The Median Hypothesis

Ran Gilad-Bachrach
Chris J. C. Burges
Microsoft Research

Technical Report: MSR-TR-2012-56 (May 2012)

Abstract

A typical learning process begins with some prior beliefs about the target concept. These beliefs are refined using evidence, typically a sample. The evidence is used to compute a "posterior belief", which assigns a probability distribution over the hypothesis class $\mathcal{F}$. Using the posterior, a hypothesis is selected to predict the labels during the generalization phase. A natural candidate is the Bayes hypothesis, that is, the weighted majority vote. However, this hypothesis is often prohibitively costly to compute. Therefore, there is a need for an algorithm to construct a hypothesis (either a single hypothesis from $\mathcal{F}$, or from a different class), given the posterior belief. Several methods have been proposed for this problem: for example, Gibbs sampling, ensemble methods, and choosing the maximum posterior. We propose a new method: choosing the *median hypothesis*. This method is close to the average Gibbs hypothesis and the Bayes hypothesis in terms of accuracy while having the same run-time efficiency, during the generalization phase, as the maximum posterior method.

In this paper, we define a measure of depth for hypotheses, from which we derive the median hypothesis. We prove generalization bounds which leverage the PAC-Bayes analysis technique. We present an algorithm to approximate the median hypotheses and we prove its correctness. Our definition of median is closely related to Tukey's median; in fact, to the best of our knowledge, our algorithm is the first polynomial approximation algorithm to the Tukey median.

1. Introduction

In the learning process, one starts with one's prior beliefs, refines them using some data (evidence) and finally chooses a hypothesis. A schematic diagram of this process is presented in Figure 1. The initial belief, given as a probability measure P over a function class $\mathcal{F}$, is refined using the evidence to form the posterior Q, which is again a probability measure over $\mathcal{F}$. Here, we draw a distinction between $\mathcal{F}$, the class of functions available to the learning algorithm, and a hypothesis, which may not even be a member of $\mathcal{F}$, although it is constructed from elements of $\mathcal{F}$. For example, many learning algorithms use a sample $\{(x_i, y_i)\}_{i=1}^m$ and employ an evaluation function of the form:

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

where l is some loss function and $r(\cdot)$ is a regularization term. The energy function, $E(f)$, can be converted into prior and posterior probabilities with density functions p and q defined respectively as:

$$p(f) := \frac{1}{Z_p}\exp(-\beta r(f)), \qquad q(f) := \frac{1}{Z_q}\exp(-\beta E(f))$$

where Z_p and Z_q are the "partition functions" and $\beta > 0$ is a constant (the reciprocal of the temperature) (Sollich (2002)).

Figure 1: A schematic view of learning

Method	Run Time Efficiency	Information content on the posterior
Gibbs sampling	High	Low
MAP	High	Low
Bayes estimator	Low (typically)	High
Ensembles (small)	High	Low
Ensembles (large)	Low	High
Median	High	High

Figure 2: Different methods of turning the posterior Q into a hypothesis.

Given the posterior measure Q, one has to pick a hypothesis to use. For example, one could choose to use the Bayes classifier which will conduct a weighted majority vote for every prediction. However, this method is often prohibitively slow due to its run-time computational complexity. One might instead choose the Maximum A Posteriori (MAP) classifier. This method will be efficient in terms of run-time computational complexity, but it is very loose in capturing the information conveyed by Q. The main shortcoming of the MAP estimator is that it focuses on the density at a single point. While doing that, it might ignore a large population that may be better represented by a different hypothesis. Moreover, the MAP estimator is not stable in the sense that a minor change to Q can result in the estimator changing in an arbitrary way. These effects are discussed in detail in Appendix A.

Another possible approach for selecting a hypothesis given Q is to sample hypotheses according to Q and thus form an ensemble. In a sense, the bagging method (Breiman (1996)) does exactly this. The run-time computational complexity of this method depends on the size of the ensemble while the ability of the ensemble to capture the information encoded in Q grows with the size of the ensemble. A schematic relating the different approaches is shown in Figure 2.

In this work we present a new method for constructing a representative hypothesis using Q. The method constructs a hypothesis which is deep (we will define 'depth' precisely in Section 2). We call this hypothesis the median hypothesis, by analogy to the notion of median in one dimension. If the median is indeed deep, it is guaranteed that whenever the weight of the majority is large enough, the median hypothesis will vote with the majority; the median can deviate from the majority only in cases where the majority is marginal. Using PAC-Bayesian results, we provide generalization bounds for these types of classifiers. Moreover, we show that in many realistic settings, for example when learning linear classifiers and using a convex energy function $E(f)$, the median is indeed deep.

In the second part of this work we present an algorithm for finding the median. The algorithm uses an unlabeled sample of points $x_1, \ldots, x_u$ and a sample of functions $f_1, \ldots, f_n$ to find a point with a large empirical depth. We prove that the gap between the empirical depth and the true depth is small and hence the hypothesis found is a good approximation to the median. In the case of linear classifiers, this algorithm approximates the Tukey Median (Tukey (1975); Donoho and Gasko (1992)). Table 1 contains a summary of the notation used in this work.

The notion of depth is a key concept in our discussion. Depth is a measure of the centrality of a point in a distribution. It has been used before in the context of learning. Rousseeuw and Hubert (1998) have defined the notion of regression depth. However, in this work we focus on classification tasks. Gilad-Bachrach et al. (2004) have used depth in the context of classification. However, their

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 test 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

work assumes that nature selects a classifier from a distribution known to the learning algorithm. Furthermore, it assumes that there is no noise in the learning process. In our work, we lift these two assumptions and present a method that works even when the target concept is chosen by an adversary and the labels are corrupted by noise.

2. Depth

Depth is a useful tool in multivariate statistics (see e.g., Tukey (1975); Liu (1990)). Given a distribution over points in R^d, a deep point is a point which is central to the distribution; the "shallow" points are the outliers. There are many different depth functions. For example, Tukey defined the Tukey depth in the following way: the depth of a point $x \in R^d$ given a distribution Q is

$$\text{Tukey-Depth}_Q(x) = \inf_{w \in R^d} Q\{x' : w \cdot (x' - x) \geq 0\} \quad . \tag{2}$$

The depth function extends the definition of the median to the multivariate case by defining the generalized median to be the deepest point according to the chosen depth function. The Tukey depth also has a minimum entropy interpretation: given x, consider all hyperplanes containing x. Each hyperplane splits the distribution Q in two, forming a Bernoulli distribution. Choose that hyperplane whose corresponding Bernoulli has minimum entropy. The Tukey depth is then the probability mass on the side of the chosen hyperplane with the lowest mass.

In this paper, 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 $\mathcal{F}$. Following Rousseeuw

and Hubert (1998), who introduced the notion of depth for regression, we define the depth for binary classification functions.

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

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

The depth of f with respect to Q is:

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

The Tukey-Depth is a special case of this definition where$\mathcal{F} = \mathbb{R}^d$ and $\mathcal{X} = \mathbb{R}^d \times \mathbb{R}$ such that $f \in \mathcal{F}$ operates on $x = (x_v, x_\theta) \in \mathcal{X}$ by $f(x) = \operatorname{sign}(f \cdot x_v - x_\theta)$. See Appendix B for details.

The depth $D_Q(f)$ is defined as the infimum over all points $x \in \mathcal{X}$. However, for our applications, we can tolerate a small fraction of 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 2. *Let $\mathcal{F}$ be a function class and 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:*

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

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

With these definitions in hand, we turn to 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 3. ***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 $f \in F$ the following holds:

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

and

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

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 classifier). 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 $x^* \in X$ we have that

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

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

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

(3) follows in the same way by defining Z to be constantly zero and δ to be zero as well. □

Theorem Deep vs. Gibbs (Theorem 3) bounds the ratio of the generalization error of the Gibbs classifier and the generalization error of a classifier as a function of the depth of the latter classifier. For example, consider the Bayes classifier. By definition, the depth of this classifier is at least one half; thus Theorem 3 recovers the well-known result that the generalization error of the Bayes classifier is at most twice larger than the generalization error of the Gibbs classifier.

Next, we combine Theorem Deep vs. Gibbs (Theorem 3) with PAC-Bayesian bounds (McAllester (1999)) to bound the difference between the training error and the test error. We use the version of the PAC-Bayesian bounds in Theorem 3.1. of Germain et al. (2009).

Theorem 4. ***Generalization Bounds***

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:

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

Furthermore, for every $\delta' > 0$, the following holds with a probability greater than $1 - \delta$ over the sample $\mathcal{S}$ sampled from ν^m:

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

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

Proof. Applying the bounds in Theorem 3 to the PAC-Bayesian bounds in Theorem 3.1 of Germain et al. (2009) yields the stated results. □

The generalization bounds theorem (Theorem 4) 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, a question remains: are there any deep functions? The following theorem shows that in many interesting settings this is indeed the case.

Theorem 5. *Let $\mathcal{X} = R^d$ and $\mathcal{F}$ be the class of linear threshold functions over $\mathcal{X}$ (see Appendix B). Let Q be a probability measure over $\mathcal{F}$ with density function $q(f)$ such that $q(f) = \frac{1}{Z}\exp\left(-E(f)\right)$ where $E(f)$ is a convex function. Then there exists a function $f \in \mathcal{F}$ such that $D_Q(f) \geq 1/e$.*

Proof. It is straight-forward to verify that $q(f)$ 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 and using the Mean Voter Theorem of Caplin and Nalebuff (1991) we conclude that the center of gravity of Q has a depth of at least $1/e$ (e is Euler's number). □

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 $\hat{D}_T(f \mid x_i) = \frac{1}{n} \sum_j 1_{f_j(x_i) = f(x_i)}$
2. return $\hat{D}_T^S(f) = \min_i \hat{D}(f \mid x_i)$

Recall that it is a 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 function 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 $1/e$.

3. Measuring depth

In the previous section we proved generalization bounds as a function of the depth function. Moreover, we showed that in many natural cases, the median function is indeed deep. This provides the motivation to seek deep functions. Finding a deep function, even in the case of linear functions, is hard. The best known algorithms have exponential dependency in the dimension (Chan (2004)). When considering approximations of the Tukey median, the best known algorithms are super-polynomial in the dimension (Clarkson et al. (1996)). In the Section 4 we present a polynomial algorithm for approximating the median for any class of binary classification functions. However, before we present this algorithm we will study the following questions: given a function f and a probability measure Q, can we compute or approximate $D_Q(f)$, the depth of f?

We suggest a straight-forward method to measure the depth of a function . The depth estimation algorithm (Algorithm 1) takes as inputs two samples. One sample, $x_1, \ldots, x_u$, is a sample of points from the domain $\mathcal{X}$. The other sample, $f_1, \ldots, f_n$, 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 4.

Theorem 6. ***Uniform convergence of depth***

Let Q be a probability measure on $\mathcal{F}$ and let μ be a probability measure on $\mathcal{X}$. Let $\epsilon, \delta > 0$.For every $f \in \mathcal{F}$ let the function f_δ be such that $f_\delta(x) = 1$ if $D_Q(f \mid x) \leq D_Q^{\delta,\mu}(f)$ 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 S and T are chosen at random from μ^u and Q^n respectively such that $u \geq 8/\delta$ then with probability

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

the following holds:

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

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 7. *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 $\mu(r) \geq \epsilon$ then $A \cap r \neq \emptyset$.*

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

Theorem 8. ***(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 $\epsilon > 0$ and assume $u \geq 8/\epsilon$. A random 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 $1 - \phi(d, 2u)\, 2^{1-\epsilon u/2}$.*

Proof. **of the uniform convergence of depth theorem (Theorem 6)**.

By a slight abuse of notation, we use f_δ both as a function and as a subset of $\mathcal{X}$ that includes every $x \in \mathcal{X}$ for which $D_Q(f \mid x) \leq D_Q^{\delta,\mu}(f)$. From Theorem 8 it follows that with a probability $\geq 1 - \phi(d, 2u)\, 2^{1-\delta m/2}$ 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 $D_Q(f \mid x_i) \leq D_Q^{\delta,\mu}(f)$. Therefore, with probability $1 - \phi(d, 2u)\, 2^{1-\delta u/2}$ over the random selection of $x_1, \ldots, x_u$:

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

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

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

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

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

Clearly, in the same setting, we also have that

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

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

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

□

4. Finding the median

In Section 3 we have seen that it is possible to measure depth. Moreover, 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 median. Recall that the 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 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 $T = \{f_j\}_{j=1}^n$ 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 point 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[1] 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 $O(u^c)$ when working on a sample of size u. The linear search described above requires $O\left(nu + u\log(u) + u^{c+1}\right)$ operations while the binary search strategy the complexity goes down 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 analyzing the MA algorithm.

Theorem 9. ***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 $\epsilon, \delta > 0$. 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*

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

the f returned by the MA algorithm is such that

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

where d is as defined in Theorem 6.

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.

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

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 $f \in \mathcal{F}$ and its depth estimation $\hat{D}_T^S(f)$

Details:

1. Foreach $i = 1, \ldots, u$ compute $p_i^+ = \frac{1}{n} |\{j : f_j(x_i) = 1\}|$ 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 $\hat{D} = q_{i^*-1}$.

Lemma 10. *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 11. *Let f be the hypothesis that MA returned and $\hat{D}$ 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. Let $i_\in = \min\{i : i \in Y(g)\}$ 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 $\hat{D}$ be the computed depth returned. If i^* is the index that the binary search returned and $i^* = 1$ 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 $1 - q_{i'} \geq 0.5$ 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 12. *If f is 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 11 we have seen that the empirical depth of a function is a function of the set of points on which it predicts 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 maximal 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 9.

Proof. **of the MA Theorem (Theorem 9)**

Parts 1, 2 and 3 of the theorem are proven by Lemmas 10, 11 and 12 respectively. The last part follows since from Theorem 6 it follows that with the stated probability $\forall g \in \mathcal{F}$:

$$D_Q(g) - \epsilon \leq \hat{D}_T^S(g) \leq D_Q^{\delta,\mu}(g) + \epsilon \ . \tag{5}$$

Let $g \in \mathcal{F}$ and assume (5) holds. Since f maximizes the empirical depth $\hat{D}_T^S(f)$ we have that

$$\forall g \in \mathcal{F}, \ \ D_Q^{\delta,\mu}(f) + \epsilon \geq \hat{D}_T^S(f) \geq \hat{D}_T^S(g) \geq D_Q(g) - \epsilon \ .$$

Therefore

$$\forall g \in \mathcal{F}, \ \ D_Q^{\delta,\mu}(f) \geq D_Q(g) - 2\epsilon \ .$$

□

One of the weaknesses of Theorem 9 is the use of the VC-dimension of the special class of functions $\mathcal{F}_\delta$. Computing this value is in many cases hard. Nevertheless, the following theorem shows that the VC-dimension of the class $\mathcal{F}$ can be used to provide performance guarantees for the algorithm.

Theorem 13. *Let $\epsilon, \delta > 0$. 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 $1 - u\exp\left(-2n\epsilon^2\right) - \phi(d, 2u)\, 2^{1-\delta u/2}$ the f returned by the MA algorithm is such that $D_Q^{\delta,\mu}(f) \geq \sup_{g \in F} D_Q(g) - 2\epsilon$ where d is the VC dimension of $\mathcal{F}$.*

Proof. Let $D = \sup_{f \in \mathcal{F}} D_Q(f)$. By abuse of notation, we define for every $f \in \mathcal{F}$ the function f_δ be such that $f_\delta(x) = 1$ if $D_Q(f \,|\, x) < d^*$ and $f_\delta(x) = -1$ otherwise. We define the class $\mathcal{F}_\delta$ such that

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

First, we show that the VC dimension of $\mathcal{F}_\delta$ is upper bounded by the VC dimension of $\mathcal{F}$. 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_δ^y 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 sampled 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

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

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

From the theory of ϵ-nets (see Theorem 8), it follows that with probability $1-\phi\left(d, 2u\right) 2^{1-\delta u/2}$ over the sample S, for every $f \in \mathcal{F}$ such that $D_Q^{\delta,\mu}\left(f\right) < D$ there exists x_i such that

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

Therefore, with probability greater than

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

we will have the following:

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

Since we already showed that the algorithm finds a function f that maximizes $\hat{D}_S^T\left(f\right)$, then we have that if f was returned by the algorithm then either $D_Q^{\delta,\mu}\left(f\right) < D$ or

$$D_Q^{\delta,\mu}\left(f\right) \geq \hat{D}_S^T\left(f\right) - \epsilon \geq \sup_g \hat{D}_S^T\left(g\right) - \epsilon \geq D - 2\epsilon \ .$$

□

5. Implementation issues

The MA algorithm is straight forward to implement provided that you have 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 (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 hypotheses 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 $Q\left[Y=1|X=x\right] = \Pr_{g\sim Q}\left[g\left(x\right)=1\right]$. 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 an RBF function, this can be viewed as an estimate of $Q\left[Y=1|X=x\right]$. This can be used directly in the above algorithms. Moreover, the results of Theorem 6 and Theorem 9 apply with $\epsilon = 0$. Note that the algorithm that is used for computing the probabilities might be infeasible for run-time application 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 $\hat{D}_T\left(f \mid x\right)$. Note that computing $\hat{D}_T\left(f \mid x\right)$ such that it is close to $D_Q\left(f \mid x\right)$ 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 $\hat{D}_T(f \mid x)$. The following definition provides the estimate for $D_Q(f \mid x)$ given a sample T sampled form Q':

Definition 14. *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 $D_Q(f \mid x) = \Pr_{g\sim Q}[g(x)]$ and $\hat{D}_T(f \mid x) = \frac{1}{n}\sum_j 1_{f_j(x)=f(x)}$ where $T = \{f_j\}_{j=1}^n$. If T is sampled from Q^n we have that

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

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

Theorem 15. *Let Q and Q' be probability measures over $\mathcal{F}$ then:*

1. *For every f, $E_{T\sim Q'^n}\left[\hat{D}_{T,\frac{dQ}{dQ'}}(f)\right] = D_Q(f \mid x)$*
2. *If $\frac{dQ}{dQ'}$ is bounded such that $\frac{dQ}{dQ'} \le c$ then*

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

Proof. To prove the first part we note that

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

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

6. Discussion

In this work we present a novel method for selecting a hypothesis that will generalize well given a posterior belief. We proposed using the median hypothesis and analyzed its performance as a function of its depth. We also presented algorithms for approximating the median and we analyzed their performance. One possible application for our algorithm is approximating the Tukey median in polynomial time. As far as we know, we are the first to provide an approximation algorithm for the Tukey median which has a polynomial dependency on the dimension.

This work can be extended in several ways. First, we are interested in conducting an empirical study to test the performance of the median hypothesis. Moreover, we are interested in extending this work beyond binary classification. The ability to approximate the Tukey median in polynomial time open new possibilities too. These are the topics of our current research.

References

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.

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

L. Breiman. Bagging predictions. *Machine Learning*, 24, 1996. URL `http://www.stat.rutgers.edu/home/rebecka/Stat687/breiman.pdf`.

A. Caplin and B. Nalebuff. Aggregation and social choice: A mean voter theorem. *Econometrica*, 59 (1):1–23, January 1991. URL `http://ideas.repec.org/a/ecm/emetrp/v59y1991i1p1-23.html`.

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

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.

D. L. Donoho and M. Gasko. Breakdown properties of location estimates based on half-space depth and projected outlyingness. *Annals of Statistics*, 1992.

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

P. Germain, A. Lacasse, F. Laviolette, M. Marchand, and S. Shanian. From pac-bayes bounds to kl regularization. In *NIPS*, 2009.

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.

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.

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

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*, 94:388–402, 1998.

P. Sollich. Bayesian methods for support vector machines: Evidence and predictive class probabilities. *Machine Learning*, 46:21–52, 2002. URL `http://www.cs.iastate.edu/ honavar/bayes-svm.pdf`.

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

Appendix A. Drawbacks of MAP

In our discussion we proposed that the maximizing the posterior might not be the optimal way to choose a function with which to make predictions. We would like to demonstrate a failure of this method on a toy example. Assume that the sample space is the set $\mathcal{X} = [1, \ldots, N]$. I.e., the sample space is a set of N discrete elements. The function class $\mathcal{F}$ is made of $N+2$ functions described as follows: for every $i \in \mathcal{X}$ the function f_i is defined to be

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

Additionally, $\mathcal{F}$ contains the constant functions f^0 and f^1 that assigns the values 1 and 0 respectively for every input. Our prior is uniform over the set $\mathcal{F}$.

Assume also that it is our belief that if the target function is f_i for some $i \in \mathcal{X}$ then there will be a noise at level $0 < \epsilon < 1/2$. That is, if the data is generated by f_i then we will see the label $y = 1$ for x with probability ϵ if $x \neq i$ and with probability $1 - \epsilon$ if $x = i$. For the functions f^0 and f^1 we assume that the noise level is $\epsilon/2$.

Assume we observed the instance x with the label $y = 1$. The posterior Q will be as follows

$$Q\left\{f^1\right\} = \frac{1 - \epsilon/2}{Z}, \;\; Q\left(f^0\right) = \frac{\epsilon/2}{Z}, \;\; Q\left\{f_x\right\} = \frac{1-\epsilon}{Z}, \;\; \forall i \neq x, Q\left\{f_i\right\} = \frac{\epsilon}{Z} \;\; .$$

Therefore, the MAP estimator is f^1. Nevertheless, for every i the probability, under Q, that the label of i is $y = 1$ is at most $\frac{2 - \frac{3}{2}\epsilon}{Z}$ while the probability that the label is $y = 0$ is at least $\frac{\left(N - \frac{1}{2}\right)\epsilon}{Z}$. Hence, if $N \geq \frac{1}{2} + \frac{8}{\epsilon}$, then according to Q the label $y = 0$ is at least 4 times more likely than the label $y = 1$. Thus, the Bayes estimator is the function f^0. Moreover, since the Bayes estimator is in the class $\mathcal{F}$, it is also the median.

As we can see in this case, despite the uniform prior, the MAP estimator is the opposite function to the Bayes estimator and the median. Moreover, if we look at the density at the function f^0,i.e. the density of the Bayes estimator and the median estimator, it turns out to have the lowest density. Therefore, in this case, rather than maximizing the density, we were better of if we were minimizing it.

In the previous example we have shown a case in which the MAP estimator seems not to be performing well. We now turn to discuss a broader issue concerning the different estimators. The estimators can be viewed as statistics of the posterior Q. One question we can ask about such a statistic is how stable is it. That is, by how much do we have to change Q to be able to get an arbitrary value for the statistic?

Assume we have an estimator **Est** which is a function that gets a probability measure Q as an input and returns a function $f : F \mapsto \{\pm 1\}$ as an output. For two probability measures Q and Q' we measure the distance between them using 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 $f \in \mathcal{F}$ we can ask what is the closest Q' to Q for which $\mathbf{Est}\left(Q'\right) = f$:

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

Finally, we define the stability of **Est** at Q to be the distance to the furthest function

$$\mathbf{stability}\left(\mathbf{Est}, Q\right) = \sup_{f \in \mathcal{F}} d\left(\mathbf{Est}, Q, f\right) \;\; .$$

The way to interpret this definition is as follows; if $s = \mathbf{stability}\left(\mathbf{Est}, Q\right)$ 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.

It may be easier to understand this definition using an example. Assume that Q is a non-atomic measure. To simplify the discussion, we assume that it is continuous with respect to Lebesgue measure. Therefore, there is an energy function $E(f)$ associated with Q such that for any measurable set A,

$$Q[A] = \int_A \exp(-E(f))\, df \quad .$$

In this setup the MAP estimator is defined to be $f^* = \arg\min_{f \in F} E(f)$. For every $\hat{f} \in \mathcal{F}$, we define E' such that

$$E'(f) = \begin{cases} E(f^*) & \text{if } f = \hat{f} \\ E\left(\hat{f}\right) & \text{if } f = f^* \\ E(f) & \text{otherwise} \end{cases}$$

We can define Q' such that $Q'[A] = \int_A \exp(-E'(f))\, df$. From the definition of the MAP estimator, it follows that $\hat{f}$ is the MAP estimator with respect to Q'. However, the total variation distance between Q and Q' is zero. Hence, the stability of the MAP estimator is zero. The following theorem lower bounds the stability of the median estimator as a function of its depth

Theorem 16. *Let Q be a posterior over $\mathcal{F}$. Assume that for every $x \in \mathcal{X}$ and $y \in \pm 1$ there exists $f \in \mathcal{F}$ such that $f(x) = y$. Let $p = \inf_{x,y} Q\{f : f(x) = y\}$. Let d be the depth of the median for Q then*

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

Proof. Let $\epsilon > 0$. There exists $(\hat{x}, \hat{y})$ such that $Q\{f : f(\hat{x}) = \hat{y}\} < p + \epsilon$. Furthermore, there exists $\hat{f}$ such that $f(\hat{x}) = \hat{y}$. Let f^* be the median of Q. Let Q' be such that $\hat{f}$ is the median of Q' hence

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

Note that for every f we have that

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

This follows since

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

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

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

Hence

$$\delta(Q, Q') \geq \frac{d - p - \epsilon}{2}$$

and thus

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

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

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

□

This analysis shows again the benefits of maximizing depth. Note that with small modifications, this theorem applies to the Bayesian estimator as well.

Appendix B. The Tukey depth as a special case

In this section we show that the Tukey depth function is a special case of the depth function defined in this work.

Assume $\mathcal{F} = \mathbb{R}^d$ and $\mathcal{X} = \mathbb{R}^{d+1}$. Every instance $x \in \mathcal{X}$ can be denoted as $x = (x_v, x_\theta)$ where $x_v \in \mathbb{R}^d$ and $x_\theta \in \mathbb{R}$. We define for $f \in \mathcal{F}$ and $x \in \mathcal{X}$

$$f(x) = \text{sign}(f \cdot x_v - x_\theta) \quad .$$

where $\text{sign}(0)$ is defined to be 1. For every $w \in \mathbb{R}^d$, and $f \in \mathcal{F}$ we consider the point $x = (w, f \cdot w)$. Note that for every $f' \in \mathcal{F}$ we have that

$$f'(x) = \text{sign}(f' \cdot w - f \cdot w) \quad .$$

Hence, $f'(x) = 1$ if and only if $f' \cdot w \geq f \cdot w$. Thus

$$Q\{f' : w \cdot (f' - f) \geq 0\} = D_Q(f \,|\, x) \quad .$$

Therefore, we have that

$$\begin{aligned} \text{Tukey-Depth}_Q(f) &= \inf_{w \in R^d} Q\{x' : w \cdot (x' - x) \geq 0\} \\ &\geq \inf_{x \in \mathcal{X}} D_Q(f \,|\, x) \\ &= D_Q(f) \quad . \end{aligned}$$

On the other hand, consider $x \in \mathcal{X}$ and fix $f \in \mathcal{F}$. Recall that $f(x) = \text{sign}(f \cdot x_v - x_\theta)$. If $x_\theta > f \cdot x_v$ we have that $f(x) = -1$ and

$$\begin{aligned} D_Q(f \,|\, x) &= Q\{f' : f' \cdot x_v < x_\theta\} \\ &\geq Q\{f' : f' \cdot x_v \leq f \cdot x_v\} \\ &= Q\{f' : (-x_v) \cdot (f' - f) \geq 0\} \quad . \end{aligned}$$

In the same fashion, if $x_\theta < f \cdot x_v$ we have that $f(x) = 1$ and

$$\begin{aligned} D_Q(f \,|\, x) &= Q(f' : f' \cdot x_v \geq x_\theta) \\ &\geq Q(f' : f' \cdot x_v \geq f \cdot x_v) \\ &= Q\{f' : x_v \cdot (f' - f) \geq 0\} \quad . \end{aligned}$$

This is sufficient to show that for this setting, $\text{Tukey-Depth}_Q(f) = D_Q(f)$.