Asymptotic optimality and minimal complexity of classification by random projection

Mireille Boutin, Evzenie Coupkova

August 17, 2021

Abstract

The generalization error of a classifier is related to the complexity of the set of functions among which the classifier is chosen. Roughly speaking, the more complex the family, the greater the potential disparity between the training error and the population error of the classifier. This principle is embodied in layman’s terms by Occam’s razor principle, which suggests favoring low-complexity hypotheses over complex ones. We study a family of low-complexity classifiers consisting of thresholding the one-dimensional feature obtained by projecting the data on a random line after embedding it into a higher dimensional space parametrized by monomials of order up to k. More specifically, the extended data is projected n-times and the best classifier among those n (based on its performance on training data) is chosen. We obtain a bound on the generalization error of these low-complexity classifiers. The bound is less than that of any classifier with a non-trivial VC dimension, and thus less than that of a linear classifier. We also show that, given full knowledge of the class conditional densities, the error of the classifiers would converge to the optimal (Bayes) error as k and n go to infinity; if only a training dataset is given, we show that the classifiers will perfectly classify all the training points as k and n go to infinity.

1 Introduction

Consider a two-class classification problem with real-valued feature vectors, where the dimension of the feature vectors is potentially very high. We seek to use training data to construct a classifier. This paper is concerned with a family of extremely simple classifiers. Specifically, we analyze a classification method consisting of projecting the data on a random line so to obtain one-dimensional data. The one-dimensional data is then classified by thresholding, using the training data to determine the best threshold. This is performed n times so to obtain n different classifiers. The best classifier among those n (based on performance on the training data) is then chosen. This yields an affine classifier, whose decision boundary is a hyperplane in the original high-dimensional space. More generally, the data can first be expanded to a higher dimensional space, by concatenating the initial feature vector with monomials in the original features, before being projected. By considering all monomials up to some order \( k \in \mathbb{N} \), the random projection and thresholding procedure yield a non-affine classifier whose decision boundary is the zero set of a polynomial of order k in the original feature space. We call this classification method thresholding after random projection. It can be viewed as a one-layer Neural Network whose parameters in the first layer are chosen at random and whose activation function is a hard-threshold sign function, with the number of different projections n corresponding to the width of the hidden layer.

Thresholding after random projection on a one-dimensional subspace has previously been used successfully to cluster high-dimensional data \([11, 5]\). Projections onto a one-dimensional subspace have also been used to develop fast approximate algorithms (e.g., \([7]\)). More generally, random projection on a subspace has been used as a pre-processing step to decrease the dimension of high-dimensional data. The Johnson-Lindenstrauss Lemma suggests that one can decrease the space dimension to a much lower one by random projection while closely preserving...
the original pairwise distances of a dataset, with high-probability. Concerning the problem of classification, it has been shown that a dataset featuring two classes separated by a large margin has a high-probability of being well separated by a random linear separator [2]. More generally, certain datasets are likely to be divided into two well-separated subsets after projection on a random line [3].

Considering the simplicity of the thresholding after random projection classification method, Occam’s razor principle suggests that such a classifier should be used for any training dataset that can be well classified after a random projection. The first part of this paper quantifies the reason why, by showing how the simplicity of this classification method is related to a low probability of classification error. Specifically, Theorem 1 provides an upper bound on the likely difference between the training error and the population error, in terms of the size of the training set \( N \) and the number of projections \( n \). This bound is independent of the space dimension, and independent of \( k \), and also lower than that for a non-random linear classifier. It also compares very favorably to the bound for any family of classifiers with a non-trivial VC dimension \( (d_{VC} > 1) \).

A simple classification method is of little use if it has a poor classification performance. In the second part of this paper, we show that, even though the thresholding after random projection classification method is extremely simple, its accuracy is asymptotically optimal. Two cases are explored. First the case where one is given full knowledge of the class conditional probability distributions, for which we show that, with \( k \) and \( n \) large enough, a classifier that is arbitrarily close to the (optimal) Bayes classifier is likely to be obtained. Second, the case where one is given a training dataset, for which we show that, for large enough \( k \) and \( n \), a perfect classification is likely to be obtained.

# 2 Minimal complexity

In this section we focus on the generalization error of the method of thresholding after random projection. Our main result is Theorem 1 which provides an upper bound on the absolute value of the difference between the training error and the population error. The bound depends on the number \( n \) of projections but neither on \( k \) nor the original space dimension. It is obtained by splitting the set of functions from which we choose a classifier into independent subsets. We also utilize the fact that the classification is carried out in one dimension, which greatly reduces the size of the set of all possible partitions. As a result, we get a much tighter bound on the generalization error than the one given by the VC dimension of the set of considered functions.

The task of constructing a classifier can be viewed as choosing a hypothesis from a set \( F \) of hypotheses, based on a training dataset \( \{ (x_i, y_i) \}_{i=1}^N \), where \( x_i \) contains the real-valued features of point \( i \), \( y_i \) is the class of point \( i \), and \( N \) is the number of points. The set \( F \) needs to be rich enough to approximate the optimal solution well, but not too rich, with respect to \( N \), or else the generalization error of the chosen classifier may turn out to be much different from the training error.

Often, the hypothesis set is a set of parametrized functions \( F = \{ f_\theta(x), \ \theta \in \Theta \} \), and one chooses the optimal parameter \( \hat{\theta} \) by minimizing the empirical error on the collected data, denoted by \( E_{train} \):

\[
\hat{f}(x) = f_\hat{\theta}(x),
\]

\[
\hat{\theta} = \arg \min_{\theta \in \Theta} E_{train}(f_\theta),
\]
where
\[ E_{\text{train}}(f_{\theta}) = \frac{1}{N} \sum_{i=1}^{N} (f_{\theta}(x_i) - y_i)^2. \]

When the label \( y \) is either 0 or 1 and the value of the function \( f(x, \theta) \) is also 0 or 1, \( E_{\text{train}} \) is the 0-1 loss:
\[ E_{\text{train}}(f_{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(f_{\theta}(x_i) \neq y_i). \]

The ultimate goal, however, is to find a classification that will be accurate on a new data set. In other words, one would like to accurately predict the class of data points that do not belong to the training set. Ideally, the error of the classifier on a new data set would be likely to be close to the training error, that is to say the error that was minimized during the process of choosing the classifier. But in actuality, it might likely be much larger. If we assume that our data are generated by a probability distribution \( \rho_{X,Y}(x,y) \) on a certain space \( \mathcal{E} \times \{0,1\} \), we can compute the overall population error of the classifier, that is to say the error that the classifier would make if it were used to classify all of the points in \( \mathcal{E} \times \{0,1\} \):
\[ E_{\text{popul}}(f) = \int_{\mathcal{E} \times \{0,1\}} \rho_{X,Y}(x,y) (f(x) - y)^2 \, d(x,y) = \int_{\mathcal{E} \times \{0,1\}} \rho_{X,Y}(x,y) \mathbb{1}(f(x) \neq y) \, d(x,y). \]

We want to guarantee that applying the classifier on a new data set will likely have an error similar to the training error. More specifically, we want that, for any given level of tolerance \( \delta \in (0,1) \), with probability at least \( 1 - \delta \), the difference between the population error and the training error is less than some small generalization term \( E_F \):
\[ |E_{\text{popul}}(f) - E_{\text{train}}(f)| \leq E_F. \]

The generalization term \( E_F \) may be expressed as a function that depends on the number of points in the training set \( N \), the tolerance level \( \delta \) and the richness of the family of functions \( \mathcal{F} \). A helpful tool to make such a bound is Hoeffding’s inequality [10], which stated that if \( X_1, \ldots, X_N \) are independent random variables whose values are bounded by the interval \([0,1]\), and \( \bar{X} = \frac{1}{N} \sum_{i=1}^{N} X_i \), then for any \( \epsilon \geq 0 \) we have
\[ P \left( |\bar{X} - E\bar{X}| \geq \epsilon \right) \leq 2e^{-2N\epsilon^2}. \]

If we consider random variables \( X_i = \mathbb{1}(f(x_i) \neq y_i) \), then \( X_i \) is bounded by \([0,1]\), because it is either 0 or 1 and \( X_i \) is independent of \( X_j \) as long as \( i \neq j \). Therefore the conditions of Hoeffding’s inequality are satisfied and since \( \bar{X} = E_{\text{train}} \) and \( E\bar{X} = E_{\text{popul}} \) we get the following result:
\[ P(\{|E_{\text{train}}(f) - E_{\text{popul}}(f)| > \epsilon\} \leq 2e^{-2N\epsilon^2}. \]

If \( \mathcal{F} \) contains a finite number \( M \) of functions, then we can use Hoeffding’s inequality for each hypothesis \( f_i \in \{f_1, \ldots, f_M\} \) and bound the supremum of all the deviations with a union bound:
\[ P \left( \sup_{i \in \{1, \ldots, M\}} |E_{\text{train}}(f_i) - E_{\text{popul}}(f_i)| > \epsilon \right) \leq P \left( \bigcup_{i=1}^{M} |E_{\text{train}}(f_i) - E_{\text{popul}}(f_i)| > \epsilon \right) \leq \prod_{i=1}^{M} P(\{|E_{\text{train}}(f_i) - E_{\text{popul}}(f_i)| > \epsilon\} \leq M \times 2e^{-2N\epsilon^2}. \]
If we denote the resulting bound by $\delta$, we can express $\varepsilon$ as

$$\varepsilon = \sqrt{\frac{1}{2N} \ln \left( \frac{2M}{\delta} \right)}.$$

As a consequence, with probability at least $1 - \delta$:

$$\sup_{i \in \{1, \ldots, M\}} |E_{\text{popul}}(f_i) - E_{\text{train}}(f_i)| \leq \sqrt{\frac{1}{2N} \ln \left( \frac{2M}{\delta} \right)}.$$

(1)

One can also obtain a bound for the case where $\mathcal{F}$ contains an infinite number of functions. In particular, when the value of $y_i$ can be either 1 or 0 (binary classifications), then the richness of the family $\mathcal{F}$ can be expressed using its VC-dimension [9], which is an integer equal to the largest number of points for which there exists an arrangement that can be shattered by the class of functions $\mathcal{F}$. In other words, if $\mathcal{F}$ has VC-dimension $d_{VC}$, then there exists a geometrical arrangement of $d_{VC}$ points, such that for any assignment of classes to these points, there exists a function from $\mathcal{F}$ that classifies each point correctly.

As an example, let us obtain the VC-dimension of the class of affine functions applied to classify points in $\mathbb{R}^2$: $\mathcal{F} = \{f_{w,b}(x) = \text{sign}(x \cdot w + b), w \in \mathbb{R}^2, b \in \mathbb{R}\}$. Consider three points in the plane that are not aligned. Then, for any class assignment, there exists a line that separates the points from different classes. However, if we take four points in the plane, then for any geometrical arrangement of points, there exists a labelling such that no line can separate points that belong to different classes. Therefore the VC-dimension of $\mathcal{F}$ is equal to three. More generally, the VC-dimension of the class of affine functions applied to classify points in $\mathbb{R}^d$ is $d + 1$ [1].

For a family of functions $\mathcal{F}$ with VC-dimension $d_{VC}$, the inequality (1) can be replaced by the following one:

$$\sup_{f \in \mathcal{F}} |E_{\text{popul}}(f) - E_{\text{train}}(f)| \leq \frac{8}{N} \ln \left( \frac{4 (2N + 1)^{d_{VC}}}{\delta} \right).$$

(2)

If the VC dimension is finite, inequality (2) gives us a guarantee that our training error can be arbitrarily close to the population error given enough sample points, since:

$$\lim_{N \to \infty} \sqrt{\frac{8}{N} \ln \left( \frac{4 (2N + 1)^{d_{VC}}}{\delta} \right)} = 0.$$

For the method of thresholding after random projection, the size of the hypothesis set is infinite, because in each iteration of the algorithm we consider all affine functions. So if the data to classify come from a $d$-dimensional euclidean space $\mathbb{R}^d$, then $d_{VC} = d + 1$. Plugging this value of the VC dimension into (2) gives us the following generalization term $E_{\mathcal{F}}$:

$$E_{\mathcal{F}} = \frac{8}{N} \ln \left( \frac{4 (2N + 1)^{d+1}}{\delta} \right).$$

(3)

A more careful analysis of the hypothesis set of the thresholding after random projection gives us a better bound than the one in (3). We use specific properties of the method of thresholding after random projection to bound the generalization error very tightly. First property is the fact
that the hypothesis set can be split into $n$ independent sets, each one gives a certain number of possible classification outcomes, but they do not influence each other. The second property is that the classification is carried out after projection into a one dimensional space. That leads to very few classification options on a finite data set and as a result to a small generalization bound.

**Theorem 1.** For the thresholding after random projection classification method (applied to data extended to monomials up to order $k$ and choosing the best among $n$ random projections), we have

$$\sup_{f \in \mathcal{F}} |E_{\text{popul}}(f) - E_{\text{train}}(f)| \leq \sqrt{\frac{8}{N} \ln \left( \frac{16 n N}{\delta} \right)}$$

with probability at least $1 - \delta$.

**Proof.** The following proof is an adaptation of the proof from [1] and [9]. Let us specify, that the training error of a classifier $f$ depends on the dataset used $\mathcal{D}$ by denoting the error by $E_{\text{train}}^\mathcal{D}(f)$. Then we need to prove the following:

$$P \left( \sup_{f \in \mathcal{F}} |E_{\text{popul}}(f) - E_{\text{train}}^\mathcal{D}(f)| \leq \sqrt{\frac{8}{N} \ln \left( \frac{16 n N}{\delta} \right)} \right) \geq 1 - \delta$$

In order to obtain this bound we first need to replace the difference between the population error and training error with the difference between training errors on two independent data sets that follow the same distribution. In order to do this, we use the following lemma.

**Lemma 1.** Consider a data set $\mathcal{D}'$ that is independent from the data set $\mathcal{D}$, but has the same number of points and follows the same distribution as $\mathcal{D}$. Denote the error of the classifier $f \in \mathcal{F}$ on this data set $\mathcal{D}'$ by $E_{\text{train}}^{\mathcal{D}'}(f)$. Then

$$P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^{\mathcal{D}'}(f) - E_{\text{train}}^{\mathcal{D}}(f) - E_{\text{popul}}(f)| \geq \frac{\varepsilon}{2} \right) \leq \frac{1}{\left(1 - 2e^{-\frac{\varepsilon^2 N}{2}} \right)} P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^{\mathcal{D}'}(f) - E_{\text{train}}^{\mathcal{D}}(f)| \geq \frac{\varepsilon}{2} \right),$$

where the probability on the right hand side is over the data sets $\mathcal{D}$ and $\mathcal{D}'$ jointly.

**Proof.**

$$P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^{\mathcal{D}'}(f) - E_{\text{train}}^{\mathcal{D}}(f)| > \frac{\varepsilon}{2} \right) \geq P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^{\mathcal{D}'}(f) - E_{\text{train}}^{\mathcal{D}}(f)| > \frac{\varepsilon}{2} \text{ and } \sup_{f \in \mathcal{F}} |E_{\text{train}}^{\mathcal{D}'}(f) - E_{\text{popul}}(f)| > \varepsilon \right) =$$

$$= P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^{\mathcal{D}'}(f) - E_{\text{train}}^{\mathcal{D}}(f)| > \varepsilon \right) \times$$

$$\times P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^{\mathcal{D}'}(f) - E_{\text{popul}}(f)| > \varepsilon \right)$$
Here we are conditioning on a set of data sets $D$ that happen with non-zero probability and have the property that:

$$\sup_{f \in \mathcal{F}} |E_{\text{train}}^D(f) - E_{\text{popul}}(f)| > \varepsilon.$$ 

Let us fix a function $f^*$ for which $|E_{\text{train}}^D(f^*) - E_{\text{popul}}(f^*)| > \varepsilon$. Then

$$P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^D(f) - E_{\text{train}}^{D'}(f)| > \frac{\varepsilon}{2} \sup_{f \in \mathcal{F}} |E_{\text{train}}^D(f) - E_{\text{popul}}(f)| > \varepsilon \right) \geq$$

$$\geq P \left( |E_{\text{train}}^D(f^*) - E_{\text{train}}^{D'}(f^*)| > \frac{\varepsilon}{2} \sup_{f \in \mathcal{F}} |E_{\text{train}}^D(f) - E_{\text{popul}}(f)| > \varepsilon \right).$$

Since

$$|E_{\text{train}}^D(f^*) - E_{\text{popul}}(f^*)| \leq |E_{\text{train}}^D(f^*) - E_{\text{train}}^{D'}(f^*)| + |E_{\text{train}}^{D'}(f^*) - E_{\text{popul}}(f^*)|$$

assuming that

$$|E_{\text{train}}^{D'}(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2}$$

$$|E_{\text{train}}^D(f^*) - E_{\text{popul}}(f^*)| > \varepsilon,$$

implies

$$|E_{\text{train}}^D(f^*) - E_{\text{train}}^{D'}(f^*)| > \frac{\varepsilon}{2},$$

therefore

$$P \left( |E_{\text{train}}^D(f^*) - E_{\text{train}}^{D'}(f^*)| > \frac{\varepsilon}{2} \sup_{f \in \mathcal{F}} |E_{\text{train}}^D(f) - E_{\text{popul}}(f)| > \varepsilon \right) \geq$$

$$\geq P \left( |E_{\text{train}}^{D'}(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2} \sup_{f \in \mathcal{F}} |E_{\text{train}}^D(f) - E_{\text{popul}}(f)| > \varepsilon \right).$$

Hoeffding’s inequality gives us that:

$$P \left( |E_{\text{train}}^{D'}(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2} \right) \geq 1 - 2e^{-\frac{\varepsilon^2}{2}}.$$ 

Hoeffding inequality applies to any dataset $D'$ and $f^*$ is independent of $D'$, therefore it also applies to any weighted average of $P(|E_{\text{train}}^{D'}(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2})$ based on $f^*$. Since $f^*$ depends on a particular $D$, we take the weighted average over all $D$ in the event:

$$A = \{D, \text{ such that } |D| = N \text{ and } \sup_{f \in \mathcal{F}} |E_{\text{train}}^D(f) - E_{\text{popul}}(f)| > \varepsilon\},$$

on which we are conditioning, where the weight comes from the probability of each $D$. Since
the bound holds for every \( \mathcal{D} \), it holds for the weighted average.

\[
P \left( |E_{\text{train}}^\mathcal{D}'(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2} \right) = \sum_{\mathcal{D} \in \mathcal{A}} P \left( |E_{\text{train}}^\mathcal{D}'(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2} | \mathcal{D}, \mathcal{A} \right) = \\
= \sum_{\mathcal{D} \in \mathcal{A}} P \left( |E_{\text{train}}^\mathcal{D}'(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2} | \mathcal{D}, \mathcal{A} \right) P(\mathcal{D}, \mathcal{A}) = \\
= \sum_{\mathcal{D} \in \mathcal{A}} P \left( |E_{\text{train}}^\mathcal{D}'(f^*) - E_{\text{popul}}(f^*)| \leq \frac{\varepsilon}{2} | \mathcal{D} \right) P(\mathcal{D}, \mathcal{A}) \geq \\
\geq \left(1 - 2e^{-\frac{1}{2}\varepsilon^2 N}\right) \sum_{\mathcal{D} \in \mathcal{A}} P(\mathcal{D}, \mathcal{A}) = \left(1 - 2e^{-\frac{1}{2}\varepsilon^2 N}\right).
\]

We can assume that \(1 - 2e^{-1/2\varepsilon^2 N} \geq 1/2\), because if it is not, then the bound in inequality (4) is satisfied automatically. Inequality (4) is equivalent to the following:

\[
P \left( \sup_{f \in \mathcal{F}} |E_{\text{popul}}(f) - E_{\text{train}}(f)| > \varepsilon \right) \leq \frac{8}{N} \ln \left( \frac{16 N}{\delta} \right) \leq \delta.
\]

If we denote the expression \(\sqrt{\frac{8}{N} \ln \frac{16 N}{\delta}}\) by \(\varepsilon\), we get:

\[
P \left( \sup_{f \in \mathcal{F}} |E_{\text{popul}}(f) - E_{\text{train}}(f)| > \varepsilon \right) \leq 16nN \exp \left( -\frac{1}{8}\varepsilon^2 N \right), \tag{5}
\]

and if \(\exp \left( -\varepsilon^2 N/2 \right) > 1/4\), then \(16nN \exp \left( -\frac{1}{8}\varepsilon^2 N \right) > 1\), so the bound in (5) holds trivially. Therefore in the context of Theorem 1, the bound from Lemma 1 implies a simple inequality:

\[
P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^\mathcal{D}(f) - E_{\text{popul}}(f)| > \varepsilon \right) \leq 2P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^\mathcal{D}'(f) - E_{\text{train}}^\mathcal{D}(f)| > \frac{\varepsilon}{2} \right).
\]

The next step of the proof is to show that the probability that the training error on two independent, equisized data sets that follow the same distribution is vastly different, can be bound by something small. Instead of considering two data sets \(\mathcal{D}\) and \(\mathcal{D}'\) both with \(N\) identically distributed and independent samples, let us consider a set \(\mathcal{S}\) of size \(2N\), which we split at random into two sets \(\mathcal{D}\) and \(\mathcal{D}'\) of equal size. Using the set \(\mathcal{S}\), we can express the probability that the training errors are different in the following way:

\[
P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^\mathcal{D}(f) - E_{\text{train}}^\mathcal{D}'(f)| > \frac{\varepsilon}{2} \right) = \sum_{\mathcal{S}} P(\mathcal{S}) \times P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^\mathcal{D}(f) - E_{\text{train}}^\mathcal{D}'(f)| > \frac{\varepsilon}{2} \right) \\
\leq \sup_{\mathcal{S}} P \left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^\mathcal{D}(f) - E_{\text{train}}^\mathcal{D}'(f)| > \frac{\varepsilon}{2} \right),
\]

where the probability on the left hand side is over \(\mathcal{D}\) and \(\mathcal{D}'\) jointly, while the the probability on the right hand side is over all of the random partitions of \(\mathcal{S}\) into \(\mathcal{D}\) and \(\mathcal{D}'\).

The next step contains the main difference between the usual proof of creating a bound using the VC dimension and our approach. We are going to use the fact that the hypothesis
set of the method of thresholding after random projection with \( n \) iterations can be split into \( n \) independent subsets:

\[
\mathcal{F} = \bigcup_{i=1}^{n} \mathcal{F}_i
\]

and each subset represents a very limited number of classifying options due to the fact that classification is obtained in one-dimensional space, to which the original data are projected. We need to count, how many different outcomes are possible if we use the hypothesis set \( \mathcal{F} \) on the dataset of size \( 2N \). By different outcomes we mean different dichotomies: ordered sequences of zeros and ones that correspond to the classes chosen for each data point. When the data are projected in some direction, they are arranged in a certain order into a line. An order is a permutation of the original order of \( 2N \) points. Then a threshold is chosen in one dimension that separates the points into two classes: to the left of the threshold and to the right of the threshold. If \( N \) was equal to 2 we could get the following 8 dichotomies in the projected space:

- (0 0 0 0)
- (1 0 0 0)
- (1 1 0 0)
- (1 1 1 0)
- (1 1 1 1)
- (0 1 1 1)
- (0 0 1 1)
- (0 0 0 1)

In general, each \( \mathcal{F}_i \) gives us \( 4N \) different possible dichotomies on \( 2N \) points. That corresponds to projection in one particular random direction. If we choose a different projection direction, it can give us different dichotomies, because the permutation of the points might be different. At most, we can get \( 4N \) more dichotomies with each projection. Therefore the number of different outcomes is smaller than \( 4nN \). For each dichotomy we can choose a representative \( f_i \in \mathcal{F} \), this way on any finite number of sample points, we have a finite number of outcomes even for an infinitely large family of functions \( \mathcal{F} \). Therefore, we can estimate the probability of a large deviation between two sample sets. If the number of all possible dichotomies is \( M \), we know that \( M \leq 4nN \) and we can write:

\[
P\left( \sup_{f \in \mathcal{F}} |E_{\text{train}}^\mathcal{D}(f) - E_{\text{train}}^\mathcal{D}'(f)| > \frac{\varepsilon}{2} \left| S \right| \right) = \sup_{f \in \{f_1, \ldots, f_M\}} P\left( \sup_{f \in \{f_1, \ldots, f_M\}} |E_{\text{train}}^\mathcal{D}(f) - E_{\text{train}}^\mathcal{D}'(f)| > \frac{\varepsilon}{2} \left| S \right| \right) \leq \sum_{i=1}^{M} P\left( |E_{\text{train}}^\mathcal{D}(f_i) - E_{\text{train}}^\mathcal{D}'(f_i)| > \frac{\varepsilon}{2} \left| S \right| \right) \leq M \times \sup_{S} \sup_{f \in \mathcal{F}} P\left( |E_{\text{train}}^\mathcal{D}(f) - E_{\text{train}}^\mathcal{D}'(f)| > \frac{\varepsilon}{2} \left| S \right| \right) \leq 4nN \times \sup_{S} \sup_{f \in \mathcal{F}} P\left( |E_{\text{train}}^\mathcal{D}(f) - E_{\text{train}}^\mathcal{D}'(f)| > \frac{\varepsilon}{2} \left| S \right| \right).
\]

The probabilities are over all of the random partitions of \( S \) into \( \mathcal{D} \) and \( \mathcal{D}' \).
Lemma 2. For any classifier \( f \) and any sample set \( S \) of size \( 2N \):

\[
P\left( \left| E_{\text{train}}^D(f) - E_{\text{popul}}^D(f) \right| > \frac{\varepsilon}{2} \right| S \right) \leq 2e^{-\frac{1}{8} \varepsilon^2 N}
\]

Proof. The lemma follows from using the result from [6] about sampling without replacement. \( \square \)

Combining the inequalities from the proof we get the following estimate:

\[
P \left( \sup_{f \in F} \left| E_{\text{train}}^D(f) - E_{\text{popul}}^D(f) \right| > \varepsilon \right) \leq 2P \left( \sup_{f \in F} \left| E_{\text{train}}^D(f) - E_{\text{train}}^{D'}(f) \right| \geq \varepsilon/2 \right) \leq 2 \sup_{S \in \mathcal{S}} P \left( \sup_{f \in F} \left| E_{\text{train}}^D(f) - E_{\text{train}}^{D'}(f) \right| \geq \varepsilon/2 \right| S \right) \leq 2 \times 4nN \sup_{S \in \mathcal{S}} P \left( \left| E_{\text{train}}^D(f) - E_{\text{train}}^{D'}(f) \right| \geq \varepsilon/2 \right| S \right) \leq 2 \times 4nN \times 2e^{-\frac{1}{8} \varepsilon^2 N} \leq 16nN e^{-\frac{1}{8} \varepsilon^2 N}
\]

Since we want the probability in (6) to be smaller or equal than \( \delta \), we get

\[
\delta \geq 16 n N e^{-\frac{1}{8} \varepsilon^2 N}
\]

leading to

\[
E_{\mathcal{F}} = \varepsilon \leq \sqrt{\frac{8}{N} \ln \frac{16nN}{\delta}}.
\]

\( \square \)

Figure 1 illustrates the relationship between the generalization term \( E_{\mathcal{F}} \) and the number of projections \( n \). As one can see from the graphs, the growth in generalization error is quite modest after a short initial window of fast increase.

Figure 1: Generalization term \( E_{\mathcal{F}} \) when \( N = 10000, \delta = 0.1 \) on the left and \( N = 100000 \) and \( \delta = 0.05 \) on the right. The number of projections \( n \) is varying between 1 and 10000.
2.1 Comparison to classifier with $d_{VC} > 1$

We now compare the generalization error term of the the method of thresholding after random projection with that of a classifier with a non-trivial VC dimension. Specifically, we consider the limit of the ratio of their generalization error terms when the number of data points in the training set converges to infinity. The result is in the following theorem.

**Theorem 2.** For large enough training sets, the generalization error term $E_F$ of the thresholding after random projection classification method is smaller than the generalization error term estimated using VC dimension for any algorithm with a non-trivial VC dimension ($d_{VC} > 1$). More specifically, when the number of samples goes to infinity, the ratio of the generalization error terms goes to $\sqrt{\frac{1}{d_{VC}}}$.

**Proof.** The ratio of the bounds for the generalization error is

\[
R(N) = \sqrt{\frac{8 N \ln \left( \frac{16nN}{\delta} \right)}{8 N \ln \left( \frac{4(2N + 1)^{d_{VC}}}{\delta} \right)}} = \sqrt{\frac{\ln(2N) + \ln \left( \frac{8n}{\delta} \right)}{\ln(2N + 1)^{d_{VC}} + \ln \left( \frac{4}{\delta} \right)}}
\]

for appropriate constants $c_1$ and $c_2$ which do not depend on $N$. Therefore

\[
\lim_{N \to \infty} R(N) = \sqrt{\frac{1}{d_{VC}}}
\]

Note that inequality (2) can be replaced by a different bound, which is tighter in cases when the VC dimension $d_{VC}$ is larger than 2. This tighter bound is based on a tighter bound on the number of different dichotomies that can be produced by a classifier with a given VC dimension. The number of dichotomies can be bounded by $\sum_{i=0}^{d_{VC}} \binom{N}{i}$. In inequality (2) we used that the sum of first $d_{VC}$ terms of binomial coefficients can be bound by $(N + 1)^{d_{VC}}$, in the following inequality we use that it can be also bound by $\left( \frac{4e}{d_{VC}} \right)^{d_{VC}}$:

\[
\sup_{f \in F} |E_{popul}(f) - E_{train}(f)| \leq \sqrt{\frac{8}{N} \ln \left( \frac{4e}{d_{VC}} \right)^{d_{VC}} \frac{2eN}{d_{VC}} \frac{1}{\delta}} = \sqrt{\frac{8}{N} \ln \left( \frac{4(2N)^{d_{VC}} (\frac{e}{d_{VC}})^{d_{VC}}}{\delta} \right)}.
\]

This bound does not change the result of the Theorem. Indeed the ratio of the generalization
error terms converges to \( \sqrt{\frac{1}{d_{VC}}} \) with this bound as well, since

\[
R(N) = \sqrt{\frac{8}{N} \ln \left( \frac{16nN}{\delta} \right)} = \sqrt{\frac{\ln(2N) + \ln \left( \frac{8n}{\delta} \right)}{d_{VC} \ln(2N) + d_{VC} \ln \left( \frac{e}{d_{VC}} \right) + \ln \left( \frac{4}{\delta} \right)}} = \sqrt{\frac{\ln(2N) + \tilde{c}_1}{d_{VC} \ln(2N) + \tilde{c}_2}},
\]

for appropriate constants \( \tilde{c}_1 \) and \( \tilde{c}_2 \), which do not depend on \( N \). Therefore

\[
\lim_{N \to \infty} R(N) = \sqrt{\frac{1}{d_{VC}}},
\]

### 2.2 Comparison with linear separation

If the feature data is \( d \)-dimensional, one can build a classifier by picking the best hyperplane following some goodness-of-fit criterion. As stated earlier, the VC dimension of this classification method is \( d + 1 \). This is one of the simplest classification methods. Yet, for the method of thresholding after random projection, we have provided a bound on the generalization error that is smaller than that for a classification method with \( d_{VC} = d + 1 \) for any \( d \geq 1 \).

To illustrate the difference, let us compare the generalization term for the method of thresholding after random projection given by our estimate and the generalization term given by an estimate that uses the VC dimension \( d_{VC} = d + 1 \). We assume that \( k = 1 \) and fix \( n \) to reach the optimal training error as discussed in Section 3 as per Formula [13]. However, since the bound on \( n \) is overly conservative, the optimal error might be achieved with a smaller \( n \).

| \( d \) | \( n \) | \( E_F(n) \) | \( E_F(d_{VC} = d + 1) \) |
|---|---|---|---|
| 2 | 16 | 0.117 | 0.163 |
| 3 | 62 | 0.121 | 0.183 |
| 4 | 258 | 0.126 | 0.200 |
| 5 | 1171 | 0.131 | 0.216 |
| 6 | 5747 | 0.135 | 0.230 |
| 7 | 30210 | 0.140 | 0.244 |
| 8 | 168778 | 0.145 | 0.256 |
| 9 | 995671 | 0.150 | 0.268 |
| 10 | 6169709 | 0.155 | 0.279 |

Table 1: Comparing the generalization terms: the generalization term \( E_F \) for the method of thresholding after random projection and the generalization term \( E_F(d_{VC} = d + 1) \) for linear classification. We use \( k = 1 \) \( \delta = 0.1 \) and \( N = 10000 \). The last column is computed using a tighter bound for \( E_F \) from [7].

As we can see from Table 1, the generalization advantage of the method of thresholding after random projection is present even for 2-dimensional data. For data in 10 dimensions, the estimate for the generalization error of the linear separation algorithm is almost twice as large as the estimate for the generalization error of the method of thresholding after random projection.
It is possible that, for a specific linear classification algorithm, there is a better way of estimating the generalization error than by merely using its VC dimension; a careful analysis of possible dichotomies might reveal better generalization properties.

3 Asymptotic optimality

In this section we show that the thresholding after random projection classification method is asymptotically optimal. We first approach this topic from a theoretical perspective, assuming full knowledge of the distributions that generate the data and optimal separation given by Bayes decision rule. We start with a case where the optimal decision function is linear. In this case using the thresholding after random projection classification method on the original data enough times (i.e., \(n\) large enough) abates the reducible error. If the optimal decision function is not linear, we show that there exists a monomial degree \(k\) for which we can find \(n\) such that the reducible error is as small as desired with probability as large as desired (Corollary 1).

Second, we look at the topic from a more practical standpoint and consider the training error on a finite data set. Our main result, Theorem 5, states that for any training set, there exists a \(k\) such that the error of the thresholding after random projection classification method on the training set converges to 0 with \(n\) growing to infinity.

Let us assume that the data to classify come from two different classes \(\omega_1\) and \(\omega_2\) that follow probability distributions with densities \(\rho_1(x) = \rho(x|\omega_1)\) and \(\rho_2(x) = \rho(x|\omega_2)\) respectively. Each class has a certain probability of occurring \(P(\omega_1)\) and \(P(\omega_2)\), called prior. The mixture of the densities can be expressed in the following way:

\[
\rho(x) = \rho_1(x)P(\omega_1) + \rho_2(x)P(\omega_2),
\]

where we denote the support of the function \(\rho(x)\) by \(B\). If the supports of the densities \(\rho_1(x)\) and \(\rho_2(x)\) have a non-empty intersection, there does not exist an algorithm that would certainly classify each point correctly. That is due to the fact that a classifier assigns one class to each point \(x\) in \(B\), but if \(x\) belongs to the support of \(\rho_1(x)\) as well as to the support of \(\rho_2(x)\) then the classification at this point will be wrong with non-zero probability for any classifier. The optimal way to choose a class for this point is to choose the one that maximizes \(\rho_i(x)P(\omega_i)\). The probability of an overall error for the classifier is minimized by choosing a class for every point according to this decision. In case it is well defined, this minimal error is called Bayes error and it can be expressed in the following way:

\[
\int_B \min\{\rho_1(x)P(\omega_1), \rho_2(x)P(\omega_2)\}dx.
\]

This error can be considered a goal for any classifier, because it is impossible to reach a smaller population error. It is important to keep in mind that it is still possible to achieve a smaller or even zero error on a training set, because it is just a subset of all possible pairs of points and classes that can be generated by the underlying distribution. When we consider the population error of a classifier, it is usually larger that the Bayes error. The difference between the error of a classifier and Bayes error is called the reducible error. Because of the random choice of projection directions, the reducible error of the method of thresholding after random projection is a random variable; we will prove that it converges to 0 in probability, as \(k\) and \(n\) grow to infinity, given full knowledge of the class distributions. Given a fixed training set, the training error of the thresholding after random projection method is also a random variable; we will show that it converges to zero in probability, as \(k\) and \(n\) grow to infinity.
Without much loss of generality, we assume that the points to classify lie inside a compact subset of $\mathbb{R}^d$, that is to say bounded and closed. For many distributions (e.g. mixture of Gaussians) this assumption does not apply. However, for distributions that have a finite first moment $\mu$, it is possible to identify a compact region so that the decision that we make outside of that region has an arbitrarily small influence on the overall error. Indeed, consider a closed and bounded set $B_\varepsilon$ defined as follows:

$$B_\varepsilon = \{ x \in \mathbb{R}^d : \| x - \mu \|_2 \leq C_\varepsilon \},$$

where $C_\varepsilon$ is such that

$$\int_{\| x - \mu \|_2 \leq C_\varepsilon} \rho(x) dx \geq 1 - \varepsilon/2.$$

Then if we manage to push the error of the algorithm on $B_\varepsilon$ under $\varepsilon/2$ (by finding appropriate monomial degree $k$ and number of projections $n$), then the error of the classifier on the whole space $\mathbb{R}^d$ will be smaller than $\varepsilon$ regardless of the decision made outside of $B_\varepsilon$. Therefore assuming that the support of the distribution is compact has an arbitrarily small effect on the generalization error.

For the method of thresholding after random projection, we generate directions for projections at random following the uniform distribution on a unit sphere. This distribution is such that every open set on the sphere has a non-zero probability: if $u$ is an open set in $S^{d-1}$ and $f_r$ denotes the distribution of the random vectors that we choose from, then:

$$p_u = \int_u f_r(r) dr > 0.$$

**Theorem 3.** If the optimal decision function given by Bayes rule is linear, then for $k = 1$ (i.e., using the original feature space coordinates without extension) the reducible error of the method of thresholding after random projection converges to 0 in probability as $n$ goes to infinity.

**Proof.** Let us denote the unit normal vector to the optimal separation hyperplane given by Bayes’ rule as $N$. If the random vector drawn $r$ is equal to $N$ then the method of thresholding after random projection will be optimal and the error will be equal to Bayes error. Let $u$ be an open neighborhood of $N$ on the unit hypersphere and let $p_u$ be the probability that $r \in u$. By assumption $p_u \neq 0$. Consider $n$ independent random samples of the vector $\{r_i\}_{i=1}^n$. The probability that all of these vectors lie outside of $u$ is equal to $(1 - p_u)^n$, and thus converges to 0 as $n$ goes to infinity. Therefore, with $n$ large enough, we can get as close to the optimal separation hyperplane as needed. The vector chosen for classification which we denote by $\hat{r}_n$ converges to the optimal one in probability:

$$\forall \varepsilon > 0 : \lim_{n \to \infty} P (\| \hat{r}_n - N \|_2 > \varepsilon) = \lim_{n \to \infty} (1 - p_u(\varepsilon))^n = 0. \quad (8)$$

Let us fix $\varepsilon > 0$ and find $n$ such that the reducible error of the method of thresholding after random projection is smaller than $\varepsilon$. The reducible error can be expressed as the following integral:

$$\frac{1}{2} \int_{A_n} \max\{\rho_1(x)P(\omega_1), \rho_2(x)P(\omega_2)\} - \min\{\rho_1(x)P(\omega_1), \rho_2(x)P(\omega_2)\} dx,$$
where $A_n$ is the set of points where the classifier does not make an optimal decision given by Bayes rule. The region $A_n$ is an intersection of two half-spaces (given by the separation hyperplanes). The area of $A_n$ can be characterized by the dihedral angle between the two hyperplanes, let us denote it by $\alpha_n$. When $\hat{r}_n$ approaches $N$, the angle between the hyperplanes approaches 0: $\alpha_n \to 0$ in the same sense as $\hat{r}_n$ converges to $N$.

Let us bound the error on the compact set $B$:

$$\frac{1}{2} \int_{A_n \cap B} \max_i \{\rho_i(x)P(\omega_i)\} - \min_i \{\rho_i(x)P(\omega_i)\} \, dx \leq M \int_{A_n \cap B} \, dx,$$

where $M$ is a finite upper bound for continuous function $\max \{\rho_1 P(\omega_1), \rho_2 P(\omega_2)\}$ on a compact set: $\max \{\rho_1 P(\omega_1), \rho_2 P(\omega_2)\} \leq M$. The integral $\int_{A_n \cap B} \, dx$ is the volume of an intersection of the set $A_n$ with the set $B$. We can provide an upper bound for this volume by considering a volume of two hyperspherical sectors with cotolatitude angle equal to $\alpha_n$ in a hypersphere with radius $B = \text{diam}(B)$. Due to [8] the hypersector has the following volume:

$$V_{\text{sector}}^d (B) = \frac{\pi^{d/2}}{2\Gamma(\frac{d}{2} + 1)} B^d f_{\sin^2(\alpha_n)} \left( \frac{d - 1}{2}, \frac{1}{2} \right),$$

where $I$ is a regularized incomplete beta function:

$$f_{\sin^2(\alpha_n)} \left( \frac{d - 1}{2}, \frac{1}{2} \right) = \frac{1}{B \left( \frac{d}{2} - 1, \frac{1}{2} \right)} \int_0^{\sin^2(\alpha_n)} u^{\frac{d-3}{2}} (1-u)^{-\frac{1}{2}} \, du,$$

which converges to 0 as $\alpha_n$ converges to 0. The volume of $A_n \cap B$ also converges to 0, since:

$$\int_{A_n \cap B} \, dx \leq 2V_{\text{sector}}^d (B).$$

Since $\hat{r}_n$ approaches $N$ in probability, the dihedral angle $\alpha_n$ approaches 0 in probability. Thresholding is a continuous function of the projection direction, which eliminates the case when $\hat{r}_n = N$, but the separating hyperplanes are parallel. Due to [10]

$$\forall \varepsilon > 0 \lim_{n \to \infty} P \left( M \int_{A_n \cup B} \, dx > \varepsilon \right) = 0$$

Therefore, due to [9] the reducible error of the method converges to 0 in probability.

If the optimal function for classification according to Bayes rule is not linear, there is a need to extend the feature space using monomials of order up to some $k > 1$ in the original data coordinates in order to achieve an arbitrarily small reducible error.

**Theorem 4.** Consider a binary classification problem with feature vectors taking values in $\mathbb{R}^d$. Assume that Bayes decision rule divides $B \subset \mathbb{R}^d$ into two measurable sets $B_1$ and $B_2$, such that the optimal classification for each $x \in B_1$ is $\omega_1$ and for each $x \in B_2$ is $\omega_2$. Consider the classifier obtained by picking a random polynomial of order $k$ in the original data $n$ times, classifying the data according to the sign of the value of the polynomial, and choosing the polynomial with minimal population error among the $n$ polynomials. Then for any given $\varepsilon > 0$ there exists a $k$ and a number of iterations $n = n(k, d)$ such that the reducible error of the chosen polynomial is smaller than $\varepsilon$ with large probability.
Proof. The proof is a modification of a proof presented in [4]. We would like to approximate the optimal decision function \( c(x) \) that gives 1 if \( x \in B_1 \) and \(-1\) if \( x \in B_2 \). Let us assume that our algorithm classifies data according to a different function \( \hat{c}(x) \). The reducible error that is produced by this algorithm can be expressed in the following way:

\[
\int_B \frac{1}{2} \left( c(x) \neq \hat{c}(x) \right) \left( \max_i \{ \rho_i(x) P(\omega_i) \} - \min_i \{ \rho_i(x) P(\omega_i) \} \right) dx.
\]

We build a function \( \hat{c}(x) \) by randomly picking the coefficients of a polynomial \( p(x) \) of some degree \( k \) and classify the data according to the rule \( \hat{c}(x) = \text{sign}(p(x)) \). We seek to find a polynomial \( p(x) \) with minimal reducible error.

Let us pick a continuous function \( g(x) \) that approximates \( c(x) \). By Lusin’s theorem, on a measure space \((\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d), \mu)\), where \( \mu \) is a Lebesgue measure, we know that \( c \) is equal to some continuous function \( g \) on an arbitrarily large set. That is, for each \( \hat{\epsilon} > 0 \) there exists a set \( \hat{B} \subset B \), such that \( \mu(B \setminus \hat{B}) < \hat{\epsilon} \) and:

\[ c(x) = g(x), \text{ for all } x \in \hat{B}. \]

The quantity \( \hat{\epsilon} \) has to be a function of the error \( \epsilon \), so that the error of the method is indeed small enough. Choosing \( \hat{c}(x) \) according to the sign of \( g(x) \) leads to the following reducible error:

\[
\int_{B \setminus \hat{B}} \frac{1}{2} \left( c(x) \neq \hat{c}(x) \right) \left( \max_i \{ \rho_i(x) P(\omega_i) \} - \min_i \{ \rho_i(x) P(\omega_i) \} \right) dx =
\]

\[
= \int_{B \setminus \hat{B}} \frac{1}{2} \left( \max_i \{ \rho_i(x) P(\omega_i) \} - \min_i \{ \rho_i(x) P(\omega_i) \} \right) dx \leq \frac{1}{2} M \mu(B \setminus \hat{B}) \leq \frac{1}{2} M \hat{\epsilon}. \tag{11}
\]

Therefore we can choose \( \hat{\epsilon} \) as \( 2 \epsilon / M \) to achieve the desired reducible error \( \epsilon \).

We can approximate the continuous function \( g \) by a polynomial \( p \) so that the classification is preserved: \( \text{sign}(g(x)) = \text{sign}(p(x)) \). Indeed, according to Stone-Weierstrass theorem since \( g \) is a continuous real-valued function defined on a closed and bounded set \( B \subset \mathbb{R}^d \) for each \( \hat{\epsilon} > 0 \) there exist a polynomial \( p(x_1, \ldots, x_d) \), such that:

\[ |g(x) - p(x)| < \hat{\epsilon}, \text{ for all } x \in B. \]

If we choose \( \hat{\epsilon} = 1/4 \), then \( \text{sign}(g(x)) = \text{sign}(p(x)) \) on the set \( \hat{B} \). The reducible error of deciding a class according to the sign of \( p(x) \) will be again smaller than \( \epsilon \) due to (11). Thus we have proven that there exists a polynomial \( p(x) \) that approximates the decision function and allows for an optimal decision (based on the Bayes rule) up to a small error \( \epsilon \).

The next question is whether we can approximate such a polynomial with a polynomial whose coefficients are chosen at random with a sufficiently high number or random draws. The optimal polynomial \( p(x) \) is of certain degree \( k \). Let us consider a polynomial transform of degree equal to \( k \) in the original data \( x \) using a mapping \( \phi : \mathbb{R}^d \to \mathbb{R}^d \):

\[ \phi(x) = (1, x_1, x_2, \ldots, x_i^a_1 x_j^a_2 \ldots x_k^a). \]

The dimension of the new space of features is \( \tilde{d} \), which depends on the original dimension \( d \) and the degree \( k \) of the optimal polynomial \( p(x) \) and is equal to \( d = \binom{d+k}{k} \). Consider randomly generated coefficients \( a^{(n)} \sim \text{Unif} \left( S^{d-1} \right) \) and constructing a polynomial as a dot product
between the generated coefficients and a polynomial transformation of order \( k \) of the original data:

\[
p_n(x) = a^{(n)} \cdot \phi(x)
\]

Our target polynomial is \( p(x) \), can we get as close to it as possible by choosing one of \( n \) randomly generated polynomials? In other words, is it true that

\[
\forall \varepsilon \exists n \text{ and } \exists \hat{p}_n(x) \in \{p_1(x), ..., p_n(x)\} : \forall x \in B \ |p(x) - \hat{p}_n(x)| \leq \varepsilon?
\]

One restriction is that \( \|a^{(n)}\| = 1 \). The norm of the coefficients \( a \) in the optimal polynomial \( p(x) \) does not have to be 1, but we can rescale the coefficients by \( \|a\| \) getting a rescaled polynomial \( \tilde{p}(x) \):

\[
\tilde{p}(x) = \frac{p(x)}{\|a\|} = \frac{a_0}{\|a\|} x_0 + \frac{a_1}{\|a\|} x_1 + ... + \frac{a_d}{\|a\|} x_d^k.
\]

Let us denote the rescaled coefficients by \( \tilde{a} \). The sign of \( \tilde{p}(x) \) is positive whenever \( p(x) \) is positive. Therefore making a decision using the rescaled polynomial \( \tilde{p}(x) \) will result in the same overall error as using the optimal polynomial \( p(x) \). Moreover, on the set \( \hat{B} \), the value of the polynomial \( \tilde{p}(x) \) is close to either \( 1/\|a\| \) or \( -1/\|a\| \): if \( x \in \hat{B} \) then

\[
\begin{cases}
\tilde{p}(x) \in \frac{1}{4\|a\|} (3, 5) & \text{if } c(x) = 1, \\
\tilde{p}(x) \in \frac{1}{4\|a\|} (-5, -3) & \text{if } c(x) = -1.
\end{cases}
\]

With probability converging to 1 (see the proof of Theorem \[3\]), we can generate coefficients \( a^{(n)} \) in an open ball around the optimal coefficients \( \tilde{a} \):

\[
\|a^{(n)} - \tilde{a}\|_\infty < \varepsilon'.
\]

Then, for every \( x \in \hat{B} \):

\[
|\hat{p}_n(x) - \tilde{p}(x)| = \left| (a_0^{(n)} - \tilde{a}_0) + (a_1^{(n)} - \tilde{a}_1) x_1 + ... + (a_d^{(n)} - \tilde{a}_d) x_d^k \right| \leq \|a^{(n)} - \tilde{a}\|_\infty (1 + |x_1| + |x_2| + ... + |x_d^k|) \leq \|a^{(n)} - \tilde{a}\|_\infty \|\phi(x)\|_1 \leq \|a^{(n)} - \tilde{a}\|_\infty \delta \|x\|^k \leq \delta \tilde{B}_k^k \varepsilon',
\]

where \( |x|_* = |x|_1 \vee 1 \) If \( B_* = \text{diam}(B) \vee 1 \) then \( |x|_* \leq B_* \). If we pick

\[
\varepsilon' = \frac{3}{8 \delta \tilde{B}_k^k} |a|
\]

in \[12\] then \( \text{sign}(\hat{p}_n(x)) = \text{sign}(\tilde{p}(x)) \) and the error of the classifier that classifies according to the sign of \( \hat{p}_n(x) \) will be smaller than \( \varepsilon \) as a result of \[11\]. The last task is to find \( n \) large enough, so that with probability \( 1 - \delta \):

\[
\|a^{(n)} - \tilde{a}\|_\infty \leq \frac{3}{8\|a\| \delta \tilde{B}_k^k}.
\]

**Lemma 3.** If we generate random vectors for projection \( n \) times, where

\[
n = \frac{\ln(\delta)}{\ln\left(1 - I \sin^2\left(4\arctan\left(\frac{3}{16\|a\| \delta \tilde{B}_k^k}\right)\right) \left(\frac{\delta - 1}{2}\right)^2\right)}.
\]

16
and $I_\phi(a, b)$ is a regularized incomplete beta function, then with probability at least $1 - \delta$, there will be one polynomial that has the same sign as the optimal one $p(x)$ on the set $\tilde{B}$.

**Proof.** The following analysis uses formulas for the area of hyperspherical cap from [8]. Since $a$ is chosen from the unit sphere in $\tilde{d}$ dimensions uniformly at random, we are interested in the probability of missing a desirable set (two hyperspherical caps) that is close enough to the optimal vector of coefficients $\tilde{a}$. The area of a cap given by an angle $\phi$ is

$$A_{\tilde{d}}^{\text{cap}} = 2\pi^{\frac{\tilde{d}-1}{2}} \frac{\sin^{\frac{\tilde{d}-2}{2}}(\theta)}{\Gamma\left(\frac{\tilde{d}-1}{2}\right)} \int_0^\phi \sin^{\frac{\tilde{d}-2}{2}}(\theta) d\theta.$$ 

The probability that we will end up in the cap is thus given by

$$\frac{A_{\tilde{d}}^{\text{cap}}}{A_{\tilde{d}}} = \frac{1}{2} I_{\sin^2 \phi} \left( \frac{\tilde{d}-1}{2}, \frac{1}{2} \right) = \frac{B\left(\sin^2 \phi; \frac{\tilde{d}-1}{2}, \frac{1}{2}\right)}{2B\left(\frac{\tilde{d}-1}{2}, \frac{1}{2}\right)},$$

where $I_{\sin^2 \phi}$ is a regularized incomplete beta function. A regularized incomplete beta function, $I_{\sin^2 \phi}$, is a cumulative distribution function of the beta distribution. The probability that we will end up outside two caps that are close to the optimal projection direction, which are symmetric with respect to the origin and have the same surface is therefore given by:

$$1 - I_{\sin^2 \phi} \left( \frac{\tilde{d}-1}{2}, \frac{1}{2} \right).$$

We want to find $n$ such that for a given $\delta$ the following inequality is true:

$$P(\text{missing the caps after } n \text{ tries}) = \left(1 - I_{\sin^2 \phi} \left( \frac{\tilde{d}-1}{2}, \frac{1}{2} \right)\right)^n \leq \delta,$$

which is equivalent to

$$n \geq \frac{\ln(\delta)}{\ln\left(1 - I_{\sin^2 \phi} \left( \frac{\tilde{d}-1}{2}, \frac{1}{2} \right)\right)}.$$ 

The inequality $\|a^{(n)} - \tilde{a}\|_\infty < \varepsilon'$ will be satisfied if $\|a^{(n)} - \tilde{a}\|_2 < \varepsilon'$, therefore the angle $\phi$ has to be smaller or equal to $4 \arctan\left(\frac{\varepsilon'}{2}\right)$. Since $\varepsilon' = 3/(8 \|a\| B_k)$ is enough to reach the desired error of the algorithm, the lower bound for $\phi$ is given by the following:

$$\phi \leq 4 \arctan\left(\frac{3}{16 \|a\| d B_k}\right),$$

therefore we need $n$ bigger than

$$n \geq \frac{\ln(\delta)}{\ln\left(1 - I_{\sin^2 \left(4 \arctan\left(\frac{3}{16 \|a\| d B_k}\right)\right)} \left( \frac{\tilde{d}-1}{2}, \frac{1}{2} \right)\right)}.$$ 

(13)
If we assume a simplified case, where \( \|a\| = 1 \) and \( B = 1 \) we get:

\[
n \geq \frac{\ln(\delta)}{\ln \left( 1 - I \sin^2 \left( \frac{3}{4} \arctan \left( \frac{\delta - 1}{\frac{1}{2}} \right) \right) \right)}.
\]

As a consequence, for any \( \varepsilon \) we can find \( k = k(\varepsilon) \) and \( n = n(k, d, \delta) \), so that with probability \( 1 - \delta \) the reducible error of the classification based on the sign of the polynomial with random coefficients that is chosen out of \( n \) tries is smaller than \( \varepsilon \). That means that the reducible error of such classification converges to zero in probability.

\[ \square \]

**Corollary 1.** Under the assumption that Bayes decision rule splits set \( B \) into measurable sets, the reducible error of the method of thresholding after random projection converges to 0 in probability as \( k \) and \( n \) converge to infinity.

**Proof.** Due to Theorem 1 for each \( \varepsilon > 0 \) there exists an order \( k \), such that with \( n \) large enough, the error of classification based on the sign of the polynomial with the smallest error out of \( n \) tries is smaller than \( \varepsilon \) with sufficiently large probability:

\[
\forall \varepsilon \exists k \text{ and } n(k) \text{ such that } \hat{p}_n(x) \in \{p_1(x), ..., p_n(x)\}
\]

is such that the error of classification using \( \text{sign}(\hat{p}_n(x)) \)

yields an error that is smaller than \( \varepsilon \).

The polynomials \( p_i(x) \) are of degree \( k \) with coefficients \( a \) picked randomly from a unit hypersphere in \( \mathbb{R}^d \). This corresponds to projecting the data on a random line and choosing the classification threshold at random. In the method of thresholding after random projection, all the coefficients are chosen at random, apart from the threshold, which corresponds to the constant term in a polynomial classifier according to sign. Recall that the threshold is optimized to minimize the population error. Therefore, if in the set with random coefficients \( \{p_1(x), ..., p_n(x)\} \) there exists a polynomial with a smaller error than \( \varepsilon \), in a set that is given by method of thresholding after random projection \( \{p_1^*(x), ..., p_n^*(x)\} \), there is also such a polynomial. Indeed, there is an equivalence between the event of polynomial \( p_i^*(x) \) appearing in the set of \( n \) randomly generated polynomials, where \( p_i^*(x) \) is exactly the same polynomial as \( p_i(x) \) apart from the absolute term. The absolute coefficient of \( p_i^* \) is found in the following way:

\[
a_0^* = \arg \min_{a \in \mathbb{R}} E_{\text{popul}}(\text{sign}(p_i(x) - a_0 + a)).
\]

That is why for each \( i \) the error of the classification according to \( p_i^*(x) \) is smaller than the error of classification according to \( p_i(x) \). And since there is an equivalence between events that are represented by the set of polynomials, the probability of getting a smaller error than \( \varepsilon \) using the method of thresholding after random projection is at least as high as the probability of this error using randomly generated polynomials.

\[ \square \]

In practice we do not know the underlying distributions of the classes, therefore we can not choose the projection direction or the threshold so that the population error is minimized.
Instead, we seek a small error on a finite set of training data points. Using this approach with a number of samples large enough we can learn the decision function that results in an error that is close to the optimal one as long as the VC dimension of our hypotheses set is finite. The next theorem shows that if we apply the method of thresholding after random projection on a training set, the error converges to zero as the order of polynomial $k$ and the number of iterations $n$ grow.

**Theorem 5.** Consider training set $D = \{x_1, ..., x_N\} \subset \mathbb{R}^d$ and training error defined as

$$E_{(k,n)} = \frac{1}{N} \sum_{i=1}^N \left( \hat{c}_{(k,n)}(x_i) - y_i \right)^2 = \frac{1}{N} \sum_{i=1}^N \mathbb{I} \left( \hat{c}_{(k,n)}(x_i) \neq y_i \right),$$

where $\hat{c}_{(k,n)}(x)$ is the classification from the method of thresholding after random projection applied on a polynomial transformation of order $k$ of the original data. Then there exists an order $k$, such that the error $E_{(k,n)}$ converges to 0 in probability as $n$ converges to infinity.

**Proof.** Let us consider a packing radius of the set $D = \{x_1, ..., x_N\}$. The packing radius is defined as a half of the infimum of distances between all of the points in the set, so that open balls with the radius of $r$ around all of the points in the set do not intersect. Since the set $D$ is finite and consists of distinct points, the packing radius $r$ is larger than 0. Let us define a continuous function $g(x)$ that is equal to 1 for the points in $D$ that belong to the first class together with the open ball around them with radius $r/2$. Similarly, for the points in $D$ that belong to the second class, as well as for their open neighbourhood, the value of the function will be $-1$:

$$\begin{cases} 
  g(x) = 1 & \text{if there exists } i \text{ such that } \|x - x_i\|_2 < \frac{r}{2} \text{ and } y_i = 1, \\
  g(x) = -1 & \text{if there exists } i \text{ such that } \|x - x_i\|_2 < \frac{r}{2} \text{ and } y_i = 0.
\end{cases}$$

For all other points in the convex hull $\text{Conv}(D)$, function $g(x)$ can be defined voluntarily as long as it is continuous. Then there exists a polynomial $p(x)$ that is as closed as desired to the function $g(x)$. In particular, there exists a polynomial $p(x)$ such that $|p(x) - g(x)| < 1/4$, therefore $\text{sign}(p(x)) = \text{sign}(g(x))$ for all $x \in D$. By Lemma 3 there exists an $n$ such that, with large probability, the classification of the method of thresholding after random projection is the same as classification using the sign of the polynomial $p(x)$. Thus for each training set $D$ there exists an order $k$, such that applying the procedure of thresholding after random projection enough times, the training error will be zero with large probability.

The convergence of the classification based on the method of thresholding after random projection to the Bayes decision function and the convergence of the classification based on thresholding after random projection to the decision function that gives zero error on any training data set are seemingly in conflict. That is because these two decision functions might (and probably will) be different. The conflict is resolved by realizing that the decision made in order to minimize the population error and the decision made in order to minimize the training error are different. In practice, the method will choose parameters based on minimizing the training error, but if we knew the underlying distribution, the parameters would have been chosen to minimize the population error. The theory described in 10 guarantees, for any classifier with a finite VC method, minimizing the training error ultimately minimizes the population error, the difference between these two becoming smaller as the number of samples $N$ increases and converges to 0 in probability as long as the VC dimension of the class of hypotheses that we use is finite. The smaller the VC dimension, the faster the convergence of the empirical - training error to the population error.
Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1826099. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation."

References

[1] Y. Abu-Mostafa, M. Magdon-Ismail, and Hsuan-Tien Lin. Learning from data. 2012.

[2] Avrim Blum. Random projection, margins, kernels, and feature-selection. In International Statistical and Optimization Perspectives Workshop” Subspace, Latent Structure and Feature Selection”, pages 52–68. Springer, 2005.

[3] Mireille Boutin and Alden Bradford. A highly likely clusterable data model with no clusters. arXiv 1909.06511, 2019.

[4] G. Cybenko. Approximation by superpositions of a sigmoidal function. Mathematics of Control, Signals, and Systems (MCSS), 2(4):303–314, December 1989.

[5] S. Han and M. Boutin. The hidden structure of image datasets. In 2015 IEEE International Conference on Image Processing (ICIP), pages 1095–1099, 2015.

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. Journal of the American Statistical Association, 58(301):13–30, 1963.

[7] Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani. Efficient search for approximate nearest neighbor in high dimensional spaces. SIAM Journal on Computing, 30(2):457–474, 2000.

[8] S Li. Concise formulas for the area and volume of a hyperspherical cap. Asian Journal of Mathematics and Statistics, 4, 01 2011.

[9] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. Theory of Probability and its Applications, 16(2):264–280, 1971.

[10] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, Berlin, Heidelberg, 1995.

[11] T. Yellamraju and M. Boutin. Clusterability and clustering of images and other “real” high-dimensional data. IEEE Transactions on Image Processing, 27(4):1927–1938, 2018.