Journal of Machine Learning Research 14 (2013) 3591-3618 Submitted 4/13; Revised 9/13; Published 12/13

Classifier Selection using the Predicate Depth

Ran Gilad-Bachrach RANG@MICROSOFT.COM
Christopher J.C. Burges CBURGES@MICROSOFT.COM
Microsoft Research
1 Microsoft Way
Redmond, WA, 98052
USA

Editor: John Shawe-Taylor

Abstract

Typically, one approaches a supervised machine learning problem by writing down an objective function and finding a hypothesis that minimizes it. This is equivalent to finding the Maximum A Posteriori (MAP) hypothesis for a Boltzmann distribution. However, MAP is not a robust statistic. We present an alternative approach by defining a median of the distribution, which we show is both more robust, and has good generalization guarantees. We present algorithms to approximate this median.

One contribution of this work is an efficient method for approximating the Tukey median. The Tukey median, which is often used for data visualization and outlier detection, is a special case of the family of medians we define: however, computing it exactly is exponentially slow in the dimension. Our algorithm approximates such medians in polynomial time while making weaker assumptions than those required by previous work.

Keywords: classification, estimation, median, Tukey depth

1. Introduction

According to the PAC-Bayesian point of view, learning can be split into three phases. First, a prior belief is introduced. Then, observations are used to transform the prior belief into a posterior belief. Finally, a hypothesis is selected. In this study, we concentrate on the last step. This allows us to propose methods that are independent of the first two phases. For example, the observations used to form the posterior belief can be supervised, unsupervised, semi-supervised, or something entirely different. The most commonly used method for selecting a hypothesis is to select the maximum a posteriori (MAP) hypothesis. For example, many learning algorithms use the following evaluation function (energy function):

$$E(f) = \sum_{i=1}^{n} l(f(x_i), y_i) + r(f) \ , \tag{1}$$

where l is a convex loss function, $\{(x_i, y_i)\}_{i=1}^{n}$ are the observations and r is a convex regularization term. This can be viewed as a prior P over the hypothesis class with density $p(f) = \frac{1}{Z_p} e^{-r(f)}$ and a posterior belief Q with density $q(f) = \frac{1}{Z_q} e^{-E[f]}$. The common practice is then to select the hypothesis that minimizes the evaluation function, that is, the MAP hypothesis. However, this choice has two significant drawbacks. First, since it considers only the maximal point, it misses much of the information encoded in the posterior belief. As a result it is straightforward to construct patho-

logical examples: in Section 2.4 we give an example where the MAP classifier solution disagrees with the Bayes optimal hypothesis on every point, and where the Bayes optimal hypothesis[1] in fact *minimizes* the posterior probability. Second, the MAP framework is sensitive to perturbations in the posterior belief. That is, if we think of the MAP hypothesis as a statistic of the posterior, it has a low breakdown point (Hampel, 1971): in fact, its breakdown point is zero as demonstrated in Section 2.2.

This motivates us to study the problem of selecting the best hypothesis, given the posterior belief. The goal is to select a hypothesis that will generalize well. Two well known methods for achieving this are the Bayes optimal hypothesis, and the Gibbs hypothesis, which selects a random classifier according to the posterior belief. However the Gibbs hypothesis is non-deterministic, and in most cases the Bayes optimal hypothesis is not a member of the hypothesis class; these drawbacks are often shared by other hypothesis selection methods. This restricts the usability of these approaches. For example, in some cases, due to practical constraints, only a hypothesis from a given class can be used; ensembles can be slow and can require large memory footprint. Furthermore stochasticity in the predictive model can make the results non-reproducible, which is unacceptable in many applications, and even when acceptable, makes the application harder to debug. Therefore, in this work we limit the discussion to the following question: given a hypothesis class $\mathcal{F}$ distributed according to a posterior belief Q, how can one select a hypothesis $f \in \mathcal{F}$ that will generalize well? We further limit the discussion to the binary classification setting.

To answer this question we extend the notions of depth and the multivariate median, that are commonly used in multivariate statistics (Liu et al., 1999), to the classification setting. The depth function measures the centrality of a point in a sample or a distribution. For example, if Q is a probability measure over $\mathbb{R}^d$, the Tukey depth for a point $x \in \mathbb{R}^d$, also known as the half-space depth (Tukey, 1975), is defined as

$$\text{TukeyDepth}_Q(x|Q) = \inf_{H \text{ s.t. } x \in H \text{ and } H \text{ is a halfspace}} Q(H) \ . \tag{2}$$

That is, the depth of a point x is the minimal measure of a half-space that contains it.[2] The Tukey depth also has a minimum entropy interpretation: each hyperplane containing x defines a Bernoulli distribution by splitting the distribution Q in two. Choose that hyperplane whose corresponding Bernoulli distribution has minimum entropy. The Tukey depth is then the probability mass of the halfspace on the side of the hyperplane with the lowest such mass.

The depth function is thus a measure of centrality. The median is then simply defined as the deepest point. It is easy to verify that in the univariate case, the Tukey median is indeed the standard median. In this work we extend Tukey's definition beyond half spaces and define a depth for any hypothesis class which we call the predicate depth. We show that the generalization error of a hypothesis is inversely proportional to its predicate depth. Hence, the median predicate hypothesis, or *predicate median,* has the best generalization guarantee. We present algorithms for approximating the predicate depth and the predicate median. Since the Tukey depth is a special case of the predicate depth, our algorithms provide polynomial approximations to the Tukey depth and the Tukey median as well. We analyze the stability of the predicate median and also discuss the case where a convex evaluation function $E(f)$ (see Equation (1)) is used to form the posterior belief. We show that in

1. The Bayes optimal hypothesis is also known as the Bayes optimal classifier. It performs a weighted majority vote on each prediction according to the posterior.
2. Note that we can restrict the half spaces in (2) to those half spaces for which x lies on the boundary.

Symbol	Description
$\mathcal{X}$	a sample space
x	an instance $x \in \mathcal{X}$
μ	a probability measure over $\mathcal{X}$
S	a sample of instances, $S = \{x_1, \ldots, x_u\}$.
$\mathcal{F}$	a function class. $f \in \mathcal{F}$ is a function $f : \mathcal{X} \mapsto \pm 1$.
f, g	functions in the function class $\mathcal{F}$
P, Q, Q'	probability measures over $\mathcal{F}$
T	a sample of functions, $T = \{f_1, \ldots, f_n\}$.
$D_Q(f \mid x)$	The depth of the function f on the instance x with respect to the measure Q.
$D_Q(f)$	The depth of the function f with respect to the measure Q.
$D_Q^{\delta,\mu}(f)$	The δ-insensitive depth of f with respect to Q and μ.
$\hat{D}_T(f \mid x)$	The *empirical depth* of f on the instance x with respect to the sample T
$\hat{D}_T^S(f)$	The *empirical depth* of f with respect to the samples T and S.
ν	A probability measure over $\mathcal{X} \times \{\pm 1\}$
$\mathcal{S}$	a sample $\{(x_i, y_i)\}_{i=1}^m$ from $(\mathcal{X} \times \{\pm 1\})^m$
$R_\nu(f)$	The generalization error of f: $R_\nu(f) = \Pr_{(x,y)\sim\nu}[f(x) \neq y]$.
$R_\mathcal{S}(f)$	The empirical error of f: $R_\mathcal{S}(f) = \Pr_{(x,y)\sim\mathcal{S}}[f(x) \neq y]$.

Table 1: A summary of the notation used in this work

this special case, the average hypothesis has a depth of at least $1/e$, independent of the dimension. Hence, it enjoys good generalization bounds.

In the first part of this work we introduce the notion of predicate depth. We discuss its properties and contrast them with the properties of the MAP estimator. In the second part we discuss algorithmic aspects. We address both the issues of approximating depth and of approximating the deepest hypothesis, that is, the predicate median. Table 1 contains a summery of the notation we use.

2. The Predicate Depth: Definitions and Properties

In this study, unlike Tukey who used the depth function on the instance space, we view the depth function as operating on the dual space, that is the space of classification functions. Moreover, the definition here extends beyond the linear case to any function class. The depth function measures the agreement of the function f with the weighted majority vote on x. A deep function is a function that will always have a large agreement with its prediction among the class $\mathcal{F}$.

Definition 1 *Let $\mathcal{F}$ be a function class and let Q be a probability measure over $\mathcal{F}$. The predicate depth of f on the instance $x \in \mathcal{X}$ with respect to Q is defined as*

$$D_Q(f \mid x) = \Pr_{g\sim Q}[g(x) = f(x)] \ .$$

The predicate depth of f with respect to Q is defined as

$$D_Q(f) = \inf_{x\in\mathcal{X}} D_Q(f \mid x) = \inf_{x\in\mathcal{X}} \Pr_{g\sim Q}[g(x) = f(x)] \ .$$

The Tukey-Depth is a special case of this definition as discussed in section 2.1. We can now define the predicate median:

Definition 2 *Let $\mathcal{F}$ be a function class and let Q be a probability measure over $\mathcal{F}$. f^* is a predicate median of $\mathcal{F}$ with respect to Q if*

$$\forall f \in \mathcal{F},\ D_Q(f) \leq D_Q(f^*)\ .$$

We show later, in Theorem 13, that if $\mathcal{F}$ is closed then the median always exists, for every probability measure Q. The depth $D_Q(f)$ is defined as the infimum over all points $x \in \mathcal{X}$. However, for our applications, we can tolerate some instances $x \in \mathcal{X}$ which have small depth, as long as most of the instances have large depth. Therefore, we define the δ-insensitive depth:

Definition 3 *Let $\mathcal{F}$ be a function class and let Q be a probability measure over $\mathcal{F}$. Let μ be a probability measure over $\mathcal{X}$ and let $\delta \geq 0$. The δ-insensitive depth of f with respect to Q and μ is defined as*

$$D_Q^{\delta,\mu}(f) = \sup_{X' \subseteq \mathcal{X}, \mu(X') \leq \delta} \inf_{x \in \mathcal{X} \setminus X'} D_Q(f\,|\,x)\ .$$

The δ-insensitive depth function relaxes the infimum in the depth definition. Instead of requiring that the function f always have a large agreement in the class $\mathcal{F}$, the δ-insensitive depth makes this requirement on all but a set of the instances with probability mass δ.

With these definitions in hand, we next provide generalization bounds for deep hypotheses. The first theorem shows that the error of a deep function is close to the error of the Gibbs classifier.

Theorem 4 ***Deep vs. Gibbs***

Let Q be a measure on $\mathcal{F}$. Let ν be a measure on $\mathcal{X} \times \{\pm 1\}$ with the marginal μ on $\mathcal{X}$. For every f the following holds:

$$R_\nu(f) \leq \frac{1}{D_Q(f)} E_{g \sim Q}[R_\nu(g)] \tag{3}$$

and

$$R_\nu(f) \leq \frac{1}{D_Q^{\delta,\mu}(f)} E_{g \sim Q}[R_\nu(g)] + \delta\ . \tag{4}$$

Note that the term $E_{g \sim Q}[R_\nu(g)]$ is the expected error of the Gibbs classifier (which is not necessarily the same as the expected error of the Bayes optimal hypothesis). Hence, this theorem proves that the generalization error of a deep hypothesis cannot be large, provided that the expected error of the Gibbs classifier is not large.

Proof For every $x^* \in \mathcal{X}$ we have that

$$\begin{aligned}
& \Pr_{g \sim Q, (x,y) \sim \nu}[g(x) \neq y\,|\,x = x^*] \\
\geq\ & \Pr_{g \sim Q, (x,y) \sim \nu}[f(x) \neq y \text{ and } g(x) = f(x)\,|\,x = x^*] \\
=\ & \Pr_{(x,y) \sim \nu}[f(x) \neq y | x = x^*] \Pr_{g \sim Q, (x,y) \sim \nu}[g(x) = f(x)\,|\,x = x^* \text{ and } f(x) \neq y] \\
=\ & \Pr_{(x,y) \sim \nu}[f(x) \neq y | x = x^*] \Pr_{g \sim Q, (x,y) \sim \nu}[g(x) = f(x)\,|\,x = x^*] \\
=\ & \Pr_{(x,y) \sim \nu}[f(x) \neq y | x = x^*] D_Q(f\,|\,x^*) \geq \Pr_{(x,y) \sim \nu}[f(x) \neq y | x = x^*] D_Q(f)\ .
\end{aligned}$$

First we prove (4). Define the set $Z = \left\{x : D_Q(f|x) < D_Q^{\delta,\mu}(f)\right\}$. Clearly $\mu(Z) \leq \delta$. By slight abuse of notation, we define the function $Z(x)$ such that $Z(x) = 1$ if $x \in Z$ and $Z(x) = 0$ if $x \notin Z$. With this definition we have

$$\begin{aligned}
\frac{1}{D_Q^{\delta,\mu}(f)} E_{g\sim Q}[R_\nu(g)] + \delta &\geq E_{x^*\sim\mu}\left[\frac{1}{D_Q^{\delta,\mu}(f)} \Pr_{g\sim Q,(x,y)\sim\nu}[g(x) \neq y | x = x^*] + Z(x^*)\right] \\
&\geq E_{x^*\sim\mu}\left[\frac{1}{D_Q^{\delta,\mu}(f)} \Pr_{(x,y)\sim\nu}[f(x) \neq y | x = x^*] D_Q(f|x^*) + Z(x^*)\right] \\
&\geq E_{x^*\sim\mu}\left[\Pr_{(x,y)\sim\nu}[f(x) \neq y | x = x^*]\right] = R_\nu(f) \ .
\end{aligned}$$

(3) follows in the same way by setting both Z and δ to be zero. ■

Theorem Deep vs. Gibbs (Theorem 4) bounds the ratio of the generalization error of the Gibbs classifier and the generalization error of a given classifier as a function of the depth of the given classifier. For example, consider the Bayes optimal classifier. By definition, the depth of this classifier is at least one half; thus Theorem 4 recovers the well-known result that the generalization error of the Bayes optimal classifier is at most twice as large as the generalization error of the Gibbs classifier.

Next, we combine Theorem Deep vs. Gibbs (Theorem 4) with PAC-Bayesian bounds (McAllester, 1999) to bound the difference between the training error and the generalization error. We use the version of the PAC-Bayesian bounds in Theorem 3.1. of Germain et al. (2009).

Theorem 5 ***Generalization Bounds***

Let ν be a probability measure on $\mathcal{X} \times \{\pm 1\}$, let P be a fixed probability measure on $\mathcal{F}$ chosen a priori, and let $\delta, \kappa > 0$. For a proportion $1-\delta$ of samples $S \sim \nu^m$,

$$\forall Q,\ \forall f,\ R_\nu(f) \leq \frac{1}{(1-e^{-\kappa}) D_Q(f)} \left(\kappa E_{g\sim Q}[R_S(g)] + \frac{1}{m}\left[KL(Q||P) + \ln\frac{1}{\delta}\right]\right) \ .$$

Furthermore, for every $\delta' > 0$, for a proportion $1-\delta$ of samples $S \sim \nu^m$,

$$\forall Q,\ \forall f,\ R_\nu(f) \leq \frac{1}{(1-e^{-\kappa}) D_Q^{\delta',\mu}(f)} \left(\kappa E_{g\sim Q}[R_S(g)] + \frac{1}{m}\left[KL(Q||P) + \ln\frac{1}{\delta}\right]\right) + \delta' \ ,$$

where μ is the marginal of ν on $\mathcal{X}$.

Proof Applying the bounds in Theorem 4 to the PAC-Bayesian bounds in Theorem 3.1 of Germain et al. (2009) yields the stated results. ■

The generalization bounds theorem (Theorem 5) shows that if a deep function exists, then it is expected to generalize well, provided that the PAC-Bayes bound for Q is sufficiently smaller than the depth of f. This justifies our pursuit to find the deepest function, that is, the median. However, the question remains: are there any deep functions? In the following section we show that this indeed the case for linear classifiers.

2.1 Depth for Linear Classifiers

In this section we discuss the special case where the hypothesis class consists of linear classifiers. Our goal is to show that deep functions exist and that the Tukey depth is a special case of the predicate depth. To that end we use a variant of linear classifiers called *linear threshold functions*. We denote by $\mathbb{S} = \left\{x \in \mathbb{R}^d \,:\, \|x\| = 1\right\}$ the unit sphere. In this setting $\mathcal{F} = \mathbb{R}^d$ and $\mathcal{X} = \mathbb{S} \times \mathbb{R}$ such that $f \in \mathcal{F}$ operates on $x = (x_v, x_\theta) \in \mathcal{X}$ by $f(x) = \text{sign}(f \cdot x_v - x_\theta)$.[3] One can think of an instance $x \in \mathcal{X}$ as a combination of a direction, denoted by x_v, and an offset x_θ.

Theorem 6 *Let $\mathcal{X} = \mathbb{S} \times \mathbb{R}$ and $\mathcal{F}$ be the class of linear threshold functions over $\mathcal{X}$. Let Q be a probability measure over $\mathcal{F}$ with density function $q(f)$ such that $q(f) = \frac{1}{Z}\exp(-E(f))$ where $E(f)$ is a convex function. Let $f^* = E_{f\sim Q}[f]$. Then $D_Q(f^*) \geq 1/e$.*

Proof From the definition of $q(f)$ it follows that it is log-concave. Borell (1975) proved that Q is log-concave if and only if q is log-concave. Hence, in the setting of the theorem, Q is log concave. The Mean Voter Theorem of Caplin and Nalebuff (1991) shows that if f is the center of gravity of Q then for every x, $D_Q(f\,|\,x) \geq 1/e$ and thus the center of gravity of Q has a depth of at least $1/e$ (e is Euler's number). Note that since $\mathcal{F} = \mathbb{R}^d$ then the center of gravity is in $\mathcal{F}$. ■

Recall that it is common practice in machine learning to use convex energy functions $E(f)$. For example, SVMs (Cortes and Vapnik, 1995) and many other algorithms use energy functions of the form presented in (1) in which both the loss function and the regularization functions are convex, resulting in a convex energy function. Hence, in all these cases, the median, that is the deepest point, has a depth of at least $1/e$.[4] This leads to the following conclusion:

Conclusion 1 *In the setting of Theorem 6, let $f^* = E_{f\sim Q}[f]$. Let ν be a probability measure on $\mathcal{X} \times \{\pm 1\}$, let P be a probability measure of $\mathcal{F}$ and let $\delta, \kappa > 0$. With a probability greater than $1-\delta$ over the sample $\mathcal{S}$ sampled from ν^m:*

$$R_\nu(f^*) \leq \frac{e}{(1-e^{-\kappa})}\left(\kappa E_{g\sim Q}[R_{\mathcal{S}}(g)] + \frac{1}{m}\left[KL(Q||P) + \ln\frac{1}{\delta}\right]\right) .$$

Proof This results follows from Theorem 5 and Theorem 6. ■

We now turn to show that the Tukey depth is a special case of the predicate depth. Again, we will use the class of linear threshold functions. Since $\mathcal{F} = \mathbb{R}^d$ in this case we will treat $f \in \mathcal{F}$ both as a function and as a point. Therefore, a probability measure Q over $\mathcal{F}$ is also considered as a probability measure over $\mathbb{R}^d$. The following shows that for any $f \in \mathcal{F}$, the Tukey-depth of f and the predicate depth of f are the same.

Theorem 7 *If $\mathcal{F}$ is the class of threshold functions then for every $f \in \mathcal{F}$:*

$$D_Q(f) = \mathit{TukeyDepth}_Q(f) .$$

3. We are overloading notation here: f is treated as both a point in $\mathbb{R}^d$and a function $f(x) : \mathbb{S} \times \mathbb{R} \to \pm 1$.
4. Note that optimization in general finds the MAP, which can be very different from (and less robust than) the median (see Section 2.4).

Proof Every closed half-space H in $\mathbb{R}^d$ is uniquely identified by a vector $z_H \in \mathbb{S}$ orthogonal to its hyperplane boundary and an offset θ_H such that

$$H = \{g : g \cdot z_H \geq \theta_H\} \ .$$

In other words, there is a $1-1$ correspondence between half-spaces and points in $\mathbb{S} \times \mathbb{R}$ such that $H \mapsto (z_H, \theta_H)$ and such that

$$g \in H \ \Leftrightarrow \ g((z_H, \theta_H)) \equiv sign(g \cdot z_H - \theta_H) = 1 \ .$$

The Tukey depth of f is the infimum measure of half-spaces that contain f:

$$\begin{aligned} \text{TukeyDepth}_Q(f) = \inf_{H: f \in H} Q(H) &= \inf_{x: f(x)=1} Q\{g : g(x) = 1\} \\ &= \inf_{x: f(x)=1} D_Q(f \,|\, x) \\ &\geq D_Q(f) \ . \end{aligned}$$

Hence, the Tukey depth cannot be larger than the predicate depth.

Next we show that the Tukey depth cannot be smaller than the predicate depth for if the Tukey depth is smaller than the predicate depth then there exists a half space H such that its measure is smaller than the predicate depth. Let $x = (z_H, \theta_H)$. Since $f \in H$ then $f(x) = 1$ and thus

$$D_Q(f) > Q(H) = Q\{g : g(x) = 1\} = D_Q(f \,|\, x) \geq D_Q(f)$$

which is a contradiction. Therefore, the Tukey depth can be neither smaller nor larger than the predicate depth and so the two must be equal. ■

2.2 Breakdown Point

We now turn to discuss another important property of the hypothesis selection mechanism: the breakdown point. Any solution to the hypothesis selection problem may be viewed as a statistic of the posterior Q. An important property of any such statistic is its stability: that is, informally, by how much must Q change in order to produce an arbitrary value of the statistic? This is usually referred to as the breakdown point (Hampel, 1971). We extend the definition given there as follows:

Definition 8 *Let **Est** be a function that maps probability measures to $\mathcal{F}$. For two probability measures Q and Q' let $\delta(Q, Q')$ be the total variation distance:*

$$\delta\left(Q, Q'\right) = \sup\left\{\left|Q(A) - Q'(A)\right| : A \text{ is measurable}\right\} \ .$$

For every function $f \in \mathcal{F}$ let $d(\textbf{Est}, Q, f)$ be the distance to the closest Q' such that $\textbf{Est}(Q') = f$:

$$d(\textbf{Est}, Q, f) = \inf\left\{\delta\left(Q, Q'\right) : \textbf{Est}\left(Q'\right) = f\right\} \ .$$

*The breakdown **Est** at Q is defined to be*

$$\textbf{breakdown}(\textbf{Est}, Q) = \sup_{f \in \mathcal{F}} d(\textbf{Est}, Q, f) \ .$$

This definition may be interpreted as follows; if $s = \textbf{breakdown}(\textbf{Est}, Q)$ then for every $f \in \mathcal{F}$, we can force the estimator **Est** to use f as its estimate by changing Q by at most s in terms of total variation distance. Therefore, the larger the breakdown point of an estimator, the more stable it is with respect to perturbations in Q.

As mentioned before, our definition of the breakdown point for an estimator stems from the work of Hampel (1971) who was the first to introduce the concept. Since then, different modifications have been suggested to address different scenarios. Davies and Gather (2005) discuss many of these definitions. He (2005) noted that one can make the breakdown point trivial, for instance, if **Est** is a fixed estimator that is not affected by its input, it has the best possible breakdown point of 1. Moreover, it suffices to have a single function f that cannot be produced as the output of **Est** to make the above definition trivial. To prevent these pathologies, Definition 8 should only be used when **Est** is such that for every f there exists Q' for which $\textbf{Est}(Q') = f$ which is the case for the estimators we study here.

The following theorem lower bounds the stability of the median estimator as a function of its depth.

Theorem 9 *Let Q be a posterior over $\mathcal{F}$. Let*

$$\begin{aligned} X' &= \{x \in \mathcal{X} \text{ s.t. } \forall f_1, f_2 \in \mathcal{F},\ f_1(x) = f_2(x)\} \text{ and} \\ p &= \inf_{x \in \mathcal{X} \setminus X', y \in \{\pm 1\}} Q\{f : f(x) = y\} \ . \end{aligned}$$

If d is the depth of the median for Q then $\textbf{breakdown}(\textbf{median}, Q) \geq \frac{d-p}{2}$.

Proof Let $\varepsilon > 0$. There exists $\hat{f}$ and $\hat{x}$ such that $Q\left\{f : f(\hat{x}) = \hat{f}(\hat{x})\right\} < p + \varepsilon$. Let f^* be the median of Q. Let Q' be such that $\hat{f}$ is the median of Q', so that

$$D_{Q'}(f^*) \leq D_{Q'}\left(\hat{f}\right) \ .$$

Note that for every f we have that

$$|D_Q(f) - D_{Q'}(f)| \leq \delta\left(Q, Q'\right) \ .$$

This follows since

$$\begin{aligned} |D_Q(f) - D_{Q'}(f)| &= \left|\inf_x \left(Q\left\{f' : f'(x) = f(x)\right\}\right) - \inf_x \left(Q'\left\{f' : f'(x) = f(x)\right\}\right)\right| \\ &\leq \delta\left(Q, Q'\right) \ . \end{aligned}$$

Since $D_Q\left(\hat{f}\right) < p + \varepsilon$ then

$$d - \delta\left(Q, Q'\right) \leq D_{Q'}(f^*) \leq D_{Q'}\left(\hat{f}\right) < p + \varepsilon + \delta\left(Q, Q'\right) \ .$$

Hence

$$\delta\left(Q, Q'\right) > \frac{d - p - \varepsilon}{2}$$

and thus

$$\textbf{breakdown}(\textbf{median}, Q) > \frac{d - p - \varepsilon}{2} \ .$$

Since this is true for every $\varepsilon > 0$ it follows that

$$\textbf{breakdown}(\textbf{median}, Q) \geq \frac{d-p}{2} \ .$$

■

2.3 Geometric Properties of the Depth Function

In this section we study the geometry of the depth function. We show that the level sets of the depth functions are convex. We also show that if the function class $\mathcal{F}$ is closed then the median exists. First, we define the term *convex hull* in the context of function classes:

Definition 10 *Let $\mathcal{F}$ be a function class and let $g, f_1, \ldots, f_n \in \mathcal{F}$. We say that g is in the convex-hull of $f_1, \ldots, f_n$ if for every x there exists $j \in 1, \ldots, n$ such that $g(x) = f_j(x)$.*

Theorem 11 ***Convexity***
Let $\mathcal{F}$ be a function class with a probability measure Q. If g is in the convex-hull of $f_1, \ldots, f_n$ then

$$D_Q(g) \geq \min_j D_Q(f_j) \ .$$

Furthermore, if μ is a measure on $\mathcal{X}$ and $\delta \geq 0$ then

$$D_Q^{\delta,\mu}(g) \geq \min_j D_Q^{\frac{\delta}{n},\mu}(f_j) \ .$$

Proof Let g be a function. If g is in the convex-hull of $f_1, \ldots, f_n$ then for every x there exists j such that $g(x) = f_j(x)$ and hence

$$D_Q(g\,|\,x) = D_Q(f_j\,|\,x) \geq \min_j D_Q(f_j) \ ,$$

thus $D_Q(g) \geq \min_j D_Q(f_j)$. For $\delta > 0$ let

$$\Delta \quad = \quad \left\{x : D_Q(g\,|\,x) \leq D_Q^{\delta,\mu}(g)\right\}$$

and for $j = 1, \ldots, n$

$$\Delta_j = \left\{x \in \Delta : f_j(x) = g(x)\right\} \ .$$

Since g is in the convex-hull of $f_1, \ldots, f_n$ implies that $\cup_j \Delta_j = \Delta$ and therefore

$$\sum_j \mu(\Delta_j) \geq \mu(\Delta) \geq \delta \ .$$

Hence, there exists j such that $\mu(\Delta_j) \geq \delta/n$ which implies that

$$D_Q^{\delta,\mu}(g) \geq D_Q^{\frac{\delta}{n},\mu}(f_j) \geq \min_j D_Q^{\frac{\delta}{n},\mu}(f_j) \ .$$

■

Next we prove the existence of the median when the function class is closed. We begin with the following definition:

Definition 12 *A function class $\mathcal{F}$ is* closed *if for every sequence $f_1, f_2, \ldots \in \mathcal{F}$ there exists $f^* \in \mathcal{F}$ such that for every $x \in \mathcal{X}$, if $\lim_{n\to\infty} f_n(x)$ exists then $f^*(x) = \lim_{n\to\infty} f_n(x)$.*

With this definition in hand we prove the following:

Theorem 13 *If $\mathcal{F}$ is closed then the median of $\mathcal{F}$ exists with respect to any probability measure Q.*

Proof Let $D = \sup_f D_Q(f)$ and let f_n be such that $D_Q(f_n) > D - 1/n$. Let $f^* \in \mathcal{F}$ be the limit of the series $f_1, f_2, \ldots$. We claim that $D_Q(f^*) = D$. Since D is the supermum of the depth values, it is clear that $D_Q(f^*) \leq D$. Note that from the construction of f^* we have that for every $x \in \mathcal{X}$ and every N there exists $n > N$ such that $f^*(x) = f_n(x)$. Therefore, if $D_Q(f^*) < D$ then there exists x such that $D_Q(f^*|x) < D$. Hence, there is a subsequence $n_k \to \infty$ such that $f_{n_k}(x) = f^*(x)$ and thus

$$D_Q(f_{n_k}) \leq D_Q(f_{n_k}|x) = D_Q(f^*|x) < D \ .$$

But this is a contradiction since $\lim_{k\to\infty} D_Q(f_{n_k}) = D$. Hence, for every x, $D_Q(f^*|x) \geq D$ and thus $D_Q(f^*) \geq D$ which completes the proof. ■

2.4 The Maximum A Posteriori Estimator

So far, we have introduced the predicate depth and median and we have analyzed their properties. However, the common solution to the hypothesis selection problem is to choose the maximum a posteriori estimator. In this section we point out some limitations of this approach. We will show that in some cases, the MAP method has poor generalization. We also show that it is very sensitive in the sense that the breakdown point of the MAP estimator is always zero.

2.4.1 LEARNING AND REGULARIZATION

The most commonly used method for selecting a hypothesis is to select the maximum a posteriori (MAP) hypothesis. For example, in Support Vector Machines (Cortes and Vapnik, 1995), one can view the objective function of SVM as proportional to the log-likelihood function of an exponential probability. From this perspective, the regularization term is proportional to the log-likelihood of the prior. SVM, Lasso (Tibshirani, 1994) and other algorithms use the following evaluation function (energy function):

$$E(f) = \sum_{i=1}^{n} l(f(x_i), y_i) + r(f) \ ,$$

where l is a convex loss function, $\{(x_i, y_i)\}_{i=1}^n$ are the observations and r is a convex regularization term. This can be viewed as if there is a prior P over the hypothesis class with a density function

$$p(f) = \frac{1}{Z_p} e^{-r(f)} \ ,$$

and a posterior belief Q with a density function

$$q(f) = \frac{1}{Z_q} e^{-E[f]} \ .$$

The common practice in these cases is to select the hypothesis that minimizes the evaluation function. Hence these methods select the MAP hypothesis.

2.4.2 THE MAP ESTIMATOR CAN GENERALIZE POORLY

Since the MAP estimator looks only at the peak of the distribution it can be very misleading. Here we give an example for which the MAP estimator disagrees with the Bayes optimal hypothesis on every instance while the median hypothesis agrees with the Bayes optimal hypothesis everywhere. Moreover, the Bayes optimal hypothesis happens to be a member of the hypothesis class. Therefore, it is also the predicate median. Hence, in this case, the MAP estimator fails to represent the belief. The rest of this sub-section is devoted to explaining the details of this example.

Assume that the sample space $\mathcal{X}$ is a set of N discrete elements indexed by integers $1,...,N$. To simplify the exposition we will collapse notation and take $\mathcal{X} = \{1,...,N\}$. The function class $\mathcal{F}$ consists of $N+1$ functions defined as follows: for every $i \in \{1,\ldots,N-1\}$ the function f_i is defined to be

$$f_i(x) = \begin{cases} 1 & \text{if } x \equiv i \text{ or } x \equiv i+1 \\ 0 & \text{otherwise} \end{cases} .$$

Additionally, $\mathcal{F}$ contains the constant functions f_- and f_+ that assign the values 0 and 1, respectively, to every input. Furthermore, assume that there is ε random label noise in the system for some $0 < \varepsilon < 1/2$, and that no further information is available. Thus, the prior is the non-informative uniform prior over the $N+1$ functions.

Assume that a training set consisting of just two examples is available, where the examples are $(x_1 = 1, y_1 = 1)$ and $(x_2 = 3, y_2 = 1)$. Given the ε random label noise, the posterior is easily computed as

$$Q\{f_+\} = \frac{(1-\varepsilon)^2}{Z},\ Q(f_-) = \frac{\varepsilon^2}{Z},\ Q\{f_{i=1,2,3}\} = \frac{\varepsilon(1-\varepsilon)}{Z},\ Q\{f_{i>3}\} = \frac{\varepsilon^2}{Z}$$

where Z is the partition function. Note that under this posterior, for every x,

$$\Pr_{f\sim Q}[f(x) = 1] \leq \frac{(1-\varepsilon)^2 + 2\varepsilon(1-\varepsilon)}{Z} = \frac{1-\varepsilon^2}{Z}$$

while

$$\Pr_{f\sim Q}[f(x) = 0] \geq \frac{(N-2)\varepsilon^2}{Z} .$$

Therefore, if $N > 1 + 1/\varepsilon^2$, for any x, the probability that it is assigned class 0 is larger than the probability that it is assigned class 1. Therefore the Bayes classifier is the function f_-. Since the Bayes classifier is in the function class, it is also the predicate median. However, the MAP estimator is the function f_+. Thus in this case the MAP estimator disagrees with the Bayes optimal hypothesis (and the predicate median) on the entire sample space. Note also that the Bayes optimal hypothesis f_0 has the lowest density in the distribution Q. Hence, in this case, the minimizer of the posterior is a better estimator than the maximizer of the posterior.

2.4.3 THE BREAKDOWN POINT OF THE MAP ESTIMATOR

In Definition 8 we defined the breakdown point of an estimator. We showed in Theorem 9 that the breakdown point of the median hypothesis is lower bounded by a function of its depth. We would like to contrast this with the breakdown point of the MAP estimator. We claim that the breakdown point of the MAP estimator is zero, for continuous concept classes.

Algorithm 1 Depth Estimation Algorithm

Inputs:

- A sample $S = \{x_1, \ldots, x_u\}$ such that $x_i \in \mathcal{X}$
- A sample $T = \{f_1, \ldots, f_n\}$ such that $f_j \in \mathcal{F}$
- A function f

Output:

- $\hat{D}_T^S(f)$ - an approximation for the depth of f

Algorithm:

1. for $i = 1, \ldots, u$ compute $\hat{D}_T(f \,|\, x_i) = \frac{1}{n} \sum_j 1_{f_j(x_i) = f(x_i)}$
2. return $\hat{D}_T^S(f) = \min_i \hat{D}(f \,|\, x_i)$

In order for the MAP estimator to be well defined, assume that Q is a Lebesgue measure such that q is the density function of Q and q is bounded by some finite M. Let $f_0 \in \mathcal{F}$ and consider $Q'$with the density function:

$$q'(f) = \begin{cases} M+1 & \text{if } f = f_0 \\ q(f) & \text{otherwise} \end{cases}.$$

While the total variation distance between Q and Q' is zero, the MAP estimator for Q' is f_0. Therefore, for every f_0 we can introduce a zero measure change to Q that will make f_0 the MAP estimator. Hence, the breakdown point of the MAP estimator is zero.

3. Measuring Depth

So far, we have motivated the use of depth as a criterion for selecting a hypothesis. Finding the deepest function, even in the case of linear functions, can be hard but some approximation techniques have been presented (see Section 4.5). In this section we focus on algorithms that measure the depth of functions. The main results are an efficient algorithm for approximating the depth uniformly over the entire function class and an algorithm for approximating the median.

We suggest a straightforward method to measure the depth of a function. The depth estimation algorithm (Algorithm 1) takes as inputs two samples. One sample, $S = \{x_1, \ldots, x_u\}$, is a sample of points from the domain $\mathcal{X}$. The other sample, $T = \{f_1, \ldots, f_n\}$, is a sample of functions from $\mathcal{F}$. Given a function f for which we would like to compute the depth, the algorithm first estimates its depth on the points $x_1, \ldots, x_u$ and then uses the minimal value as an estimate of the global depth. The depth on a point x_i is estimated by counting the fraction of the functions $f_1, \ldots, f_n$ that make the same prediction as f on the point x_i. Since samples are used to estimate depth, we call the value returned by this algorithm, $\hat{D}_T^S(f)$, the empirical depth of f.

Despite its simplicity, the depth estimation algorithm can provide good estimates of the true depth. The following theorem shows that if the x_i's are sampled from the underlying distribution over $\mathcal{X}$, and the f_j's are sampled from Q, then the empirical depth is a good estimator of the true

depth. Moreover, this estimate is uniformly good over all the functions $f \in \mathcal{F}$. This will be an essential building block when we seek to find the median in Section 3.1.

Theorem 14 ***Uniform convergence of depth***

Let Q be a probability measure on $\mathcal{F}$ and let μ be a probability measure on $\mathcal{X}$. Let $\varepsilon, \delta > 0$. For every $f \in \mathcal{F}$ let the function f_δ be such that $f_\delta(x) = 1$ if $D_Q(f|x) \leq D_Q^{\delta,\mu}(f)$ and $f_\delta(x) = -1$ otherwise. Let $\mathcal{F}_\delta = \{f_\delta\}_{f\in\mathcal{F}}$ .Assume $\mathcal{F}_\delta$ has a finite VC dimension $d < \infty$ and define $\phi(d,k) = \sum_{i=0}^{d} \binom{k}{i}$ if $d < k$, $\phi(d,k) = 2^k$ otherwise. If S and T are chosen at random from μ^u and Q^n respectively such that $u \geq 8/\delta$ then with probability

$$1 - u\exp\left(-2n\varepsilon^2\right) - \phi(d, 2u)\, 2^{1-\delta u/2}$$

the following holds:

$$\forall f \in \mathcal{F},\ D_Q(f) - \varepsilon \leq D_Q^{0,\mu}(f) - \varepsilon \leq \hat{D}_T^S(f) \leq D_Q^{\delta,\mu}(f) + \varepsilon$$

where $\hat{D}_T^S(f)$ is the empirical depth computed by the depth measure algorithm.

First we recall the definition of ε-nets of Haussler and Welzl (1986):

Definition 15 *Let μ be a probability measure defined over a domain $\mathcal{X}$. Let R be a collection of subsets of $\mathcal{X}$. An ε-net is a finite subset $A \subseteq \mathcal{X}$ such that for every $r \in R$, if $\mu(r) \geq \varepsilon$ then $A \cap r \neq \emptyset$.*

The following theorem shows that a random set of points forms an ε-net with high probability if the VC dimension of R is finite.

Theorem 16 ***Haussler and Welzl, 1986, Theorem 3.7 therein*** *Let μ be a probability measure defined over a domain $\mathcal{X}$. Let R be a collection of subsets of $\mathcal{X}$ with a finite VC dimension d. Let $\varepsilon > 0$ and assume $u \geq 8/\varepsilon$. A sample $S = \{x_i\}_{i=1}^u$ selected at random from μ^u is an ε-net for R with a probability of at least $1 - \phi(d, 2u)\, 2^{1-\varepsilon u/2}$.*

Proof of Theorem 14

By a slight abuse of notation, we use f_δ to denote both a function and a subset of $\mathcal{X}$ that includes every $x \in \mathcal{X}$ for which $D_Q(f|x) \leq D_Q^{\delta,\mu}(f)$. From Theorem 16 it follows that with probability $\geq 1 - \phi(d, 2u)\, 2^{1-\delta u/2}$ a random sample $S = \{x_i\}_{i=1}^u$ is a δ-net for $\{f_\delta\}_{f\in\mathcal{F}}$. Since for every $f \in \mathcal{F}$ we have $\mu(f_\delta) \geq \delta$ we conclude that in these cases,

$$\forall f \in \mathcal{F},\ \exists i \in [1, \ldots, u] \text{ s.t. } x_i \in f_\delta\ .$$

Note that $x_i \in f_\delta$ implies that $D_Q(f|x_i) \leq D_Q^{\delta,\mu}(f)$. Therefore, with probability $1 - \phi(d, 2u)\, 2^{1-\delta u/2}$ over the random selection of $x_1, \ldots, x_u$:

$$\forall f \in \mathcal{F},\ D_Q(f) \leq \min_i D(f|x_i) \leq D_Q^{\delta,\mu}(f)\ .$$

Let $f_1, \ldots, f_n$ be an i.i.d. sample from Q. For a fixed x_i, using Hoeffding's inequality,

$$\Pr_{f_1,\ldots,f_n}\left[\left|\frac{1}{n}\left|f_j : f_j(x_i) = 1\right| - \mu\{f : f(x_i) = 1\}\right| > \varepsilon\right] \leq 2\exp\left(-2n\varepsilon^2\right)\ .$$

Hence, with a probability of $1-u\exp\left(-2n\varepsilon^2\right)$,

$$\forall i,\ \left|\frac{1}{n}\left|f_j : f_j(x_i)=1\right| - \mu\{f\in\mathcal{F} : f(x_i)=1\}\right| \leq \varepsilon\ .$$

Clearly, in the same setting, we also have that

$$\forall i,\ \left|\frac{1}{n}\left|f_j : f_j(x_i)=-1\right| - \mu\{f\in\mathcal{F} : f(x_i)=-1\}\right| \leq \varepsilon\ .$$

Thus, with a probability of at least $1-u\exp\left(-2n\varepsilon^2\right)-\phi(d,2u)\,2^{1-\varepsilon u/2}$ over the random selection of $x_1,\ldots,x_u$ and $f_1,\ldots,f_n$ we have that

$$\forall f\in\mathcal{F},\ \hat{D}_T^S(f) \leq D_Q^{\delta,\mu}(f)+\varepsilon\ .$$

Note also, that with probability 1 there will be no i in the sample such that $D_Q(f\,|\,x_i) < D_Q^{0,\mu}(f)$. Therefore, it is also true that

$$\forall f\in\mathcal{F},\ D_Q^{0,\mu}(f)-\varepsilon \leq \hat{D}_T^S(f)$$

while it is always true that $D_Q(f) \leq D_Q^{0,\mu}(f)$. ■

In Theorem 14 we have seen that the estimated depth uniformly converges to the true depth. However, since we are interested in deep hypotheses, it suffices that the estimate is accurate for these hypotheses, as long as "shallow" hypotheses are distinguishable from the deep ones. This is the motivation for the next theorem:

Theorem 17 *Let Q be a probability measure on $\mathcal{F}$ and let μ be a probability measure on $\mathcal{X}$. Let $\varepsilon,\delta>0$. Assume $\mathcal{F}$ has a finite VC dimension $d<\infty$ and define $\phi(d,k)$ as before. Let $D=\sup_{f\in\mathcal{F}} D_Q(f)$. If S and T are chosen at random from μ^u and Q^n respectively such that $u \geq 8/\delta$ then with probability*

$$1-u\exp\left(-2n\varepsilon^2\right)-\phi(d,2u)\,2^{1-\delta u/2}$$

the following holds:

1. *For every f such that $D_Q^{\delta,\mu}(f) < D$ we have that $\hat{D}_S^T(f) \leq D_Q^{\delta,\mu}(f)+\varepsilon$*
2. *For every f we have that $\hat{D}_S^T(f) \geq D_Q^{0,\mu}(f) \geq D_Q(f)-\varepsilon$*

where $\hat{D}_T^S(f)$ is the empirical depth computed by the depth measure algorithm.

The proof is very similar to the proof of Theorem 14. The key however, is the following lemma:

Lemma 18 *Let $D=\sup_{f\in\mathcal{F}} D_Q(f)$. For every $f\in\mathcal{F}$ let f_δ be such that $f_\delta(x)=1$ if $D_Q(f\,|\,x)<D$ and $f_\delta(x)=-1$ otherwise. Let $\mathcal{F}_\delta$ be*

$$\mathcal{F}_\delta = \{f_\delta\}_{f:D_Q^{\delta,\mu}(f)<D}\ .$$

Then the VC dimension of $\mathcal{F}_\delta$ is upper bounded by the VC dimension of $\mathcal{F}$.

Proof Assume that $x_1,\ldots,x_m$ are shattered by $\mathcal{F}_\delta$. Therefore, for every sequence $y \in \{\pm 1\}^m$ there exists f^y such that f^y_δ induces the labels y on $x_1,\ldots,x_m$. We claim that for every $y \neq y'$, the function f^y and $f^{y'}$ induce different labels on $x_1,\ldots,x_m$ and hence this sample is shattered by $\mathcal{F}$. Let $y \neq y'$ and assume, w.l.o.g. that $y_i = 1$ and $y'_i = -1$. Therefore x_i is such that

$$D_Q\left(f^y \,|\, x_i\right) < D \leq D_Q\left(f^{y'} \,|\, x_i\right) \ .$$

From the definition of the depth on the point x_i, it follows that $D_Q\left(f^y \,|\, x_i\right) \neq D_Q\left(f^{y'} \,|\, x_i\right)$ if and only if $f^y(x_i) \neq f^{y'}(x_i)$. Therefore, the sample $x_1,\ldots,x_m$ being shattered by $\mathcal{F}_\delta$ implies that it is also shattered by $\mathcal{F}$. Hence, the VC dimension of $\mathcal{F}_\delta$ is bounded by the VC dimension of $\mathcal{F}$. ■

Proof of Theorem 17.

From the theory of ε-nets (see Theorem 16), and Lemma 18 it follows that with probability $1-\phi(d,2u)\,2^{1-\delta u/2}$ over the sample S, for every $f \in \mathcal{F}$ such that $D_Q^{\delta,\mu}(f) < D$ there exists x_i such that

$$D_Q\left(f \,|\, x_i\right) \leq D_Q^{\delta,\mu}(f) < D \ .$$

Therefore, with probability greater than $1-\phi(d,2u)\,2^{1-\delta u/2}$, for every f such that $D_Q^{\delta,\mu}(f) < D$ we have that $\hat{D}_S^T(f) \leq D_Q^{\delta,\mu}(f) + \varepsilon$.

To prove the second part, note that with probability 1, for every x and every f, $D_Q\left(f \,|\, x\right) \geq D_Q^{0,\mu}(f)$. Thus if $\hat{D}_S^T(f) < D_Q(f)$ it is only because the inaccuracy in the estimation $\hat{D}_T\left(f \,|\, x_i\right) = \frac{1}{n}\sum_j 1_{f_j(x_i)=f(x_i)}$. We already showed, in the proof of Theorem 14, that with probability of $1-u\exp\left(-2n\varepsilon^2\right)$ over the sample T,

$$\forall i, f, \ \left|\frac{1}{n}\sum_j 1_{f_j(x_i)=f(x_i)} - D_Q\left(f \,|\, x\right)\right| < \varepsilon \ .$$

Hence,

$$\forall f, \ \hat{D}_S^T(f) \geq D_Q^{0,\mu}(f) - \varepsilon \ .$$

■

3.1 Finding the Median

So far we discussed ways to measure the depth. We have seen that if the samples S and T are large enough then with high probability the estimated depth is accurate uniformly for all functions $f \in \mathcal{F}$.

We use these findings to present an algorithm which approximates the predicate median. Recall that the predicate median is a function f which maximizes the depth, that is $f = \arg\max_{f\in\mathcal{F}} D_Q(f)$. As an approximation, we will present an algorithm which finds a function f that maximizes the empirical depth, that is $f = \arg\max_{f\in\mathcal{F}} \hat{D}_T^S(f)$.

The intuition behind the algorithm is simple. Let $S = \{x_i\}_{i=1}^u$. A function that has large empirical depth will agree with the majority vote on these points. However, it might be the case that such a function does not exist. If we are forced to find a hypothesis that does not agree with the majority on

Algorithm 2 Median Approximation (MA)

Inputs:

- A sample $S = \{x_1, \ldots, x_u\} \in \mathcal{X}^u$ and a sample $T = \{f_1, \ldots, f_n\} \in \mathcal{F}^n$.
- a learning algorithm $\mathcal{A}$ that given a sample returns a function consistent with it if such a function exists.

Outputs:

- a function $f \in \mathcal{F}$ which approximates the predicate median, together with its depth estimation $\hat{D}_T^S(f)$

Details:

1. Foreach $i = 1, \ldots, u$ compute $p_i^+ = \frac{1}{n}\left|\left\{j : f_j(x_i) = 1\right\}\right|$ and $q_i = \min\left\{p_i^+, 1 - p_i^+\right\}$.
2. Sort $x_1, \ldots, x_u$ such that $q_1 \geq q_2 \geq \ldots \geq q_m$
3. Foreach $i = 1, \ldots, u$ let $y_i = 1$ if $p_i^+ \geq 0.5$ otherwise, let $y_i = -1$.
4. Use binary search to find i^*, the smallest i for which $\mathcal{A}$ can find a consistent function f with the sample $S^i = \{(x_k, y_k)\}_{k=i}^u$
5. If $i^* \equiv 1$ return f and depth $\hat{D} = 1 - q_1$ else return f and depth $\hat{D} = q_{i^*-1}$.

some instances, the empirical depth will be higher if these points are such that the majority vote on them wins by a small margin. Therefore, we take a sample $T = \left\{f_j\right\}_{j=1}^n$ of functions and use them to compute the majority vote on every x_i and the fraction q_i of functions which disagree with the majority vote. A viable strategy will first try to find a function that agrees with the majority votes on all the points in S. If such a function does not exist, we remove the point for which q_i is the largest and try to find a function that agrees with the majority vote on the remaining points. This process can continue until a consistent function[5] is found. This function is the maximizer of $\hat{D}_T^S(f)$. In the Median Approximation algorithm, this process is accelerated by using binary search. Assuming that the consistency algorithm requires $O(u^c)$ for some c when working on a sample of size u, then the linear search described above requires $O\left(nu + u\log(u) + u^{c+1}\right)$ operations while invoking the binary search strategy reduces the complexity to $O(nu + u\log(u) + u^c \log(u))$.

The Median Approximation (MA) algorithm is presented in Algorithm 2. One of the key advantages of the MA algorithm is that it uses a consistency oracle instead of an oracle that minimizes the empirical error. Minimizing the empirical error is hard in many cases and even hard to approximate (Ben-David et al., 2003). Instead, the MA algorithm requires only access to an oracle that is capable of finding a consistent hypothesis if one exists. For example, in the case of a linear classifier, finding a consistent hypothesis can be achieved in polynomial time by linear programming while finding a hypothesis which approximates the one with minimal empirical error is NP hard. The rest of this section is devoted to an analysis of the MA algorithm.

5. A function is defined to be consistent with a labeled sample if it labels correctly all the instances in the sample.

Theorem 19 ***The MA Theorem***

The MA algorithm (Algorithm 2) has the following properties:

1. *The algorithm will always terminate and return a function $f \in \mathcal{F}$ and an empirical depth $\hat{D}$.*
2. *If f and $\hat{D}$ are the outputs of the MA algorithm then $\hat{D} = \hat{D}_T^S(f)$.*
3. *If f is the function returned by the MA algorithm then $f = \arg\max_{f \in \mathcal{F}} \hat{D}_T^S(f)$.*
4. *Let $\varepsilon, \delta > 0$. If the sample S is taken from μ^u such that $u \geq 8/\delta$ and the sample T is taken from Q^n then with probability of at least*

$$1 - u\exp\left(-2n\varepsilon^2\right) - \phi(d, 2u)\, 2^{1-\delta u/2} \tag{5}$$

the f returned by the MA algorithm is such that

$$D_Q^{\delta,\mu}(f) \geq \sup_{g \in F} D_Q^{0,\mu}(g) - 2\varepsilon \geq \sup_{g \in F} D_Q(g) - 2\varepsilon$$

where d is the minimum between the VC dimension of $\mathcal{F}$ and the VC dimension of the class $\mathcal{F}_\delta$ defined in Theorem 14.

To prove the MA Theorem we first prove a series of lemmas. The first lemma shows that the MA algorithm will always find a function and will return it.

Lemma 20 *The MA algorithm will always return a hypothesis f and a depth $\hat{D}$*

Proof It is sufficient to show that the binary search will always find $i^* \leq u$. Therefore, it is enough to show that there exists i such that $\mathcal{A}$ will return a consistent function f with respect to S^i. To see that, recall that $S^u = \{(x_u, y_u)\}$. Therefore, the sample contains a single point x_u with the label y_u such that at least half of the functions in T are such that $f_j(x_u) = y_u$. Therefore, there exists a function f consistent with this sample. ■

The next lemma proves that the depth computed by the MA algorithm is correct.

Lemma 21 *Let f be the hypothesis that MA returned and let $\hat{D}$ be the depth returned. Then $\hat{D} = \hat{D}_T^S(f)$.*

Proof For any function g, denote by $Y(g) = \{i : g(x_i) = y_i\}$ the set of instances on which g agrees with the proposed label y_i. $\hat{D}_T^S(g)$, the estimated depth of g, is a function of $Y(g)$ given by:

$$\hat{D}_T^S(g) = \min\left(\min_{i \in Y(g)}(1 - q_i), \min_{i \notin Y(g)} q_i\right) .$$

Since the q_i's are sorted, we can further simplify this term. Letting $i_\in = \min\{i : i \in Y(g)\}$ and $i_\notin = \max\{i : i \notin Y(g)\}$, then

$$\hat{D}_T^S(g) = \min\left((1 - q_{i_\in}), q_{i_\notin}\right) .$$

In the above term, if $Y(g)$ includes all i's we consider the term $q_{i_\notin}$ to be one. Similarly, if $Y(g)$ is empty, we consider $q_{i_\in}$ to be zero.

Let f be the hypothesis returned by MA and let $\hat{D}$ be the returned computed depth. If $i^*$is the index that the binary search returned and if $i^* = 1$ then $Y(f) = [1,\ldots,u]$ and $\hat{D}_T^S(f) = 1 - q_1$ which is exactly the value returned by MA. Otherwise, if $i^* > 1$ then $i^* - 1 \notin Y(f)$ but $[i^*,\ldots,u] \subseteq Y(f)$. Since $q_{i^*-1} \leq 0.5$ but for every i' it holds that $1 - q_{i'} \geq 0.5$ so we have that $\hat{D}_T^S(f) = q_{i^*-1}$ which is exactly the value returned by FMA. ■

The next lemma shows that the MA algorithm returns the maximizer of the empirical depth.

Lemma 22 *Let f be the function that the MA algorithm returned. Then*

$$f = \arg\max_{f \in \mathcal{F}} \hat{D}_T^S(f) \ .$$

In the proof of Lemma 21 we have seen that the empirical depth of a function is a function of the set of points on which it agrees with the majority vote. We use this observation in the proof of this lemma too.

Proof Let i^* be the value returned by the binary search and let f be the function returned by the consistency oracle. If $i^* = 1$ then the empirical depth of f is the maximum possible. Hence we may assume that $i^* > 1$ and $\hat{D}_T^S(f) = q_{i^*-1}$.

For a function $g \in \mathcal{F}$, if there exists $i > i^*$ such that $g(x_i) \neq y_i$ then $\hat{D}_T^S(g) \leq q_{i-1} \leq q_{i^*-1} \leq \hat{D}_T^S(f)$. However, if $g(x_i) = y_i$ for every $i \geq i^*$ it must be that $g(x_{i^*-1}) \neq y_{i^*-1}$ or else the binary search phase in the MA algorithm would have found $i^* - 1$ or a larger set. Therefore, $\hat{D}_T^S(g) = q_{i^*-1} = \hat{D}_T^S(f)$. ■

Finally we are ready to prove Theorem 19.

Proof of the MA Theorem (Theorem 19)

Parts 1, 2 and 3 of the theorem are proven by Lemmas 20, 21 and 22 respectively. Therefore, we focus here on the last part.

Let f be the maximizer of $\hat{D}_T^S(f)$ and let $D = \sup_f D_Q^{0,\mu}(f)$. From Theorems 14 and 17 it follows that if d is at least the smaller of the VC dimension of $\mathcal{F}$ and the VC dimension of $\mathcal{F}_\delta$, then with the probability given in (5) we have that

$$\hat{D}_T^S(f) \geq \max_g \hat{D}_T^S(g) \geq \sup_g D_Q^{0,\mu}(g) - \varepsilon = D - \varepsilon \ .$$

Moreover, if $D_Q^{\delta,\mu}(f) < D$ then $\hat{D}_T^S(f) \leq D_Q^{\delta,\mu}(f) + \varepsilon$. Therefore, either

$$D_Q^{\delta,\mu}(f) \geq D$$

or

$$D_Q^{\delta,\mu}(f) \geq \hat{D}_T^S(f) - \varepsilon \geq D - 2\varepsilon$$

which completes the proof. ■

3.2 Implementation Issues

The MA algorithm is straightforward to implement provided that one has access to three oracles: (1) An oracle capable of sampling unlabeled instances $x_1,\ldots,x_u$. (2) An oracle capable of sampling hypotheses $f_1,\ldots,f_n$ from the belief distribution Q. (3) A learning algorithm $\mathcal{A}$ that returns a hypothesis consistent with the sample of instances (if such a hypothesis exists).

The first requirement is usually trivial. In a sense, the MA algorithm converts the consistency algorithm $\mathcal{A}$ to a semi-supervised learning algorithm by using this sample. The third requirement is not too restrictive. In a sense, many learning algorithms would be much simpler if they required a hypothesis which is consistent with the entire sample as opposed to a hypothesis which minimizes the number of mistakes (see, for example, Ben-David et al., 2003). The second requirement, that is sampling hypotheses, is challenging.

Sampling from continuous hypothesis classes is hard even in very restrictive cases. For example, even if Q is uniform over a convex body, sampling from it is challenging but theoretically possible (Fine et al., 2002). A closer look at the MA algorithm and the depth estimation algorithm reveals that these algorithms use the sample of functions in order to estimate the marginal $Q[Y=1|X=x]=\Pr_{g\sim Q}[g(x)=1]$. In some cases, it is possible to directly estimate this value. For example, many learning algorithms output a real value such that the sign of the output is the predicted label and the amplitude is the margin. Using a sigmoid function, this can be viewed as an estimate of $Q[Y=1|X=x]$. This can be used directly in the above algorithms. Moreover, the results of Theorem 14 and Theorem 19 apply with $\varepsilon=0$. Note that the algorithm that is used for computing the probabilities might be infeasible for run-time applications but can still be used in the process of finding the median.

Another option is to sample from a distribution Q' that approximates Q (Gilad-Bachrach et al., 2005). The way to use a sample from Q' is to reweigh the functions when computing $\hat{D}_T(f|x)$. Note that computing $\hat{D}_T(f|x)$ such that it is close to $D_Q(f|x)$ is sufficient for estimating the depth using the depth measure algorithm (Algorithm 1) and for finding the approximated median using the MA algorithm (Algorithm 2). Therefore, in this section we will focus only on computing the empirical conditional depth $\hat{D}_T(f|x)$. The following definition provides the estimate for $D_Q(f|x)$ given a sample T sampled from Q':

Definition 23 *Given a sample T and the relative density function $\frac{dQ}{dQ'}$ we define*

$$\hat{D}_{T,\frac{dQ}{dQ'}}(f)=\frac{1}{n}\sum_j \frac{dQ(f_j)}{dQ'(f_j)}1_{f_j(x)=f(x)}\ .$$

To see the intuition behind this definition, recall that $D_Q(f|x)=\Pr_{g\sim Q}[g(x)=f(x)]$ and $\hat{D}_T(f|x)=\frac{1}{n}\sum_j 1_{f_j(x)=f(x)}$ where $T=\{f_j\}_{j=1}^n$. If T is sampled from Q^n we have that

$$E_{T\sim Q^n}\left[\hat{D}_T(f|x)\right] \;=\; \frac{1}{n}\sum_j E\left[1_{f_j(x)=f(x)}\right]=\frac{1}{n}\sum_j \Pr[f_j(x)=f(x)]=D_Q(f|x)\ .$$

Therefore, we will show that $\hat{D}_{T,\frac{dQ}{dQ'}}(f)$ is an unbiased estimate of $D_Q(f|x)$ and that it is concentrated around its expected value.

Theorem 24 *Let Q and Q' be probability measures over $\mathcal{F}$. Then:*

1. *For every* f, $E_{T\sim Q^n}\left[\hat{D}_{T,\frac{dQ}{dQ'}}(f)\right] = D_Q(f\,|\,x)$

2. *If* $\frac{dQ}{dQ'}$ *is bounded such that* $\frac{dQ}{dQ'} \leq c$ *then*

$$\Pr_{T\sim Q^n}\left[\left|\hat{D}_{T,\frac{dQ}{dQ'}}(f) - D_Q(f\,|\,x)\right| > \varepsilon\right] < 2\exp\left(-\frac{2n\varepsilon^2}{c^2}\right) .$$

Proof To prove the first part we note that

$$\begin{aligned} E_{T\sim Q'^n}\left[\hat{D}_{T,\frac{dQ}{dQ'}}(f)\right] &= E_{T\sim Q'^n}\left[\frac{1}{n}\sum_j \frac{dQ(f_j)}{dQ'(f_j)} 1_{f_j(x)=f(x)}\right] \\ &= E_{g\sim Q'}\left[\frac{dQ(g)}{dQ'(g)} 1_{g(x)=f(x)}\right] = \int_g \frac{dQ(g)}{dQ'(g)} 1_{g(x)=f(x)} dQ'(g) \\ &= \int_g 1_{g(x)=f(x)} dQ(g) = D_Q(f\,|\,x) \ . \end{aligned}$$

The second part is proved by combining Hoeffding's bound with the first part of this theorem. ■

3.3 Tukey Depth and Median Algorithms

To complete the picture we demonstrate how the algorithms presented here apply to the problems of computing Tukey's depth function and finding the Tukey median. In section 2.1 we showed how to cast the Tukey depth as a special case of the predicate depth. We can use this reduction to use Algorithm 1 and Algorithm 2 to compute the Tukey depth and approximate the median respectively. To compute these values, we assume that one has access to a sample of points in $\mathbb{R}^d$, which we denote by $f_1,\ldots,f_n$. We also assume that one has access to a sample of directions and biases of interest. That is, we assume that one has access to a sample of x_i's such that $x_i \in \mathbb{S}\times\mathbb{R}$ where $\mathbb{S}$ is the unit sphere. Hence, we interpret x_i as a combination of a d-dimensional unit vector x_i^v and an offset term x_i^θ. Algorithm 3 shows how to use these samples to estimate the Tukey depth of a point $f\in\mathbb{R}^d$. Algorithm 4 shows how to use these samples to approximate the Tukey median. The analysis of these algorithms follows from Theorems 17 and 19 recalling that the VC dimension of this problem is d.

Computing the Tukey depth requires finding the infimum over all possible directions. As other approximation algorithm do (see Section 4.5) the algorithm presented here finds a minimum over a sample of possible directions represented by the sample S. When generating this sample, it is natural to select x_i^v uniformly from the unit sphere. According to the algorithms presented here one should also select x_i^θ at random. However, for the special case of the linear functions we study here, it is possible to find the minimal depth over all possible selections of x_i^θ once x_i^v is fixed. This can be done by counting the number of f_j's such that $f_j\cdot x_i^v > f\cdot x_i^v$ and the number of f_j's such that $f_j\cdot x_i^v < f\cdot x_i^v$ and taking the minimal value between these two. We use this in the algorithm presented here.

Algorithm 4 selects a set of random directions $x_1,\ldots,x_u$. The median f should be central in every direction. That is, if we project $f_1,\ldots,f_n$ and f on x_i then the projection of f should be close

Algorithm 3 Tukey Depth Estimation

Inputs:

- A sample $S = \{x_1, \ldots, x_u\}$ such that $x_i \in \mathbb{S}$
- A sample $T = \{f_1, \ldots, f_n\}$ such that $f_j \in \mathbb{R}^d$
- A point $f \in \mathbb{R}^d$

Output:

- $\hat{D}_T^S(f)$ - an approximation for the depth of f

Algorithm:

1. for $i = 1, \ldots, u$ compute $\hat{D}_T(f \,|\, x_i) = \frac{1}{n} \min\left(\left|f_j : f_j \cdot x_i > f \cdot x_i\right|, \left|f_j : f_j \cdot x_i < f \cdot x_i\right|\right)$
2. return $\hat{D}_T^S(f) = \min_i \hat{D}(f \,|\, x_i)$

Algorithm 4 Tukey Median Approximation

Inputs:

- A sample $S = \{x_1, \ldots, x_u\}$ such that $x_i \in \mathbb{S}$
- A sample $T = \{f_1, \ldots, f_n\}$ such that $f_j \in \mathbb{R}^d$
- A linear programs solver $\mathcal{A}$ that given a set of linear constraints finds a point that is consistent with the constraints if such a point exists.

Outputs:

- A point $f \in \mathbb{R}^d$ and its depth estimation $\hat{D}_T^S(f)$

Details:

1. Foreach $i = 1, \ldots, u$ and $j = 1, \ldots, n$ compute $f_j \cdot x_i$
2. Let $s_i^1, \ldots s_i^n$ be the sorted values of $f_j \cdot x_i$.
3. Use binary search to find the smallest $k = 0, \ldots, n/2$ for which $\mathcal{A}$ can find f such that

$$\forall i \;\; s_i^{\lfloor \frac{n}{2} \rfloor - k} \leq f \cdot x_i < s_i^{\lceil \frac{n}{2} \rceil + k}$$

4. Return the f that $\mathcal{A}$ found for the smallest k in (3).

to the median of the projection, that is, it should have high one dimensional depth. Therefore, we can start by seeking f with the highest possible depth in every direction. If such f does not exist we can weaken the depth requirement in each direction and try again until we can find a candidate f. Algorithm 4 accelerates this process by using binary search. Note that since the above procedures use only inner products, kernel versions are easily constructed.

4. Relation to Previous Work

In this section we survey the relevant literature. Since depth plays an important role in multivariate statistics it has been widely studied, see Liu et al. (1999), for example, for a comprehensive introduction to statistical depth and its applications in statistics and the visualization of data. We focus only on the part that is related to the work presented here. To make this section easier to follow, we present each related work together with its contexts. Note however that the rest of this work does not build upon information presented in this section and thus a reader can skip this section if he wishes to do so.

4.1 Depth for Functional Data

López-Pintado and Romo (2009) studied depth for functions. The definitions of depth used therein is closer in spirit to the simplicial depth in the multivariate case (Liu, 1990). As a consequence it is defined only for the case where the measure over the function space is an empirical measure over a finite set of functions. Zuo (2003) studied the projection based depth. For a point x in $\mathbb{R}^d$ and a measure ν over $\mathbb{R}^d \times \mathbb{R}$, the depth of x with respect to ν is defined to be

$$\begin{aligned} D_\nu(x) &= \left(1 + \sup_{\|u\|=1} \phi(u,x,\nu)\right)^{-1} \quad \text{where} \\ \phi(u,x,\nu) &\equiv \frac{|u \cdot x - \mu(\nu|_x)|}{\sigma(\nu|_x)} \end{aligned}$$

where μ is a measure of dislocation and σ is a measure of scaling. The functional depth we present in this work can be presented in similar form by defining, for a function f,

$$\begin{aligned} D(f,\nu) &= \left(1 + \sup_{x \in X} \phi(f,x,\nu)\right)^{-1} \quad \text{where} \\ \phi(f,x,\nu) &= |f(x) - E_{g\sim\nu}[g(x)]| \ . \end{aligned}$$

Fraiman and Muniz (2001) introduced an extension of univariate depth to function spaces. For a real function f, the depth of f is defined to be $E_x[D(f(x))]$ where $D(\cdot)$ is the univariate depth function. It is not clear how to use this definition in the binary classification setting. Since the range of the functions contains only two possible values, the univariate rank is of limited utility. However, if we choose the rank function such that the rank of a value is the probability that a function will assign this value, we arrive at a similar definition to the one we propose. The main difference is that Fraiman and Muniz (2001) define the depth as an average over all x's, while in our setting we take the infimum. This plays a key role in our analysis.

4.2 Depth and Classification

Ghosh and Chaudhuri (2005a) used depth for classification purposes. Given samples of data from the different classes, one creates depth functions for each of the classes. At inference time, the depth of a point x is computed with respect to each of the samples. The algorithm associates an instance x with the class in which x is deepest. Ghosh and Chaudhuri (2005a) prove generalization bounds in the case in which each class has a elliptic distribution. Cuevas et al. (2007) used a similar approach and compared the performance of different depth functions empirically. Jörnsten (2004) used a similar approach with an L_1 based depth function. Billor et al. (2008) proposed another variant of this technique.

Ghosh and Chaudhuri (2005b) introduced two variants of depth functions to be used for learning linear classifiers. Let $\{(x_i, y_i)\}_{i=1}^n$ be the training data such that $x_i \in \mathbb{R}^d$ and $y_i \in \pm 1$. In the first variant, the depth of a linear classifier $\alpha \in \mathbb{R}^d$ is defined to be

$$U(\alpha) = \frac{1}{n^+ n^-} \sum_{i:y_i=1} \sum_{j:y_j=-1} \mathbb{I}[\alpha \cdot (x_i - x_j) > 0]$$

where n^+ and n^- are the numbers of positive (and negative) examples, and $\mathbb{I}$ is the indicator function. The regression based depth function is defined to be

$$\Delta(\alpha, \beta) = \frac{\pi^+}{n^+} \sum_{i:y_i=1} \mathbb{I}[\alpha \cdot x_i + \beta > 0] + \frac{\pi^-}{n^-} \sum_{i:y_i=-1} \mathbb{I}[\alpha \cdot x_i + \beta < 0]$$

where π^+ and π^- are positive scalars that sum to one. It is easy to see that the regression based depth defined here is the balanced misclassification probability. The authors showed that as the sample size goes to infinity, the maximal depth classifier is the optimal linear classifier. However, since minimizing this quantity is known to be hard, the authors suggesting using the logistic function as a surrogate to the indicator function. Therefore, these methods are very close (and in some cases identical) to logistic regression.

Gilad-Bachrach et al. (2004) used the Tukey depth to analyze the generalization performance of the Bayes Point Machine (Herbrich et al., 2001). This work uses depth in a similar fashion to the way we use it in the current study. However, the definition of the Bayes depth therein compares the generalization error of a hypothesis to the Bayes classifier in a way that does not allow the use of the PAC-Bayes theory to bound the gap between the empirical error and the generalization error. As a result the analysis in Gilad-Bachrach et al. (2004) was restricted to the realizable case in which the empirical error is zero.

4.3 Regression Depth

Rousseeuw and Hubert (1999) introduced the notion of regression depth. They discussed linear regression but their definition can be extended to general function classes in the following way: Let $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathbb{R}\}$ be a function class and let $S = \{(x_i, y_i)\}_{i=1}^n$ be a sample such that $x_i \in \mathcal{X}$ and $y_i \in \mathbb{R}$. We say that the function $f \in \mathcal{F}$ has depth zero ("non-fit" in Rousseeuw and Hubert, 1999) if there exists $g \in \mathcal{F}$ that is strictly better than f on every point in S. That is, for every point (x_i, y_i) one of the following applies:

$$\begin{aligned} &i. \quad f(x_i) < g(x_i) \leq y_i \\ &ii. \quad f(x_i) > g(x_i) \geq y_i \ . \end{aligned}$$

A function $f \in \mathcal{F}$ is said to have a depth d if d is the minimal number of points that should be removed from S to make f a non-fit.

Christmann (2006) applied the regression depth to the classification task. He used the logit function to convert the classification task to a regression problem. He showed that in this setting the regression depth is closely related to the logistic regression problem and the well known risk minimization technique.

4.4 An Axiomatic Definition of Depth

Most of the applications of depth for classification define the depth of a classifier with respect to a given sample. This is true for the regression depth as well. In that respect, the empirical accuracy of a function is a viable definition of depth. However, in this study we define the depth of a function with respect to a probability measure over the function class. Following Zuo and Serfling (2000) we introduce a definition of a depth function for the classification setting. Our definition has four conditions, or axioms. The first condition is an invariance requirement, similar to the affine invariance requirement in multivariate depth functions. In our setting, we require that if there is a symmetry group acting on the hypothesis class then the same symmetry group acts on the depth function too. The second condition is a symmetry condition: it requires that if the Bayesian optimal hypothesis happens to be a member of the hypothesis class then the Bayesian optimal hypothesis is the median hypothesis. The third condition is the monotonicity condition. This requires that if f_1 and f_2 are two hypotheses such that f_1 is strictly closer to the Bayesian optimal hypothesis than f_2, then f_1 is deeper than f_2. The final requirement is that the depth function is not trivial, that is, that the depth is not a constant value.

Definition 25 *Let $\mathcal{F} = \{f : \mathcal{X} \mapsto \pm 1\}$ and let Q be a probability measure over $\mathcal{F}$. $D_Q : \mathcal{F} \mapsto \mathbb{R}^+$ is a depth function if it has the following properties:*

1. ***Invariance:*** *If $\sigma : \mathcal{X} \mapsto \mathcal{X}$ is a symmetry of $\mathcal{F}$ in the sense that for every $f \in \mathcal{F}$ there exists $f_\sigma \in \mathcal{F}$ such that $f(\sigma(x)) = f_\sigma(x)$ then for every Q and f: $D_Q(f) = D_{Q_\sigma}(f_\sigma)$. Here, Q_σ is such that for every measurable set $A \subseteq \mathcal{F}$ we let $A_\sigma = \{f_\sigma : f \in A\}$ and have that $Q(A) = Q_\sigma(A_\sigma)$.*
2. ***Symmetry:*** *if there exists $f^* \in \mathcal{F}$ such that $\forall x \in \mathcal{X},\ Q\{f : f(x) = f^*(x)\} \geq 1/2$ then $D_Q(f^*) = \sup_f D_Q(f)$.*
3. ***Monotonicity:*** *if there exists $f^* \in \mathcal{F}$ such that $D_Q(f^*) = \sup_f D_Q(f)$ then for every $f_1, f_2 \in \mathcal{F}$, if $f_1(x) \neq f^*(x) \Longrightarrow f_2(x) \neq f^*(x)$ then $D_Q(f_1) \geq D_Q(f_2)$.*
4. ***Non-trivial:*** *for every $f \in \mathcal{F}$, there exist Q such that f is the unique maximizer of D_Q.*

Our definition attempts to capture the same properties that Zuo and Serfling (2000) considered, with a suitable adjustment for the classification setting. It is a simple exercise to verify that the predicate depth meets all of the above requirements.

4.5 Methods for Computing the Tukey Median

Part of the contribution of this work is the proposal of algorithms for approximating the predicate depth and the predicate median. The Tukey depth is a special case of the predicate depth and therefore we survey the existing literature for computing the Tukey median here. Chan (2004) presented

optimal algorithms for computing the Tukey median. Chan presented a randomized algorithm that can find the Tukey median for a sample of n points with expected computational complexity of $O(n \log n)$ when the data is in $\mathbb{R}^2$ and $O\left(n^{d-1}\right)$ when the data is in $\mathbb{R}^d$ for $d > 2$. It is conjectured that these results are optimal for finding the exact median. Massé (2002) analyzed the asymptotic behavior of the empirical Tukey depth. The empirical Tukey depth is the Tukey depth function when it is applied to an empirical measure sampled from the true distribution of interest. He showed that the empirical depth converges uniformly to the true depth with probability one. Moreover, he showed that the empirical median converges to the true median at a rate that scales as $1/\sqrt{n}$. Cuesta-Albertos and Nieto-Reyes (2008) studied the random Tukey depth. They proposed picking k random directions and computing the univariate depth of each candidate point for each of the k directions. They defined the random Tukey depth for a given point to be the minimum univariate depth of this point with respect to the k random directions. In their study, they empirically searched for the number of directions needed to obtain a good approximation of the depth. They also pointed out that the random Tukey depth uses only inner products and hence can be computed in any Hilbert space.

Note that the empirical depth of Massé (2002) and the random Tukey depth of Cuesta-Albertos and Nieto-Reyes (2008) are different quantities. In the empirical depth, when evaluating the depth of a point x, one considers every possible hyperplane and evaluates the measure of the corresponding half-space using only a sample. On the other hand, in the case of random depth, one evaluates only k different hyperplanes. However, for each hyperplane it is assumed that the true probability of the half-space is computable. Therefore, each one of these approaches solves one of the problems involved in computing the Tukey depth. However, in reality, both problems need to be solved simultaneously. That is, since scanning all possible hyperplanes is computationally prohibited, one has to find a subset of representative hyperplanes to consider. At the same time, for each hyperplane, computing the measure of the corresponding half-space is prohibitive for general measures. Thus, an approximation is needed here as well. The solution we present addresses both issues and proves the convergence of the outcome to the Tukey depth, as well as giving the rate of convergence.

Since finding the deepest point is hard, some studies focus on just finding a deep point. Clarkson et al. (1996) presented an algorithm for finding a point with depth $\Omega(1/d^2)$ in polynomial time. For the cases in which we are interested, this could be insufficient. When the distribution is log-concave, there exists a point with depth $1/e$, independent of the dimension (Caplin and Nalebuff, 1991). Moreover, for any distribution there is a point with a depth of at least $1/d+1$ (Carathèodory's theorem).

4.6 PAC-Bayesian Bounds

Our works builds upon the PAC-Bayesian theory that was first introduced by McAllester (1999). These results were further improved in a series of studies (see, for example, Seeger, 2003; Ambroladze et al., 2007; Germain et al., 2009). These results bound, with high probability, the gap between the empirical error of a stochastic classifier based on a posterior Q to the expect error of this classifier in terms of the KL-divergence between Q and the prior P. Some of these studies demonstrate how this technique can be applied to the class of linear classifier and how to improve the bounds by using parts of the training data to learn a prior P to further tighten the generalization bounds.

In the current study we use different approaches which result in different type of bounds. Theorem 4 shows a multiplicative bound on the error of a classifier with respect to the error of the

Gibbs classifier. For example, if the posterior is log-concave and the hypothesis is the mean of the posterior, then the multiplicative factor is $e \cong 2.71$. This bound contains no additive components, therefore, if the generalization error of the Gibbs classifier is small, this new bound may be superior compared to bounds which have additive components (Ambroladze et al., 2007; Germain et al., 2009). The structure of this bound is closer to the consistency bounds for the Nearest Neighbor algorithm (Fix and Hodges Jr, 1951). However, unlike consistency bounds, the bound of Theorem 4 applies to any sample size and any method of obtaining the data.

Another aspect of Theorem 4 is that the bound applies to any classification function. That means that it does not assume that the classifier comes from the same class on which the Gibbs classifier is defined, neither does it make any assumptions on the training process. For example, the training error does not appear in this bound.

Theorem 5 uses the PAC-Bayesian theory to relate the training error of the Gibbs classifier to the generalization error of a deep classifier. In Ambroladze et al. (2007) and Germain et al. (2009) the posterior Q is chosen to be a unit variance Gaussian around the linear classifier of interest. Using the same posterior in Theorem 5 will result in inferior results since there is an extra penalty of factor 2 due to the $1/2$ depth of the center of the Gaussian. However, our bound provides more flexibility in choosing the posterior Q in the tradeoff between the empirical error, the KL divergence and the depth. It is left as an open problem to determine if one can derive better bounds by using this flexibility.

5. Discussion

In this study we addressed the hypothesis selection problem. That is, given a posterior belief over the hypothesis class, we examined the problem of choosing the best hypothesis. To address this challenge, we defined a depth function for classifiers, the predicate depth, and showed that the generalization of a classifier is tied to its predicate depth. Therefore, we suggested that the deepest classifier, the predicate median, is a good candidate hypothesis to select. We analyzed the breakdown properties of the median and showed it is related to the depth as well. We contrasted these results with the more commonly used maximum a posteriori classifier.

In the second part of this work we discussed the algorithmic aspects of our proposed solution. We presented efficient algorithms for uniformly measuring the predicate depth and for finding the predicate median. Since the Tukey depth is a special case of the depth presented here, it also follows that the Tukey depth and the Tukey median can be approximated in polynomial time by our algorithms.

Our discussion was limited to the binary classification case. It will be interesting to see if this work can be extended to other scenarios, for example, regression, multi-class classification and ranking. These are open problems at this point.

References

A. Ambroladze, E. Parrado-Hernández, and J. Shawe-Taylor. Tighter PAC-Bayes bounds. *Advances in neural information processing systems*, 19:9, 2007.

S. Ben-David, N. Eiron, and P. M. Long. On the difficulty of approximately maximizing agreements. *J. Comput. Syst. Sci.*, 66(3):496–514, 2003.

N. Billor, A. Abebe, A. Turkmen, and S. V. Nudurupati. Classification based on depth transvariations. *Journal of classification*, 25(2):249–260, 2008.

C. Borell. Convex set functions in d-space. *Periodica Mathematica Hungarica*, 6:111–136, 1975. ISSN 0031-5303. 10.1007/BF02018814.

A. Caplin and B. Nalebuff. Aggregation and social choice: A mean voter theorem. *Econometrica*, 59(1):1–23, January 1991.

T. M. Chan. An optimal randomized algorithm for maximum Tukey depth. In *SODA*, pages 430–436, 2004.

A. Christmann. Regression depth and support vector machine. *DIMACS series in discrete mathematics and theoretical computer science*, 72:71–86, 2006.

K. L. Clarkson, D. Eppstein, G. L. Miller, C. Sturtivant, and S. H. Teng. Approximating center points with iterative radon points. *Int. J. Comput. Geometry Appl.*, 6(3):357–377, 1996.

C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995.

J. A. Cuesta-Albertos and A. Nieto-Reyes. The random Tukey depth. *Computational Statistics & Data Analysis*, 52, 2008.

A. Cuevas, M. Febrero, and R. Fraiman. Robust estimation and classification for functional data via projection-based depth notions. *Computational Statistics*, 2007.

P. L. Davies and U. Gather. Breakdown and groups. *The Annals of Statistics*, 33(3):977–1035, 2005.

S. Fine, R. Gilad-Bachrach, and E. Shamir. Query by committee, linear separation and random walks. *Theor. Comput. Sci.*, 284(1):25–51, 2002.

E. Fix and J. L. Hodges Jr. Discriminatory analysis-nonparametric discrimination: Consistency properties, 1951.

R. Fraiman and G. Muniz. Trimmed means for functional data. *Test*, 2001.

P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and S. Shanian. From PAC-Bayes bounds to KL regularization. In *ICML*, 2009.

A. K. Ghosh and P. Chaudhuri. On maximum depth and related classifiers. *Scandinavian Journal of Statistics*, 32(2):327–350, 2005a.

A. K. Ghosh and P. Chaudhuri. On data depth and distribution-free discriminant analysis using separating surfaces. *Bernoulli*, 11(1):1–27, 2005b.

R. Gilad-Bachrach, A. Navot, and N. Tishby. Bayes and Tukey meet at the center point. In *Proceedings of the Conference on Learning Theory (COLT)*, pages 549–563. Springer, 2004.

R. Gilad-Bachrach, A. Navot, and N. Tishby. Query by committee made real. In *NIPS*, 2005.

F. R. Hampel. A general qualitative definition of robustness. *The Annals of Mathematical Statistics*, 42(6):1887–1896, 1971.

D. Haussler and E. Welzl. Epsilon-nets and simplex range queries. In *Symposium on Computational Geometry*, pages 61–71, 1986. doi: 10.1145/10515.10522.

X. He. Discussion: Breakdown and groups. *The Annals of Statistics*, 33(3):998–1000, 2005.

R. Herbrich, T. Graepel, and C. Campbell. Bayes point machines. *The Journal of Machine Learning Research*, 1:245–279, 2001.

R. Jörnsten. Clustering and classification based on the L1 data depth. *Journal of Multivariate Analysis*, 90, 2004.

R. Y. Liu. On a notion of data depth based on random simplices. *Annals of Statistics*, 1990.

R. Y. Liu, J. M. Parelius, and K. Singh. Multivariate analysis by data depth: descriptive statistics, graphics and inference,(with discussion and a rejoinder by Liu and Singh). *The Annals of Statistics*, 27(3):783–858, 1999.

S. López-Pintado and J. Romo. On the concept of depth for functional data. *Journal of The American Statistical Association*, 104, 2009.

J. C. Massé. Asymptotics for the Tukey median. *Journal of Multivariate Analysis*, 81, 2002.

D. A. McAllester. Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363, 1999.

P. J. Rousseeuw and M. Hubert. Regression depth. *Journal of the American Statistical Association*, 94(446):388–402, 1999.

M. Seeger. PAC-Bayesian generalisation error bounds for gaussian process classification. *The Journal of Machine Learning Research*, 3:233–269, 2003.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

J. Tukey. Mathematics and the picturing of data. In *Proceedings of the International Congress of Mathematicians*, 1975.

Y. Zuo. Projection-based depth functions and associated medians. *The Annals of Statistics*, 2003.

Y. Zuo and R. Serfling. General notions of statistical depth function. *The Annals of Statistics*, 28 (2):461–482, 2000.