lemma 14 *Suppose the data dependent class $\mathcal{G}$ is z-increasing. Then*

$$\mathcal{N}_n(z_m, \mathcal{G}) \leq \mathcal{N}_{\acute{n}}(z_m, \mathcal{G})$$

and

$$S_{n/m}(\mathcal{G}) \leq S_{\acute{n}/m}(\mathcal{G})$$

for all $n \leq \acute{n} \leq m$. In particular,

$$S_{n/m}(\mathcal{G}) \leq S_{m/m}(\mathcal{G}).$$

Proof The z-increasing assumption $\mathcal{G}_{z_n} \subseteq \mathcal{G}_{z_{\acute{n}}}$ when $z_n \subseteq z_{\acute{n}}$ implies that

$$\mathcal{N}_n(z_m, \mathcal{G}) = \left| \{\{z_m(1), ..., z_m(m)\} \cap I_f : f \in \mathcal{G}_{w_n}, w_n \subset z_m\} \right|$$

$$\leq \left| \{\{z_m(1), ..., z_m(m)\} \cap I_f : f \in \mathcal{G}_{z_{\acute{n}}}, z_{\acute{n}} \subset z_m\} \right| = \mathcal{N}_{\acute{n}}(z_m, \mathcal{G})$$

and taking the supremum over z_m finishes the proof. ∎

We define the data dependent VC dimension of order m as

$$VC_m(\mathcal{G}) = \max\{n : \exists z_n \subseteq z_m : \mathcal{G}_{z_m} \text{shatters } z_n\}.$$

In words $VC_m(\mathcal{G})$ is the size of the largest subset of some m points which is shattered by the data dependent class on those m points. The Sauer lemma (Devroye et al., 1996) can be directly applied. In particular

$$S_{n/m}(\mathcal{G}) \leq S_{m/m}(\mathcal{G}) \leq m^{VC_m(\mathcal{G})} + 1.$$

5.1.2 Symmetric $\mathcal{F}$

We say that $\mathcal{F}$ is z-symmetric if

$$\mathcal{F}^q_{z_n} = \mathcal{F}^q_{\sigma(z_n)}$$

for every permutation σ of the n points. In this section we will show that when $\mathcal{F}$ is z-symmetric, μ^*_∞ is measurable and constant *wp1*. As we mentioned in the introduction, such measurability considerations are appropriate through most of this work, but feel these technicalities would obscure the presentation. However, the following result depends so critically on measurability assumptions that we include these technicalities here. We now assume that $\mathcal{F}$ is z-symmetric. Before we state the next theorem we need to introduce some terminology. A universally measurable set in a measurable space is a set M which is measurable for the completion of every measure. In particular, for each measure ν their exists measurable sets

$$A \subset M \subset B$$

such that $\nu(A) = \nu(B)$. Consequently, we can and will be a little sloppy and say $\nu(M) = \nu(A) = \nu(B)$ for the universally measurable set M. A universally measurable function between measurable spaces is one such that the preimage of a measurable set is a universally measurable set. For a specific measure, we can consider universally measurable sets as measurable for the completion of the measure (Ash, 1972). Indeed not only do they form a σ-algebra, but it is elementary to show that if $\{M_n\}$ is a countable collection of universally measurable sets such that

$$A_n \subset M_n \subset B_n$$

with $\nu(A_n) = \nu(B_n)$ for each n, then $\bigcup_n M_n$ and $\bigcap_n M_n$ are both universally measurable,

$$\bigcup_n A_n \subseteq \bigcup_n M_n \subseteq \bigcup_n B_n,$$

$$\bigcap_n A_n \subseteq \bigcap_n M_n \subseteq \bigcap_n B_n,$$

$\nu(\bigcup_n A_n) = \nu(\bigcup_n B_n)$, and $\nu(\bigcap_n A_n) = \nu(\bigcap_n B_n)$.

Theorem 15 *Suppose that Z is a Suslin space with Borel measure μ. Suppose that the data dependent sequence $\mathcal{F}$ is z-symmetric and image admissible Suslin. Then μ^*_∞ is universally measurable and μ^*_∞ = constant wp1.*

Proof Since $\mathcal{F}$ is image admissible Suslin, for fixed q, Ξ_n^q is measurable. Since Ξ_n^q is positive Fubini's theorem tells us that

$$\mu\left(f(T_q(y_n)_{z_n})\right) = \mu\left(\Xi^q(y_n, z_n, \cdot)\right)$$

is jointly measurable. By the theorem of Sainte-Beuve as presented by Dudley (1999)

$$\mu_q^*(z_n) = \inf_{f \in \mathcal{F}_{z_n}} \mu(f) = \inf_{y_n \in Y_n} \mu\left(f(T_q(y_n)_{z_n})\right)$$

is universally measurable on Z_n and extends naturally to a universally measurable function on Z_∞. We work on Z_∞ for the rest of this proof. In particular, for every c the set $\{\mu_{q,n}^* \geq c\}$ is universally measurable. Consequently, there exists measurable sets

$$A_{q,n} \subset \{\mu_{q,n}^* \geq c\} \subset B_{q,n}$$

such that $\mathcal{P}_{Z_\infty}(A_{q,n}) = \mathcal{P}_{Z_\infty}(B_{q,n})$. Since $\mathcal{F}$ is z-symmetric $\{\mu_{q,n}^* \geq c\}$ is invariant under permutation of the first n positions in Z_∞. If we let S_n^+ denote the upper symmetric envelope of a set (the union over all permutations of the first n position indices) and S_n^- the lower symmetric envelope it is clear from the fact that universally measurable sets behave like measurable sets with respect to unions and intersections that $\mathcal{A}_{q,n} = S_n^+(A_{q,n})$ and $\mathcal{B}_{q,n} = S^-(B_{q,n})$ are measurable, symmetric in there first n positions, and

$$A_{q,n} \subseteq \mathcal{A}_{q,n} \subseteq \{\mu_{q,n}^* \geq c\} \subseteq \mathcal{B}_{q,n} \subseteq B_{q,n}.$$

Consequently $\mathcal{P}_{Z_\infty}(\mathcal{A}_{q,n}) = \mathcal{P}_{Z_\infty}(\mathcal{B}_{q,n})$. Now consider

$$\mu^*(z_\infty) = \inf_q \limsup_{n \to \infty} \mu_q^*(z_n) = \inf_q \inf_N \sup_{n \geq N} \mu_q^*(z_n)$$

so that

$$\{\mu_\infty^* \geq c\} = \bigcap_q \bigcap_N \bigcup_{n \geq N} \{\mu_{q,n}^* \geq c\}.$$

It is clearly universally measurable and if we define $\mathcal{A}_\infty = \bigcap_q \bigcap_N \bigcup_{n \geq N} \mathcal{A}_{q,n}$ and $\mathcal{B}_\infty = \bigcap_q \bigcap_N \bigcup_{n \geq N} \mathcal{B}_{q,n}$,

$$\mathcal{A}_\infty \subseteq \{\mu_\infty^* \geq c\} \subseteq \mathcal{B}_\infty$$

and

$$\mathcal{P}_{Z_\infty}(\mathcal{A}_\infty) = \mathcal{P}_{Z_\infty}(\mathcal{B}_\infty).$$

Since $\mathcal{A}_{q,n}$ is symmetric in its first n components, $\bigcup_{n \geq N} \mathcal{A}_{q,n}$ is symmetric in the first N. Consequently, $\bigcap_N \bigcup_{n \geq N} \mathcal{A}_{q,n}$ is symmetric for any finite permutation and the same is true for $\mathcal{A}_\infty$. This argument works just as well for $\mathcal{B}_\infty$. Therefore both $\mathcal{A}_\infty$ and $\mathcal{B}_\infty$ are measurable and symmetric. By the Hewitt-Savage Zero-One law (Shiryaev, 1980)

$$\mathcal{P}_{Z_\infty}(\mathcal{A}_\infty) = \mathcal{P}_{Z_\infty}(\mathcal{B}_\infty) = 0 \text{ or } 1$$

and the proof is finished. ■

Consequently, when $\mathcal{F}$ is both z-symmetric and z-increasing we obtain that $\mu(f_{z_n}^*)$ converges to $\mu_\infty^* = \inf_{f \in \mathcal{F}} \mu(f)$ *wp1* and that $\mu_\infty^* = \inf_{f \in \mathcal{F}} \mu(f) = \text{constant}$ *wp1*.

Note that in the course of proving this theorem we have shown that for a fixed symmetric image admissible Suslin data dependent class, the limit $\mu_\infty^* = \limsup_{n \to \infty} \mu^*(z_n)$ is constant *wp1*.

5.2 Structural risk minimization for multi-sphere classifiers

When the structural risk minimization theorem is applied it would be very nice to be able to characterize μ^*_∞. For example, when is μ^*_∞ the minimum value over the unconstrained sequence $\mathcal{UF} = \bigcup_{q,n,z_n \in Z_n} \mathcal{F}^q_{z_n}$ *wp1*? We can provide no general facts in this regard at present but in the this section we show that this is true for multi-sphere classifiers. We first show that e^*_∞ is the Bayes optimal and then show we can apply the Structural Risk Minimization Theorem 13. The analysis below works for closed or open balls, balls all of the same radii, and when the complements of the union of the balls is included in the class of functions.

Theorem 16 *Consider the data dependent multi-sphere classifiers on a Polish space X with Borel measure μ on $Z = (X, \{0, 1\})$. Let $e_{\mathcal{B}}$ denote the Bayes error. Let the structural risk minimization procedure be applied with $S_{q,n} = (2n)^{2q}$ to produce the classifier $f^*_{z_n}$. Then $e(f^*_{z_n}) \underset{n\to\infty}{\longrightarrow} e_{\mathcal{B}}$ wp1.*

Proof
To prove this theorem we have to do two things. The first is to show that e^*_∞ is universally measurable and that $e^*_\infty = e_{\mathcal{B}}$ *wp1*. The second is to show that we can apply the Structural Risk Minimization Theorem 13. We now proceed with the first.

Lemma 17 *Consider a Borel measurable random variable $Z = (X, \{0, 1\})$ with X Polish and Borel measure μ. Then for the data dependent multi-sphere classifiers, e^*_∞ is universally measurable and $e^*_\infty = e_{\mathcal{B}}$ wp1.*

Proof We now show that for the multi-sphere classifiers that the evaluation function Ξ^q is jointly measurable for each q. The data dependent classes $\acute{\mathcal{C}}$ of functions on Z are the data dependent classes $\bigcup_{q\le n} \acute{\mathcal{C}}^q$ where $\mathcal{C}^q_{x_n}$ is the set of functions on X which are one on the union of a set of q balls placed on the x component of the data samples. It is clear that $\acute{\mathcal{C}}$ is both z-symmetric and z-increasing. Consider the class $\mathcal{B}_{x_q}$ for a fixed set of q points x_q. We define the parameter space of radii $W_q = \Re^q_+$ and the parameterization that $T_{x_q}(w_q)(z) = \max_{1\le i\le q} I_{|x - x_q(i)| \le w_q(i)}$ where I is the indicator function. For each i the set $I_{|x-x_q(i)|\le w_q(i)}$ is closed and so measurable in W_q, Z_q, X. Consequently the evaluation function

$$\Xi^q(w_q, z_q, X) = T_{x_q}(w_q)(z) = \max_{1\le i\le q} I_{|x-x_q(i)|\le w_q(i)}$$

is jointly measurable. Since the measurable sets are closed under complement, the evaluation function for the class corresponding to the functions on Z is measurable on W_q, Z^q, Z. Consequently, the conditions of Theorem 15 are satisfied and so e^*_∞ is universally measurable and e^*_∞ is constant *wp1*. What is left is to show that this constant is $e_{\mathcal{B}}$.

To this end note that

$$e_{\mathcal{B}} \le e^*_q(z_n)$$

so that

$$e_{\mathcal{B}} \le e^*(z_\infty).$$

Now we prove that

$$e^*_\infty \le e_{\mathcal{B}}$$

wp1 which will then finish the proof. The idea is as follows. We keep an eye on the generalization error (Equation 4) $e(f) = p(1) + \mu_0(f)p(0) - \mu_1(f)p(1)$ and show that for any f we can use a data dependent placement of balls to form a classifier f^* such that $\mu_0(f^*)$ is never much larger than $\mu_0(f)$ while at the same time show that for large enough n with high probability $\mu_1(f^*)$ is not much smaller than $\mu_1(f)$. For the first part we utilize the continuity of the measure of a decreasing family of measurable sets and for the second the fact that the measure of a measurable set on a Polish space can be approximated from below by the measure of compact measurable sets. This compactness allows us to show that with high probability the randomly placed set of balls will have μ_1 measure not too much smaller than that of f. Consequently from the generalization error formula we see that with high probability we can control the increase in generalization error over the generalization error of any measurable set through a random placement of balls. We now carry out the details.

Let $\epsilon_1 > 0$. Then there exists a measurable f_{ϵ_1} such that

$$e(f_{\epsilon_1}) \leq e_{\mathcal{B}} + \epsilon_1. \tag{5}$$

Since X is Polish it follows (see e.g. Ash, 1972) that f_{ϵ_1} can be approximated from below by compact sets. That is, given any $\epsilon_2 > 0$ there is a compact $h_{\epsilon_2} \leq f_{\epsilon_1}$ such that

$$\mu_1(f_{\epsilon_1}) - \epsilon_2 \leq \mu_1(h_{\epsilon_2}) \leq \mu_1(f_{\epsilon_1}). \tag{6}$$

Since it is compact, for any covering of h_{ϵ_2} of open balls $B_i^{\epsilon_3}$ of radius ϵ_3 there is a finite subcovering. From this subcovering choose the positive measure subcovering such that $\mu_1(B_i^{\epsilon_3}) > 0$ for each ball in the subcover. Note that the positive measure subcover is not really a covering. However $\mu_1(\bigcup B_i^{\epsilon_3}) \geq \mu_1(h_{\epsilon_2})$ for the positive measure subcover. Consider the classifier $f_{z_n}^{2\epsilon_3} \in \mathcal{F}_{z_n}$ which places closed balls $\bar{B}^{2\epsilon_3}$ of radius $2\epsilon_3$ at each of the data points which satisfy $h_{\epsilon_2}(x) = 1$. The triangle inequality implies that

$$\bar{B}^{2\epsilon_3}(x) \supset B^{\epsilon_3}$$

whenever $x \in B^{\epsilon_3}$ so that if the sample z_n has at least one point in every ball $B_i^{\epsilon_3}$ of the positive measure subcover, then $f_{z_n}^{2\epsilon_3} \geq h_{\epsilon_2}$ and so

$$\mu_1(f_{z_n}^{2\epsilon_3}) \geq \mu_1(h_{\epsilon_2}). \tag{7}$$

From Equation 6 we see that for $\epsilon_2 < \mu_1(f_{\epsilon_1})$, $\mu_1(h_{\epsilon_2}) > 0$ and since the positive measure subcover is finite we know then for large enough n with high probability there will be at least one point in every ball. That is, given an $\epsilon_4 > 0$ there exists an $M(\epsilon_4)$ such that

$$\mathcal{P}_{Z_\infty}\Big(\mu_1(f_{z_n}^{2\epsilon_3}) \geq \mu_1(h_{\epsilon_2})\Big) \geq 1 - \epsilon_4$$

for all $n \geq M(\epsilon_4)$. Now $f_{z_n}^{2\epsilon_3} \leq N_{2\epsilon_3}(h_{\epsilon_2})$ where $N_\epsilon(A)$ is the set of all points at distance less that or equal to ϵ from the set A. Consequently, $\mu_0(f_{z_n}^{2\epsilon_3}) \leq \mu_0(N_{2\epsilon_3}(h_{\epsilon_2}))$ and so $\mu_0(f_{z_n}^{2\epsilon_3}) \leq \mu_0(N_{2\epsilon_3}(f_{\epsilon_1}))$. Since in addition, $\mu_1(h_{\epsilon_2}) \geq \mu_1(f_{\epsilon_1}) - \epsilon_2$, if we put this together in the generalization formula we obtain

$$\begin{aligned}\mathcal{P}_{Z_\infty}\Big(e(f_{z_n}^{2\epsilon_3}) \leq e(f_{\epsilon_1}) + p(0)\Big(\mu_0(N_{2\epsilon_3}(h_{\epsilon_2})) - \mu_0(h_{\epsilon_2})\Big)& \\ + p(0)\Big(\mu_0(h_{\epsilon_2}) - \mu_0(f_{\epsilon_1})\Big) + p(1)(\epsilon_2)\Big) &\geq 1 - \epsilon_4.\end{aligned}$$

Since $e(f_{\epsilon_1}) \leq e_{\mathcal{B}} + \epsilon_1$ and $h_{\epsilon_2} \leq f_{\epsilon_1}$ we further obtain

$$\mathcal{P}_{Z_\infty}\Big(e(f_{z_n}^{2\epsilon_3}) \leq e_{\mathcal{B}} + \epsilon_1 + p(0)\Big(\mu_0(N_{2\epsilon_3}(h_{\epsilon_2})) - \mu_0(h_{\epsilon_2})\Big) + p(1)(\epsilon_2)\Big) \geq 1 - \epsilon_4.$$

Since $e^*(z_n) \leq e(f_{z_n}^{2\epsilon_3})$ we obtain

$$\mathcal{P}_{Z_\infty}\Big(e_n^* \leq e_{\mathcal{B}} + \epsilon_1 + p(0)\Big(\mu_0(N_{2\epsilon_3}(h_{\epsilon_2})) - \mu_0(h_{\epsilon_2})\Big) + p(1)(\epsilon_2)\Big) \geq 1 - \epsilon_4$$

where, as before, since we know that e_n is universally measurable the probability statement is for the probability of the two measurable sets that trap the event. The proof of Theorem 15 showed that the limit of a decreasing sequence of universally measurable sets was universally measurable and that the probability of the limit is the limit of the sequence of probabilities. Consequently,

$$\mathcal{P}_{Z_\infty}\Big(e_\infty^* \leq e_{\mathcal{B}} + \epsilon_1 + p(0)\Big(\mu_0(N_{2\epsilon_3}(h_{\epsilon_2})) - \mu_0(h_{\epsilon_2})\Big) + p(1)(\epsilon_2)\Big) = 1.$$

Since a metric space is Hausdorff, the compact set h_{ϵ_2} is closed and so the $\mu_0(N_{2\epsilon_3}(h_{\epsilon_2}))$ converges to $\mu_0(h_{\epsilon_2})$ as ϵ_3 goes to zero. Consequently we can choose ϵ_3 small enough so that

$$\mathcal{P}_{Z_\infty}\Big(e_\infty^* \leq e_{\mathcal{B}} + \epsilon_1 + \epsilon_2\Big) = 1.$$

Letting ϵ_2 and ϵ_1 go to zero, we obtain

$$\mathcal{P}_{Z_\infty}\Big(e_\infty^* \leq e_{\mathcal{B}}\Big) = 1$$

and the proof is finished. ■

We have already shown $\acute{\mathcal{C}}$ to be image admissible Suslin and have observed that it is also both z-symmetric and z-increasing. From Theorems 13,15, and Lemma 17 all that is left is to show that the bound on the shatter coefficient $S_{q,n}$ satisfies the conditions of Theorem 13. In Section 4.2 we showed that $S_{n/2n}(\acute{\mathcal{C}}^q) \leq (2n)^{2q}$. Thus, a direct calculation with $S_{q,n} = (2n)^{2q}$ shows that the conditions of Theorem 13 are satisfied. We note that these shattering bounds are dimension independent, whereas if the balls could be located anywhere this would not be the case.

■

6. A framework for classification

The empirical process results in the preceding sections form the basis of a new framework to analyze the performance of statistical learning paradigms. In this framework classifier design is decomposed into two components; the first component is the restriction to the data dependent hypothesis class and the second is empirical risk minimization within that class. Analysis within this framework involves the the study of the decomposition induced on both error and computation. There are several goals we might hope to accomplish with this framework: tighter performance bounds for existing learning paradigms, simpler (or more elegant) proofs, greater flexibility in the analysis, and the discovery of new mechanisms for controlling error and/or computation (which may lead to new learning paradigms). In this section we discuss progress along these fronts using the VC framework as a benchmark for comparison. Specifically, we make comparisons between error bounds for traditional classes C, and their data dependent counterparts $\mathcal{C}$, where the data dependency is introduced in different ways. A natural way of defining a data dependent class is to impose a constraint on a traditional class C to obtain $\mathcal{C}_{z_n}$. If this is not the case we define $C = \cup_{z_n}\mathcal{C}_{z_n}$ as the traditional hypothesis class.

We now examine the structure of the induced error decomposition. It is common to break the generalization error into two components, *approximation error* and *estimation error*, which are defined as follows. The approximation error for C is

$$A = \inf_{f \in C} e(f) - e_{\mathcal{B}}$$

and the estimation error is

$$E = e(\hat{f}) - \inf_{f \in C} e(f)$$

where $\hat{f} \in C$ minimizes the empirical error.

The approximation error for the data dependent class $\mathcal{C}$ is

$$A_D = \inf_{f \in \mathcal{C}_{z_n}} e(f) - e_{\mathcal{B}}$$

and the estimation error is

$$E_D = e(\tilde{f}) - \inf_{f \in \mathcal{C}_{z_n}} e(f),$$

where $\tilde{f} \in \mathcal{C}_{z_n}$ minimizes the empirical error. The data dependent approximation error A_D splits into

$$A_D = A + E_{DD}$$

where the data dependency error is

$$E_{DD} = \inf_{f \in \mathcal{C}_{z_n}} e(f) - \inf_{f \in C} e(f).$$

Consequently we see the increase in approximation error due to data dependency. The data dependent estimation error splits into

$$E_D = E - E_{DD} + e(\tilde{f}) - e(\hat{f})$$

giving

$$A_D + E_D = A + E + e(\tilde{f}) - e(\hat{f}).$$

The right hand side shows the possible benefits of data dependent learning which we analyze through the terms on the left hand side.

Analysis of the data dependent approximation error is similar to that for traditional classifiers, except for the data dependency error term. Our treatment of this random variable has just begun. We have shown that its infinite sample limit is constant when the data dependency is symmetric in its dependence on the n-sample. Although we have provided no general conditions on the data dependency that allow computation of this constant, we have shown that it is zero when we employ the structural risk minimization over multi-sphere classifiers as described in Section 5.2. In this case we have also shown that the infinite sample limit of the data dependent approximation error is zero.

In both the traditional and data dependent theory estimation error is controlled in terms of the shatter coefficient. While the traditional theory uses the Sauer lemma to provide a simple bound for the shatter coefficient in terms of the VC dimension (Devroye et al., 1996) we have discovered no such simplification in the data dependent case. Indeed, we have not yet found a way to bound the shatter coefficient in terms of general characteristics of the data dependent class. In fact we suspect incorporation of the expectation process in the shatter bounds will become important. Such an approach leads to problems in the field of probabilistic geometry (Ambartzumian, 1990). However, there are few distribution independent results in this field, the most notable exception being the result of Rogers (1978) concerning the probability that the convex hulls of two samples do not intersect. Although relevant, we have been unable to extend it to more than two dimensions.

We argue that the change in shatter coefficient due to data dependency may not always be significant. Let C be a finite VC class and let the corresponding data dependent class be the subset of C that achieves the minimum empirical error on the n-sample,

$$\mathcal{C}_{z_n} = \{\hat{f} : \hat{f} = \arg\min_{f \in C} \hat{e}(f)\}$$

Since $\mathcal{C}_{z_n} \subseteq C$ it is clear that $S_{n/2n}(\mathcal{C})$ is less than or equal to the shatter coefficient $S(n, C)$ of C, but the question is whether $S_{n/2n}(\mathcal{C})$ is sufficiently smaller (uniformly) to guarantee a reduction in estimation error. The existence of lower bounds on the generalization error for empirical error minimization (Devroye et al., 1996; Vapnik, 1998) that closely match the upper bounds with which we wish to compare suggest not.

We now discuss examples where the improvement is significant. We begin by considering the sphere classifiers where the multi-sphere class is restricted to a single sphere. The traditional class C, whose center and radius are unrestricted, can be represented as a generalized linear classifier with VC dimension $\leq d+2$ (Cover, 1965) (where d is the dimension of the sample space) and so has shatter coefficient $S(n, C) \leq n^{d+2}$. Forcing the sphere to be centered at a data point represents a clear restriction on the function class. Note that traditional VC theory accounts for the reduction in complexity that results from placing the center at any fixed location. Indeed, in this case $S(n, C) \leq n$, but the center location must be chosen before we see the data. On the other hand, a data dependent framework like the one in this paper is required if we wish to account for the reduction in complexity that results from centering the sphere at a data point. We have shown that $S_{n/2n}(\mathcal{C}) \leq (2n)^2$ which demonstrates that the complexity reduction is manifested in a shattering bound that is independent of dimension. While this class restriction may give tighter control on the estimation error, it is likely to increase the approximation error. However, the dimension independent nature of the estimation error allows us to employ methods for reducing approximation error that were not previously available to us. For example we can map to a higher dimensional space and implement the single sphere classifier there (in much the same way we do with support vector machines). This can be accomplished in a computationally efficient way through the use of a kernel to compute distances (Schölkopf, 2001). Finally we note that forcing the sphere to be centered at a data sample gives rise to a significant reduction in the computational requirements for learning. Computing the optimal position and radius of a single ball is NP-Hard (Johnson & Preparata, 1978), but determining the optimal center sample and its radius can be accomplished in (low order) polynomial time (e.g. the brute force algorithm runs in $O(dn^2 \log n)$ time). In summary, the data dependent framework has allowed us to quantify the reduction in class complexity obtained by forcing the ball to be centered at a data point, provided dimension independent error bounds, and consequently led us to consider a new learning paradigm that is computationally tractable. Similar conclusions can be drawn for the set of simple linear classifiers introduced in Section 4.1.

It is possible that the significant decrease in estimation error in these examples is countered by a large increase in approximation error. However, we can reduce the approximation error substantially by using multiple balls (or linear classifiers). For example we consider variations of the q-sphere classifiers introduced in Section 4.2. Theorem 16 tells us that under very general conditions we can drive the approximation (and estimation) error to zero asymptotically (*wp1*). In practice, where q and n are finite, since the estimation error bounds remain independent of dimension, we can combine the technique of mapping to a higher dimension through a kernel with multiple balls to help reduce approximation error. While this is an attractive idea, computational requirements may limit its utility. Theorem 9 states that the problem of finding the best q out of n balls and their radii is NP-Hard [1]. Consequently it may be beneficial to consider variations of the data dependent class $\mathcal{C}^q_{z_n}$ that allow us to trade approximation and/or estimation error for computational resources.

For example suppose we restrict $\mathcal{C}^q_{z_n}$ by either fixing the radii or fixing the subset of the samples where the balls are centered. These restrictions lead to a natural reduction in the shatter coefficient for the corresponding data dependent class as shown is Section 4.2. Note however, that when we fix the subset where the balls are centered symmetry is violated and the infinite sample limit of the data dependent optimal generalization error may not be constant (see the proof of Theorem 15). While these variations may remain NP-Hard(see Section 4.2) it is possible that they admit a polynomial time approximation that is acceptable for use in practice. They may also be solved by an algorithm whose expected (or typical) run time is polynomial.

Finally, we obtain a computationally tractable variant of this problem by fixing the q samples where the balls are centered and employ a single radius (the same for each ball) that is chosen to minimize the empirical error.

In summary, we have introduced a new framework which, although incomplete, has demonstrated its utility by allowing us to more thoroughly quantify the trade-offs between performance and computational complexity, and has lead to the discovery of new families of classifiers with dimension independent performance bounds and efficient learning procedures.

1. Note that since there are at most $\binom{n}{q} = O(n^q)$ choices for the center points, and for each of these no more than n^q choices for the radii, the problem is polynomial for fixed q. Even so, it is not practical for q greater than about 3.

Acknowledgments

We would like to thank Leonid Gurvits for a very helpful conversation. Clint Scovel gratefully acknowledges support from the Los Alamos DOE Program in Applied Mathematics. Adam Cannon gratefully acknowledges support from Los Alamos National Laboratory.

References

R.V. Ambartzumian. *Factorization Calculus and Geometric Probability*. Cambridge University Press, Cambridge, 1990.

R.B. Ash. *Real Analysis and Probability*. Academic Press, New York, 1972.

P.L. Bartlett, S. Boucheron, and G. Lugosi. Model selection and error estimation. *preprint*, 2000. (A shorter version of this paper was apparently presented at COLT 2000.).

S. Boucheron, G. Lugosi, and P. Massart. A sharp inequality with applications. *Random Structures and Algorithms*, 16:277–292, 2000.

K. Buescher and P. Kumar. Learning by canonical smooth estimation, part ii: Learning and choice of model complexity. *IEEE Transactions on Automatic Control*, 41:557–569, 1996.

A.H. Cannon and L.J. Cowen. Approximation algorithms for the class cover problem. In *Proceedings of the 6th International Symposium on Mathematics and Artificial Intelligence*, 2000.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

T. Cover. Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition. *IEEE Transactions on Electronic Computers*, 14:326–334, 1965.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000.

L. Devroye. Automatic pattern recognition: A study of the probability of error. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 10(4):530–543, 1988.

L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer-Verlag, 1996.

R.M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, New York, 1999.

Y. Freund. Self bounding learning algorithms. In *COLT; Proceedings of the Workshop on Computational Learning Theory*. Morgan Kaufmann, 1998.

Y. Gat. A bound concerning the generalization ability of a certain class of learning algorithms. Technical Report 548, University of California, Berkeley, 1999.

Dorit S. Hochbaum. *Approximation Algorithms for NP-Hard Problems.* PWS Publishing Company, 1997.

D. Hush and C. Scovel. On the vc dimension of bounded margin classifiers. *Machine Learning*, 45:33–44, 2001.

D.S. Johnson and F.P. Preparata. The densest hemisphere problem. *Theoretical Computer Science*, 6:93–107, 1978.

V.I. Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

V.I. Koltchinskii, C.T. Abdallah, M. Ariola, P. Dorato, and D. Panchenko. Improved sample complexity estimates for statistical learning control of uncertain systems. *IEEE Transactions on Automatic Control*, 45(12):2383–2388, 2000.

D.J. Marchette and C.E. Priebe. Characterizing the scale dimension of a high dimensional classification problem. *Pattern Recognition*. forthcoming.

C.E. Priebe, J.G. DeVinney, and D.J. Marchette. On the distribution of the domination number for random class cover catch digraphs. *Statistics and Probability Letters*, 55(3): 239–246, 2001.

L.C.G. Rogers. The probability that two samples in the plane have disjoint convex hulls. *Journal of Applied Probability*, 15:790–802, 1978.

B. Schölkopf. The kernel trick for distances. *Advances in Neural Information Processing Systems*, 13:301–307, 2001. Editors: T.K. Leen, T.G. Dietterich, and V. Tresp.

R.J. Serfling. Probability inequalities for the sum in sampling without replacement. *Annals of Statistics*, 2(1):39–48, 1974.

J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998. Also appeared in *NeuroCOLT Technical Report*, NC-TR-96-053, 1996.

J. Shawe-Taylor and N. Cristianini. On the generalization of soft margin algorithms. Technical Report NC2-TR-2000-082, NeuroCOLT2 Technical Report Series, 2000.

A.N. Shiryaev. *Probability.* Springer-Verlag, New York, 1980.

V.N. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, Inc., New York, 1998.

G. Young and A.S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3:19–22, 1938.