Bayes and Tukey Meet at the Center Point

Ran Gilad-Bachrach[†] Amir Navot[‡] Naftali Tishby[†‡]

[†]School of Computer Science and Engineering
[‡]Interdisciplinary Center for Neural Computation
The Hebrew University, Jerusalem, Israel
ranb,anavot,tishby@cs.huji.ac.il

Abstract. The Bayes classifier achieves the minimal error rate by constructing a weighted majority over all concepts in the concept class. The *Bayes Point* [1] uses the single concept in the class which has the minimal error. This way, the *Bayes Point* avoids some of the deficiencies of the Bayes classifier. We prove a bound on the generalization error for *Bayes Point Machines* when learning linear classifiers, and show that it is at most ~ 1.71 times the generalization error of the Bayes classifier, independent of the input dimension and length of training. We show that when learning linear classifiers, the *Bayes Point* is almost identical to the *Tukey Median* [2] and *Center Point* [3]. We extend these definitions beyond linear classifiers and define the *Bayes Depth* of a classifier. We prove generalization bound in terms of this new definition. Finally we provide a new concentration of measure inequality for multivariate random variables to the *Tukey Median*.

1 Introduction

In this paper we deal with supervised concept learning in a Bayesian framework. The task is to learn a concept c from a concept class $\mathcal{C}$. We assume that the target c is randomly chosen from $\mathcal{C}$ according to a known probability distribution ν. The Bayes classifier is known to be optimal in this setting, i.e. it achieves the minimal possible expected loss. However the Bayes classifier suffers from two major deficiencies. First, it is usually computationally infeasible, since each prediction requires voting over all parameters. The second problem is the possible inconsistency of the Bayes classifier [4], as it is often outside of the target class. Consider for example the following scenario: Alice, Bob and Eve would like to vote on the linear order of three items $\mathcal{A}, \mathcal{B}$ and $\mathcal{C}$. Alice suggests $\mathcal{A} < \mathcal{B} < \mathcal{C}$, Bob suggests $\mathcal{C} < \mathcal{A} < \mathcal{B}$, and Eve suggests $\mathcal{B} < \mathcal{C} < \mathcal{A}$. Voting among the three, as the Bayes classifier does, will lead to $\mathcal{A} < \mathcal{B}$, $\mathcal{B} < \mathcal{C}$ and $\mathcal{C} < \mathcal{A}$ which does not form a linear order.

The computational infeasibility and possible inconsistency of the Bayes optimal classifier are both due to the fact that it is not a single classifier from the given concept class but rather a weighted majority among concepts in the class. These drawbacks can be resolved if one selects a single classifier in the proper class (or a proper ordering in the previous example). Indeed, once the single concept is selected, its predictions are usually both efficient and consistent. It is, however, no longer Bayes optimal. Our problem is to find the single member of the concept class which best approximates the optimal Bayes classifier.

Herbrich, Graepel and Campbell [1] have recently studied this problem. They called the single concept which minimizes the expected error the *Bayes Point*. Specifically for the case of linear classifiers, they designed the *Bayes Point Machine* (BPM), which employs the center of gravity of the version space (which is convex in this case) as the candidate classifier. This method has been applied successfully to various domains, achieving comparable results to those obtained by Support Vector Machines [5].

1.1 The results of this paper

Theorem 1 provides a generalization bound for *Bayes Point Machines*. We show that the expected generalization error of BPM is greater than the expected generalization error of the Bayes classifier by a factor of at most $(e-1) \simeq 1.71$. Since the Bayes classifier obtains the minimal expected generalization error we conclude that BPM is "almost" optimal. Note that this bound is independent of the input dimension and it holds for any size of the training sequence. These two factors, i.e. input dimension and training set size, affect the error of BPM only through the error of the optimal Bayes classifier. The error of *Bayes Point Machines* can also be bounded in the online mistake bound model. In theorem 2 we prove that the mistake bound of BPM is at most $\frac{n}{-\log(1-1/e)} \log \frac{2R}{r}$, where n is the input dimension, R is a bound on the norm of the input data points, and r is a margin term. This bound is different from Novikoff's well known mistake bound for the perceptron algorithm [6] of R^2/r^2. In our new bound, the dependency on the ratio R/r is logarithmic, whereas Novikoff's bound is dimension independent.

The proofs of theorems 1 and 2 follow from a definition of the proximity of a classifier to the Bayes optimal classifier. In the setting of linear classifier the proximity measure is a simple modification of the *Tukey Depth* [2]. The *Tukey Depth* measures the *centrality* of a point in $\mathbb{R}^n$. For a Borell probability measure ν over $\mathbb{R}^n$ the *Tukey Depth* (or halfspace depth) of $x \in \mathbb{R}^n$ is defined as

$$D(x) = \inf \left\{ \nu\left(H\right) \text{ s.t. H is half space and } x \in H \right\} \quad , \tag{1}$$

i.e. the depth of x is the minimal probability of an half space which contains x. Using this definition Donoho and Gasko [7] defined the *Tukey Median* as the point x which maximizes the depth function $D(x)$ (some authors refer to this median as the *Center Point* [3]).

Donoho and Gasko [7] studied the properties of the *Tukey Median*. They showed that the median always exists but need not be unique. They also showed that for any measure ν over $\mathbb{R}^n$, the depth of the *Tukey Median* is at least $\frac{1}{n+1}$. Caplin and Nalebuff [4] proved the *Mean Voter Theorem*. This theorem (using different motivations and notations) states that if the measure ν is log-concave then the center of gravity of ν has a depth of at least $1/e$. ν is log-concave if it conforms with

$$\nu\left(\lambda A+(1-\lambda)B\right)\geq\nu\left(A\right)^{\lambda}\nu\left(B\right)^{1-\lambda}\ .$$

For example, uniform distributions over convex bodies are log-concave, normal and chi-square distributions are log-concave as well. See [8] for a discussion and examples of log-concave measures (a less detailed discussion can be found in appendix A).

The lower bound of $1/e$ for the depth of the center of gravity for log-concave measures is the key to our proofs of the bounds for BPM. The intuition behind the proofs is that any "deep" point must generalize well. This can be extended beyond linear classifiers to general concept classes. We define the *Bayes Depth* of a hypothesis and show in theorem 3 that the expected generalization error of any classifier can be bounded in terms of its *Bayes Depth*. This bound holds for any concept class, including multi-class classifiers.

Finally we provide a new concentration of measure inequality for multivariate random variables to their *Tukey Median*. This is an extension of the well known concentration result of scalar random variables to the median [9].

This paper is organized as follows. In section 2 the *Bayes Point Machine* is introduced and the generalization bounds are derived. In section 3 we extend the discussion beyond linear classifiers. We define the *Bayes Depth* and prove generalization bounds for the general concept class setting. A concentration of measure inequality for multivariate random variables to their *Tukey Median* is provided in section 4. Further discussion of the results is provided in section 5. Some background information regarding concave measures can be found in appendix A. The statement of the Mean Voter Theorem is given in appendix B.

1.2 Preliminaries and Notation

Throughout this paper we study the problem of concept learning with Bayesian prior knowledge. The task is to approximate a concept $c\in\mathcal{C}$ which was chosen randomly using a probability measure ν. The Bayes classifier (denoted by $h_{\mathtt{opt}}$) assigns the instance x to the class with minimal expected loss:

$$h_{\mathtt{opt}}\left(x\right)=\arg\min_{y}E_{c\sim\nu}\left[l\left(y,c\left(x\right)\right)\right] \tag{2}$$

where l is some loss function $l:\mathcal{Y}\times\mathcal{Y}\rightarrow\mathbb{R}$. The Bayes classifier is optimal among all possible classifiers since it minimizes the expected generalization error:

$$\mathrm{error}\left(h\right)=E_{x}\left[E_{c\sim\nu}\left[l\left(h\left(x\right),c\left(x\right)\right)\right]\right] \tag{3}$$

The Bayes classifier achieves the minimal possible error on each individual instance x and thus also when averaging over x. If a labeled sample is available the Bayes classifier uses the posterior induced by the sample, and likewise the expected error is calculated with respect to the same posterior. If the concepts in $\mathcal{C}$ are stochastic then the loss in (2) and (3) should be averaged over the internal randomness of the concepts.

2 Bayes Point Machine

Herbrich, Graepel and Campbell [1] introduced the *Bayes Point Machine* as a tool for learning classifiers. They defined the Bayes Point as follows:

Definition 1. *Given a concept class $\mathcal{C}$, a loss function $l : \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}$ and a posterior ν over $\mathcal{C}$, the* Bayes Point *is:*

$$\arg\min_{h \in \mathcal{C}} E_x \left[E_{c \sim \nu} \left[l\left(h(x), c(x)\right) \right] \right]$$

Note that $E_x \left[E_{c \sim \nu} \left[l\left(h(x), c(x)\right) \right] \right]$ is the average error of the classifier h, as defined in (3), and thus the *Bayes Point*, as defined in definition 1, is simply the classifier in $\mathcal{C}$ which minimizes the average error, while the Bayes optimal rule minimizes the same term without the restriction of choosing h from $\mathcal{C}$.

When applying to linear classifiers with the zero-one loss function[1], [1] assumed a uniform distribution over the class of linear classifiers. Furthermore they suggested that the center of gravity is a good approximation of the *Bayes Point*. In theorem 1 we show that this is indeed the case. The center of gravity is indeed a good approximation of the *Bayes Point*.

We will consider the case of linear classifiers through the origin. In this case the sample space is $\mathbb{R}^n$ and a classifier is half-space through the origin. Formally, any vector $\theta \in \mathbb{R}^n$ represents a classifier. Given an instance $x \in \mathbb{R}^n$ the corresponding label is $+1$ if $\theta \cdot x > 0$ and -1 otherwise. Note that if $\lambda > 0$ then the vector θ and the vector $\lambda\theta$ represent the same classifier; hence we may assume that θ is in the unit ball.

Given a sample of labeled instances, the *Version Space* is defined as the set of classifiers consistent with the sample:

$$\text{Version-Space} = \{\theta \,:\, \|\theta\| \leq 1 \;\text{ and }\; y_i \theta \cdot x_i > 0 \quad \text{forall } 1 \leq i \leq m\}$$

This version space is the intersection of the unit ball with a set of linear constraints imposed by the observed instances and hence it is convex. The posterior is the restriction of the original prior to the version space. Herbrich et al. [1] suggested using the center of gravity of the version space as the hypothesis of the learning algorithm which they named the *Bayes Point Machine*. They suggested a few algorithms which are based on random walks in the version space to approximate the center of gravity.

2.1 Generalization Bounds for Bayes Point Machines

Our main result is a generalization bound for the Bayes Point Machine learning algorithm.

[1] The zero-one loss function is zero whenever the predicted class and the true class are the same. Otherwise, the loss is one.

Theorem 1. *Let ν be a continuous log-concave measure*[2] *over the unit ball in $\mathbb{R}^n$ (the prior) and assume that the target concept is chosen according to ν. Let BPM be a learning algorithm such that after seeing a batch of labeled instances S returns the center of gravity of ν restricted to the version space as a hypothesis $h_{\texttt{bpm}}$. Let $h_{\texttt{opt}}(\cdot)$ be the Bayes optimal classifier. For any $x \in \mathbb{R}^n$ and any sample S*

$$\Pr_c \left[h_{\texttt{bpm}}(x) \neq c(x) \,|S \right] \leq (e-1) \Pr_c \left[h_{\texttt{opt}}(x) \neq c(x) \,|S \right]$$

Theorem 1 proves that the generalization error of $h_{\texttt{bpm}}$ is at most $(e-1) \sim 1.7$ times larger than the best possible. Note that this bound is dimension free. There is no assumption on the size of the training sample S or the way it was collected. However, the size of S, the dimension and maybe other properties influence the error of $h_{\texttt{opt}}$ and thus affect the performance of BPM.

Proof. If ν is log-concave, then any restriction of ν to a convex set is log-concave as well. Since the version space is convex, the posterior induced by S is log-concave. Let $x \in \mathbb{R}^n$ be an instance for which the prediction is unknown. Let H be the set of linear classifiers which predict that the label of x is $+1$, therefore

$$H = \{\theta \,:\, \theta \cdot x \geq 0\}$$

and hence H is a half-space. Algorithm $h_{\texttt{opt}}$ will predict that the label of x is $+1$ iff $\nu\,(H|S) \geq 1/2$. W.l.o.g. assume that $\nu(H|S) \geq 1/2$. We consider two cases.

First assume that $\nu(H|S) > 1 - 1/e$. From theorem 6 and the definition of the depth function (1) it follows that any half space with measure $> 1 - 1/e$ must contain the center of gravity. Hence the prediction made by $h_{\texttt{bpm}}$ is the same as the prediction made by $h_{\texttt{opt}}$.

The second case is when $1/2 \leq \nu\,(H|S) \leq 1 - 1/e$. If BPM predicts that the label is $+1$, then it suffers from the same error as $h_{\texttt{opt}}$. If $h_{\texttt{bpm}}$ predicts that the label of x is -1 then:

$$\frac{\Pr_c \left[h_{\texttt{bpm}}(x) \neq c(x) \,|S \right]}{\Pr_c \left[h_{\texttt{opt}}(x) \neq c(x) \,|S \right]} = \frac{\nu\,(H|S)}{1 - \nu\,(H|S)} \leq \frac{1 - 1/e}{1/e} = e - 1$$

Note that if $\nu\,(H|S) < 1/2$ the prediction of $h_{\texttt{opt}}$ will be that the label of x is -1 and we can apply the same proof to

$$\bar{H} = \{\theta \,:\, \theta \cdot x \leq 0\}$$

$\square$

2.2 Computational Complexity

Theorem 1 provides a justification for the choice of the center of gravity in the *Bayes Point Machine* [1]. Herbrich et al. [1] suggested algorithms for approxi-

[2] See appendix A for discussion and definitions of concave measures. Note however, that the uniform distribution over the version space is always log-concave.

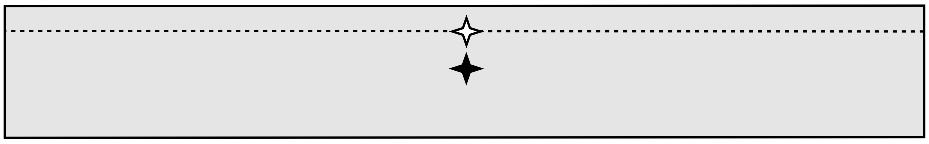

Fig. 1. Although the white point is close (distance wise) to the *Tukey Median* (in black), it does not have large depth, as demonstrated by the dotted line.

mating the center of gravity. In order for our bounds to follow for the approximation, it is necessary to have some lower bound on the *Tukey Depth* of the approximating point. For this purpose, Euclidean proximity is not good enough (see figure 1). Bertsimas and Vempala [10] have suggested a solution for this problem. The algorithm they suggest requires $O^*(n^4)$ operations where n is the input dimension. However it is impractical due to large constants. Nevertheless, the research in this field is active and faster solutions may emerge.

2.3 Mistake Bound

The On-line Mistake-Bound model is another common framework in statistical learning. In this setting the learning is an iterative process, such that at iteration i, the student receives an instance x_i and has to predict the label y_i. After making this prediction, the correct label is revealed. The goal of the student is to minimize the number of wrong predictions in the process.

The following theorem proves that when learning linear classifiers in the online model, if the student makes its predictions using the center of gravity of the current version space, then the number of predictions mistakes is at most $\frac{n}{-\log(1-1/e)} \log \frac{2R}{r}$ where R is a radius of a ball containing all the instances and r is a margin term. Note that the algorithm of the perceptron has a bound of R^2/r^2 in the same setting [6]. Hence the new bound is better when the dimension n is finite (i.e. small).

Theorem 2. *Let $\{(x_i, y_i)\}_{i=1}^{\infty} \subset \mathbb{R}^n \times \{-1, 1\}$ be a sequence such that $\|x_i\|_2 \leq R$ and there exists $r > 0$ and a unit vector $\theta \in \mathbb{R}^n$ such that $y_i x_i \cdot \theta \geq r$ for any i. Let BPM be an algorithm that predicts the label of the next instance x_{m+1} to be the label assigned by the center of gravity of the intersection of the version space induced by $\{(x_i, y_i)\}_{i=1}^{m}$ and the unit ball. The number of prediction mistakes that BPM makes is at most $\frac{n}{-\log(1-1/e)} \log \frac{2R}{r}$.*

Proof. Recall that the version space is the set of all linear classifiers (inside the unit ball) which correctly classifies all instances seen so far. The proof track is as follows: first we will show that the volume of the version space is bounded from below. Second, we will show that whenever a mistake occurs, the volume of the version space reduces by a constant factor. Combining these two together, we conclude that the number of mistakes is bounded.

Let θ be a unit vector such that $y_i x_i \cdot \theta \geq r$. Note that if $\|\theta' - \theta\|_2 < r/R$ then $y_i x_i \cdot \theta' > 0$. Therefore, there exists a ball of radius $r/2R$ inside the unit ball of

$\mathbb{R}^n$ such that all θ' in this ball correctly classify all x_i's. Hence, the volume of the version space is at least $(r/2R)^n V_n$ where V_n is the volume of the n-dimensional unit ball.

Assume that BPM made a mistake while predicting the label of x_i. W.l.o.g. assume that BPM predicted that the label is $+1$. Let $H = \{\theta : \theta \cdot x_i \geq 0\}$, since the center of gravity is in H, and the *Tukey Depth* of the center of gravity $\geq 1/e$, the volume of H is at least $1/e$ of the volume of the version space. This is true since the version space is convex and the uniform measure over convex bodies is log-concave.

Therefore, whenever BPM makes a wrong prediction, the volume of the version space reduces by a factor of $(1 - 1/e)$ at least. Assume that BPM made k wrong predictions while processing the sequence $\{(x_i, y_i)\}_{i=1}^m$ then we have that the volume of the version space is at most $V_n \left(1 - \frac{1}{e}\right)^k$ and at least $V_n \left(\frac{r}{2R}\right)^n$ and thus we conclude that

$$k \leq \frac{n}{-\log\left(1 - \frac{1}{e}\right)} \log \frac{2R}{r}$$

□

3 The Bayes Depth

As we saw in the previous section the *Tukey Depth* plays a key role in bounding the error of *Bayes Point Machine* when learning linear classifiers. We would like to extend these results beyond linear classifiers; thus we need to extend the notion of depth. Recall that the *Tukey Depth* (1) measures the centrality of a point with respect to a probability measure. We say that a point $x \in \mathbb{R}^n$ has depth $D = D(x)$ if when standing at x and looking in any direction, the points you will see have a probability measure of D at least. The question is thus how can we extend this definition to other classes? How should we deal with multi-class partitions of the data, relative to the binary partitions in the linear case? For this purpose we define *Bayes Depth*:

Definition 2. *Let $\mathcal{C}$ be a concept class such that $c \in \mathcal{C}$ is a function $c : \mathcal{X} \to \mathcal{Y}$. Let $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a loss function, and let ν be a probability measure over $\mathcal{C}$. The* Bayes Depth *of a hypothesis h is*

$$D_{Bayes}(h) = \inf_x \frac{\min_{y \in \mathcal{Y}} E_{c \sim \nu}\left[l\left(y, c\left(x\right)\right)\right]}{E_{c \sim \nu}\left[l\left(h\left(x\right), c\left(x\right)\right)\right]} \tag{4}$$

The denominator in (4) is the expected loss of h when predicting the class of x, while the numerator is the minimal possible expected loss, i.e. the loss of the Bayes classifier. Note that the hypothesis h need not be a member of the concept class $\mathcal{C}$. Furthermore, it need not be a deterministic function; if h is stochastic then the loss of h should be averaged over its internal randomness.

An alternative definition of depth is provided implicitly in definition 1. Recall that Herbrich et al. [1] defined the *Bayes Point* h as the point which minimizes the term

$$E_x\left[E_{c \sim \nu}\left[l\left(h(x), c(x)\right)\right]\right] \tag{5}$$

when l is some loss function. Indeed the concept which minimizes the term in (5) is the concept with minimal average loss, and thus this is a good candidate for a depth function. However, evaluating this term requires full knowledge of the distribution of the sample points. This is usually unknown and in some cases it does not exist since the sample point might be chosen by an adversary.

3.1 Examples

Before going any further we would like to look at a few examples which demonstrate the definition of *Bayes Depth*.

Example 1. Bayesian prediction rule

Let h be the Bayesian prediction rule, i.e. $h(x) = \min_{y \in \mathcal{Y}} E_{c' \sim \nu}\left[l\left(y, c'\left(x\right)\right)\right]$. It follows from the definition of depth that $D_{\text{Bayes}}\left(h\right) = 1$. Note that any prediction rule cannot have a depth greater than 1.

Example 2. MAP on finite concept classes

Let $\mathcal{C}$ be a finite concept class of binary classifiers and let l be the zero-one loss function. Let $h = \arg\max_{c \in \mathcal{C}} \nu(\mathcal{C})$, i.e. h is the Maximum A-Posteriori. Since $\mathcal{C}$ is finite we obtain $\nu(h) \geq 1/|\mathcal{C}|$. Simple algebra yields $D_{\text{Bayes}}\left(h\right) \geq \frac{1}{|\mathcal{C}|-1}$.

Example 3. Center of Gravity

In this example we go back to linear classifiers. The sample space consists of tuples (x, b) such that $x \in \mathbb{R}^n$ and $b \in \mathbb{R}$. A classifier is a vector $w \in \mathbb{R}^n$ such that the label w assigns to (x, b) is $\text{sign}(w \cdot x + b)$. The loss is the zero-one loss as before. Unlike the standard setting of linear classifiers the offset b is part of the sample space and not part of the classifier. This setting has already been used in [11]. In this case the *Bayes Depth* is a normalized version of the *Tukey Depth*:

$$D_{\text{Bayes}}\left(w\right) = \frac{D\left(w\right)}{1 - D\left(w\right)}$$

Example 4. Gibbs Sampling

Our last example uses the Gibbs prediction rule which is a stochastic rule. This rule selects at random $c \in \mathcal{C}$ according to ν and uses it to predict the label of x. Note that Haussler et al. [12] already analyzed this special case using different notation. Let h be the Gibbs stochastic prediction rule such that $\Pr\left[h(x) = y\right] = \nu\left\{c : c(x) = y\right\}$. Let l be the zero-one loss function Assume that $\mathcal{Y} = \{-1, +1\}$, and denote by $p = \nu\left\{c : c(x) = +1\right\}$. We obtain $D_{\text{Bayes}}\left(h\right) \geq \inf_{p \in (0,1)} \frac{\min(p, 1-p)}{2p(1-p)} = 0.5$.

3.2 Generalization Bounds

Theorems 1 and 2 are special cases of a general principle. In this section we show that a "deep" classifier, i.e. a classifier with large *Bayes Depth*, generalizes well. We will see that both the generalization error, in the batch framework, and the mistake bound, in the online framework, can be bounded in terms of the *Bayes Depth*.

Theorem 3. *Let $\mathcal{C}$ be a parameter space and let ν be a probability measure (prior or posterior) over $\mathcal{C}$ and l be a loss function. Let h be a classifier then for any probability measure over $\mathcal{X}$*

$$E_{c\sim\nu}E_x\left[l\left(h\left(x\right),c\left(x\right)\right)\right] \leq \frac{1}{D_{Bayes}\left(h\right)}E_{c\sim\nu}E_x\left[l\left(h_{\mathtt{opt}}\left(x\right),c\left(x\right)\right)\right] \tag{6}$$

where $h_{\mathtt{opt}}(\cdot)$ is the optimal predictor, i.e. the Bayes prediction rule.

The generalization bound presented in (6) differs from the common PAC bounds (e.g. [13, 14, ...]). The common bounds provide a bound on the generalization error based on the empirical error. (6) gives a multiplicative bound on the ratio between the generalization error and the best possible generalization error. A similar approach was used by Haussler et al. [12]. They proved that the generalization error of the Gibbs sampler is at most twice as large as the best possible.

Proof. Let $x \in X$ and let $D = D_{\text{Bayes}}\left(h\right)$ be the depth of h. Thus ,

$$D \leq \frac{\min_{y\in\mathcal{Y}} E_{c'\sim\nu}\left[l\left(y,c'\left(x\right)\right)\right]}{E_{c'\sim\nu}\left[l\left(h\left(x\right),c'\left(x\right)\right)\right]}$$

Therefore,

$$\begin{aligned} E_{c'\sim\nu}\left[l\left(h\left(x\right),c'\left(x\right)\right)\right] &\leq \frac{1}{D}\min_{y\in\mathcal{Y}} E_{c'\sim\nu}\left[l\left(y,c'\left(x\right)\right)\right] \\ &= \frac{1}{D}E_{c'\sim\nu}\left[l\left(h_{\mathtt{opt}}\left(x\right),c'\left(x\right)\right)\right] \end{aligned} \tag{7}$$

Averaging (7) over x we obtain the stated result. □

We now turn to prove the extended version of theorem 2, which deals with the online setting. This analysis resembles the analysis of the *Halving* algorithm [15]. However, the algorithm presented avoids the computational deficiencies of the *Halving* algorithm.

Theorem 4. *Let $\{(x_i, y_i)\}_{i=1}^{\infty}$ be a sequence of labeled instances where $x_i \in \mathcal{X}$ and $y_i \in \{\pm 1\}$. Assume that there exists a probability measure ν over a concept class $\mathcal{C}$ such that $\nu\left\{c \in \mathcal{C} : \forall i\ c(x_i) = y_i\right\} \geq \gamma > 0$. Let L be a learning algorithm such that given a training set $S = \{(x_i, y_i)\}_{i=1}^{m}$, L returns a hypothesis h which is consistent with S and such that $D_{Bayes}\left(h\right) \geq D_0 > 0$ (with respect to the measure ν restricted to the version-space and the zero-one loss). Then the algorithm which predicts the label of a new instance using the hypothesis returned by L on the data seen so far will make at most*

$$\frac{\log 1/\gamma}{\log\left(1 + D_0\right)}$$

mistakes.

Proof. Assume that the algorithm presented made a mistake in predicting the label of x_m. Denote by V_{m-1} the version space at this stage; then

$$V_{m-1} = \{c \in \mathcal{C} \,:\, \forall 1 \leq i < m,\; c(x_i) = y_i\}$$

from the definition of the version space and the assumptions of this theorem we have that $\nu(V_{m-1}) \geq \gamma$. We will consider two cases. One is when the majority of the classifiers are misclassifies x_m, and the second is when only the minority misclassifies. If the majority made a mistake then $\nu(V_m) \leq \frac{1}{2}\nu(V_{m-1})$.

However if the minority made a mistake, the hypothesis h returned by L is in the minority, but since $D_{\text{Bayes}}(h) \geq D_0$ we obtain

$$D_0 \geq \frac{\nu\{c \in V_{m-1} \,:\, c(x_m) = -y_m\}}{\nu\{c \in V_{m-1} \,:\, c(x_m) = y_m\}} \tag{8}$$

Note that the denominator in (8) is merely $\nu(V_m)$ while the numerator is $\nu(V_{m-1}) - \nu(V_m)$. Thus

$$D_0 \leq \frac{\nu(V_{m-1}) - \nu(V_m)}{\nu(V_m)} = \frac{\nu(V_{m-1})}{\nu(V_m)} - 1$$

and thus $\nu(V_m) \leq \frac{1}{1+D_0}\nu(V_{m-1})$.

If there were k wrong predictions on the labels of $x_1, \ldots, x_m$ then

$$\nu(V_m) \leq \max\left(\frac{1}{2}, \frac{1}{1+D_0}\right)^k$$

while $\gamma \leq \nu(V_m)$ and thus, since D_0 is upper bounded by 1, we conclude

$$k \leq \frac{\log \gamma}{\log \frac{1}{1+D_0}}$$

□

4 Concentration of Measure for Multivariate Random Variables to the Tukey Median

In previous sections we have seen the significance of the *Tukey Depth* [2] in proving generalization bounds. Inspired by this definition we also used the extended *Bayes Depth* to prove generalization bounds on general concept classes and loss functions. However, the *Tukey Depth* has many other interesting properties. For example, Donoho and Gasko [7] defined the *Tukey Median* as the point which achieves the best *Tukey Depth*. They showed that such a point always exists, but it need not be unique. The *Tukey Median* has high breakdown point [7] which means that it is resistant to outliers, much like the univariate median.

In this section we use *Tukey Depth* to provide a novel concentration of measure inequality for multivariate random variables. The theorem states that any

Lipschitz[3] function from a product space to $\mathbb{R}^n$ is concentrated around its *Tukey Median*.

Theorem 5. *Let $\Omega_1, \ldots, \Omega_d$ be measurable spaces and let $\mathcal{X} = \Omega_1 \times \ldots \times \Omega_d$ be the product space with P being a product measure. Let $F : \mathcal{X} \longrightarrow \mathbb{R}^n$ be a multivariate random variable such that F is a Lipschitz function in the sense that for any $x \in \mathcal{X}$ there exists $a = a(x) \in \mathbb{R}^d_+$ with $\|a\|_2 = 1$ such that for every $y \in \mathcal{X}$*

$$\|F(x) - F(y)\|_2 \leq \sum_{i \,:\, x_i \neq y_i} a_i \tag{9}$$

Assume furthermore that F is bounded such that $\|F(x) - F(y)\| \leq M$.

Let $z \in \mathbb{R}^n$ then for any $r > 0$

$$P_x\left[\|F(x) - z\| \geq r\right] \leq \left(\frac{4M}{r}\right)^n \frac{1}{D(z)} e^{-r^2/16} \tag{10}$$

where $D(z)$ is the Tukey Depth *of z with respect to the push forward measure induced by F.*

Proof. Let $w \in \mathbb{R}^n$ be in the unit ball. From (9), it follows that if $a = a(x)$ then for any $y \in \mathbb{R}^n$

$$F(x) \cdot w - F(y) \cdot w = (F(x) - F(y)) \cdot w \leq \|F(x) - F(y)\| \, \|w\| \leq \sum_{i \,:\, x_i \neq y_i} a_i$$

which means that the functional $x \to F(x) \cdot w$ is Lipschitz. Let $z \in \mathbb{R}^n$ then $\Pr_{x \sim P}\left[F(x) \cdot w \leq z \cdot w\right] \geq D(z)$. Using Talagrand's theorem [16] we conclude that

$$\Pr_{x \sim P}\left[F(x) \cdot w \geq z \cdot w + r/2\right] \leq \frac{1}{D(z)} e^{-r^2/16}$$

clearly this will hold for any vector w such that $\|w\| \leq 1$.

Let W be a minimal $r/2M$ covering of the unit sphere in $\mathbb{R}^n$, i.e. for any unit vector u there exists $w \in W$ such that $\|u - w\| \leq r/2M$. W.l.o.g. W is a subset of the unit ball, otherwise project all the points in W onto the unit ball. Since W is minimal then $|W| \leq (4M/r)^n$. Using the union bound over all $w \in W$ it follows that

$$\Pr_{x \sim P}\left[\exists w \in W,\ \ F(x) \cdot w \geq z \cdot w + r/2\right] \leq \left(\frac{4M}{r}\right)^n \frac{1}{D(z)} e^{-r^2/16}$$

Finally we claim that if x is such that $\|F(x) - z\| \geq r$ then there exists $w \in W$ such that $F(x) \cdot w \geq z \cdot w + r/2$. For this purpose we assume that $z \in \mathrm{conv}(F(X))$ otherwise the statement is trivial since $D(z) = 0$. Let

$$u = \frac{F(x) - z}{\|F(x) - z\|}$$

[3] Lipschitz is in Talagrand's sense. See e.g [9, pg 72-79].

then u is a unit vector and

$$F(x) \cdot u - z \cdot u = (F(x) - z) \cdot u = \|F(x) - z\| \geq r$$

Since w is a cover of the unit sphere and u is a unit vector, there exist $w \in W$ such that $\|w - u\| \leq r/2M$.

$$\begin{aligned} F(x) \cdot w - z \cdot w &= (F(x) - z) \cdot w \\ &= (F(x) - z) \cdot u + (F(x) - z) \cdot (w - u) \\ &\geq r - \|F(x) - z\| \, \|w - u\| \\ &\geq r - (M)\,(r/2M) \\ &= r/2 \end{aligned}$$

and thus $F(x) \cdot w \geq z \cdot w + r/2$. Hence,

$$\begin{aligned} \Pr_x \left[\|F(x) - z\| \geq r\right] &\leq \Pr_x \left[\exists w \in W, \;\; F(x) \cdot w \geq z \cdot w + r/2\right] \\ &\leq \left(\frac{4M}{r}\right)^n \frac{1}{D(z)} e^{-r^2/16} \end{aligned}$$

□

Corollary 1. *In the setting of theorem 5, if m_F is the* Tukey Median *of F, i.e. the* Tukey Median *of the push-forward measure induced by F then for any $r > 0$*

$$P_x \left[\|F(x) - m_F\| \geq r\right] \leq \left(\frac{4M}{r}\right)^n (n+1)\, e^{-r^2/16}$$

Proof. From Helly's theorem [3] it follows that $D\,(m_F) \geq 1/\,(n+1)$ for any measure on $\mathbb{R}^n$. Substitute this in (10) to obtain the stated result. □

Note also that any Lipschitz function is bounded since

$$\|F(x) - F(y)\| \leq \sum_{i: x_i \neq y_i} a_i \leq \sqrt{d}$$

hence M in the above results is bounded by $\sqrt{d}$.

5 Summary and discussion

In this paper we present new generalization bounds for *Bayes Point Machines* [1]. These bounds apply the mean voter theorem [4] to show that the generalization error of *Bayes Point Machines* is greater than the minimal possible error by at most a factor of $(e-1) \sim 1.71$. We also provide a new on-line mistake bound of $\frac{n}{-\log(1-1/e)} \log(2R/r) \sim 2.18n \ln(2R/r)$ for this algorithm.

The notion of *Bayes Point* is extended beyond linear classifiers to a general concept class. We defined the *Bayes Depth* in the general supervised learning

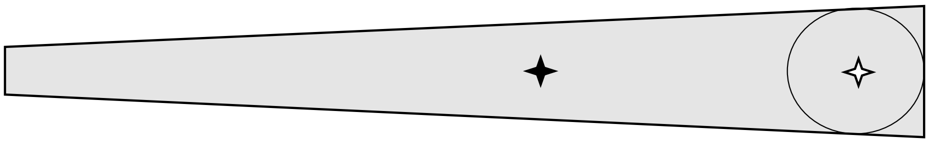

Fig. 2. A comparison of the *Tukey Median* (in black) and the maximal margin point (in white). In this case, the maximal margin point has small *Tukey Depth*

context, as an extension of the familiar *Tukey Depth*. We give examples for calculating the *Bayes Depth* and provide a generalization bound which is applicable to this more general setting. Our bounds hold for multi-class problems and for any loss function.

Finally we provide a concentration of measure inequality for multivariate random variables to their *Tukey Median*. This inequality suggests that the center of gravity is indeed a good approximation to the *Bayes Point*. This provides additional evidence for the fitness of the *Tukey Median* as the multivariate generalization of the scalar median (see also [17] for a discussion on this issue).

The nature of the generalization bounds presented in this paper is different from the more standard bounds in machine learning. Here we bound the multiplicative difference between the learned classifier and the optimal Bayes classifier. This multiplicative factor is a measure of the efficiency of the learning algorithm to exploit the available information. On the other hand, the more standard PAC-like bounds [13, 14, ...], provide an additive bound, on the difference between the training error and the generalization error, with high confidence. The advantage of additive bounds is in their performance guaranty. Nevertheless, empirically it is known that PAC bounds are very loose due to their worst case distributional assumptions. The multiplicative bounds are tighter than the additive ones in these cases.

The bounds for linear *Bayes Point Machines* and the use of *Tukey Depth* can provide another explanation for the success of *Support Vector Machines* [5]. Although the depth of the maximal margin classifier can be arbitrarily small (see figure 2), if the version space is "round" the maximal margin point is close to the *Tukey Median*. We argue that in many cases this is indeed the case.

There seems to be a deep relationship between *Tukey Depth* and *Active Learning*, especially through the *Query By Committee* (QBC) algorithm [11]. The concept of information gain, as used by Freund et al. [11] to analyze the QBC algorithm, is very similar to *Tukey Depth*. This and other extensions are left for further research.

Acknowledgments

We thank Ran El-Yaniv, Amir Globerson and Nati Linial for useful comments. RGB is supported by the Clore foundation. AN is supported by the Horowitz foundation. This work was supported in part by the IST Programme of the

European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views.

References

1. Herbrich, R., Graepel, T., Campbell, C.: Bayes point machines. Journal of Machine Learning Research (2001)
2. Tukey, J.: mathematics and picturing data. In: proceeding international congress of mathematics. Number 2 (1975) 523–531
3. Matoušek, J.: Lectures on discrete geometry. Springer-Verlag (2002)
4. Caplin, A., Nalebuff, B.: Aggregation and social choice: A mean voter theorem. Exonometrica **59** (1991) 1–23
5. Vapnik, V.: Statistical Learning Theory. Wiley (1998)
6. Novikoff, A.B.J.: On convergence proofs on perceptrons. In: Proceedings of the Symposium on the Mathematical Theory of Automata. Volume 12. (1962) 615–622
7. Donoho, D., Gasko, M.: Breakdown properties of location estimates based on halfspace depth and projected outlyingness. Annals of Statistics **20** (1992) 1803–1827
8. Bagnoli, M., Bergstrom, T.: Log-concave probability and its applications. http://www.econ.ucsb.edu/~tedb/Theory/logconc.ps (1989)
9. Ledoux, M.: The Concentration of Measure Phenomenon. American Mathematical Society (2001)
10. Bertsimas, D., Vempala, S.: Solving convex programs by random walks. In: STOC. (2002) 109–115
11. Freund, Y., Seung, H., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Macine Learning **28** (1997) 133–168
12. Haussler, D., Kearns, M., Schapie, R.E.: Bounds on the sample complexity of bayesian learning using information theory and the vc dimension. Machine Learning **14** (1994) 83–113
13. Vapnik, V., Chervonenkis, A.Y.: On the uniform covergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications **16** (1971) 264–280
14. Bartlett, P., Mendelson, S.: Rademacher and gaussian complexities: risk bounds and structural results. Journal of Machine Learning Research **3** (2002) 463–482
15. Littlestone, N.: Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. In: In 28th Annual Symposium on Foundations of Computer Science. (1987) 68–77
16. Talagrand, M.: Concentration of measure and isoperimetric inequalities in product space. Publ. Math. I.H.E.S. **81** (1995) 73–205
17. Zuo, Y., Serfling, R.: General notions of statistical depth function. The Annals of Statistics **28** (2000) 461–482
18. Prekopa, A.: Logarithmic concave measures with applications to stochastic programming. Acta Sci. Math. (Szeged) **32** (1971) 301–315
19. Borell, C.: Convex set functions in d-space. Periodica Mathematica Hungarica **6** (1975) 111–136

A Concave Measures

We provide a brief introduction to concave measures. See [8, 4, 18, 19] for more information about log-concavity and log-concave measures.

Definition 3. *A probability measure ν over $\mathbb{R}^n$ is said to be* log-concave *if for any measurable sets A and B and every $0 \leq \lambda \leq 1$ the following holds:*

$$\nu\left(\lambda A + (1-\lambda)B\right) \geq \nu(A)^{\lambda}\,\nu(B)^{1-\lambda}$$

Note that many common probability measures are log-concave, for example uniform measures over compact convex sets, normal distributions, chi-square and more. Moreover the restriction of any log-concave measure to a convex set is a log-concave measure.

In some cases, there is a need to quantify concavity. The following definition provides such a quantifier.

Definition 4. *A probability measure ν over $\mathbb{R}^n$ is said to be* ρ-concave *if for any measurable sets A and B and every $0 \leq \lambda \leq 1$ the following holds:*

$$\nu\left(\lambda A + (1-\lambda)B\right) \geq \left[\lambda\nu(A)^{\rho} + (1-\lambda)\nu(B)^{\rho}\right]^{1/\rho}$$

A few facts about ρ-concave measures:

- If ν is ρ-concave with $\rho = \infty$ then $\nu(\lambda A + (1-\lambda)B) \geq \max(\nu(A), \nu(B))$.
- If ν is ρ-concave with $\rho = -\infty$ then $\nu(\lambda A + (1-\lambda)B) \geq \min(\nu(A), \nu(B))$.
- If ν is ρ-concave with $\rho = 0$ then $\nu(\lambda A + (1-\lambda)B) \geq \nu(A)^{\lambda}\nu(B)^{1-\lambda}$, in this case ν is called log-concave.

B Mean Voter Theorem

Caplin and Nalebuff [4] proved the Mean Voter Theorem in the context of the voting problem. They did not phrase their theorem using *Tukey Depth* but the translation is trivial. Hence, we provide here (without proof) a rephrased version of their theorem.

Theorem 6. *(Caplin and Nalebuff) Let ν be a ρ-concave measure over $\mathbb{R}^n$ with $\rho \geq -1/(n+1)$. Let z be the center of gravity of ν, i.e. $z = E_{x\sim\nu}[x]$. Then*

$$D(z) \geq \left(\frac{n + 1/\rho}{n + 1 + 1/\rho}\right)^{n+1/\rho} \tag{11}$$

where $D(\cdot)$ is the Tukey Depth.

First note that when $\rho \to 0$ the bound in (11) approches $1/e$; hence for log-concave measures $D(z) \geq 1/e$. However, this bound is better than $1/e$ in many cases, i.e. when $\rho > 0$. This fact can be used to obtain an improved version of theorems 1 and 2.