Federal State Autonomous Educational Institution for Higher Education
National Research University Higher School of Economics

Faculty of Computer Science
Educational Program
Applied Mathematics and Information Science

TERM PAPER
Research project
Empirical Risk Bounds and Convex Optimization Methods on Real Data

Prepared by the student of group 187, 3rd year of study,
Иванушко Михаил Павлович / Ivanushko Mikhail Pavlovitch

Supervisor:
PhD, Assistant professor: Faculty of Computer Science / Big Data and Information Retrieval School, Bruno Frederik Bauwens

Moscow 2021

Abstract

We investigate empirical risk bounds for the linear-kernel SVM algorithm on real data, computing Rademacher complexity of empirical margin loss and other heuristics.

Keywords

SVM, Empirical Risk, Rademacher Complexity, Empirical Margin Loss

Аннотация

Мы изучаем значения верхних границ эмпирического риска в контексте линейного алгоритма SVM на реальных данных. Для этого мы рассчитываем "Rademacher complexity" различных классов функций и другие евристики.

1 Introduction and goals

1.1 Goals

The main goal of this coursework is to develop material for the Statistical Learning Course taught here at HSE. All of the graphics, notes and code will be contributed to the course. We have chosen to explore the topic of empirical risk bounds on real data in the context of binary classification with the SVM algorithm.

1.2 Tasks

A large portion of this work is based on chapters 3, 5 in "Foundations of Machine Learning" by Mehryar Mohri [1]; Namely, Theorem 5.8 (Margin bound for binary classification) and other theorems and definitions related to it. This theorem provides a bound on risk for the SVM algorithm in terms of empirical risk.

If we could magically turn every underlying inequality in that theorem to an equality, then the SVM algorithm would be the best learning algorithm, because SVM is an empirical risk minimizer. Having tight bounds would also allow us to explicitly know the optimal value for parameter C, which is usually chosen by cross-validation.

Our task is to check how "bad" the underlying inequalities are on real data, how they depend on parameter C and the resulting margin size ρ and intercept b. Having gotten this picture, we could in the future explore different objective functions for the SVM algorithm in the hopes of "tightening" the bounds.

2 Parameter C and margin size in SVM

When training an SVM, the free parameter C determines the weight of slack variables in the loss function. In practice, C picked by cross validation.

A small value of C results in a hyperplane with a large margin. As we increase C, the resulting margin size decreases. The relation between C and the resulting margin size is not linear. Often, changing C will not reduce the margin size below a certain value (see Figure 2.1).

Figure 2.1: Margin size and absolute value of intercept versus parameter C

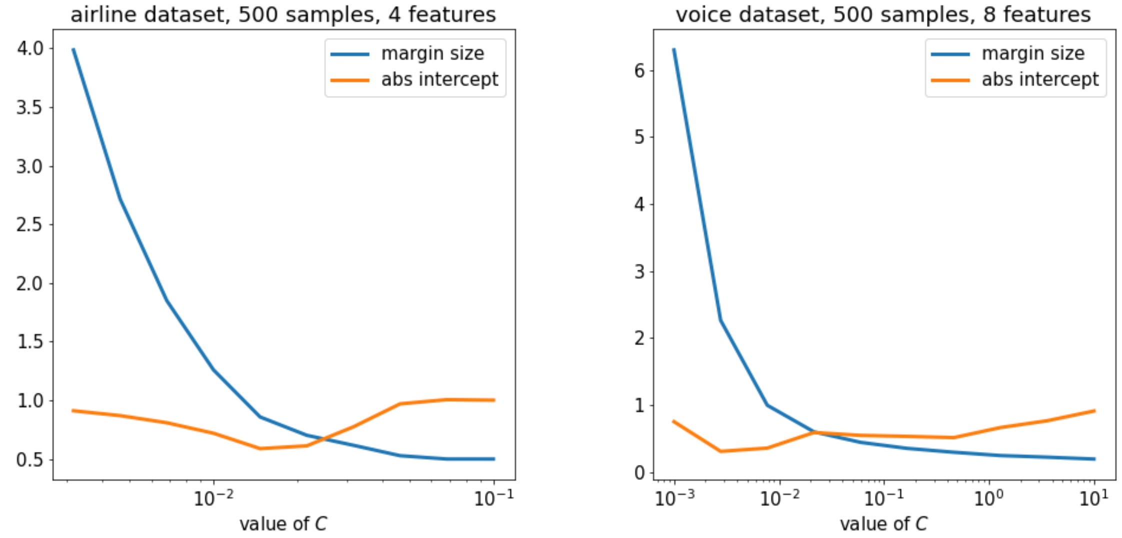
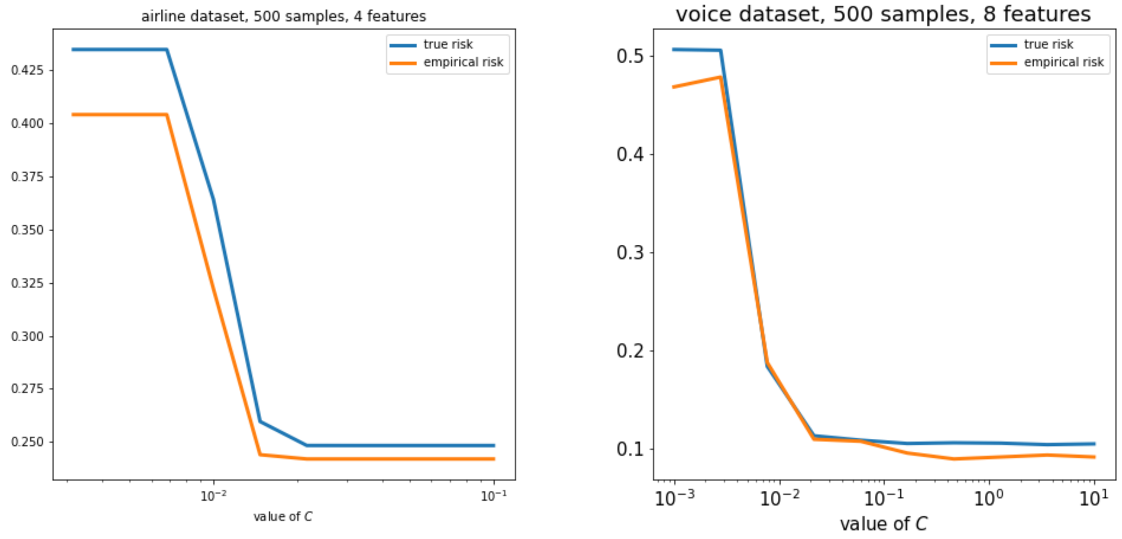


Figure 2.2: Risk on training and test sets versus parameter C



When C is large, the SVM algorithm prioritizes high accuracy on the training set over a large margin, which can result in overfitting.

The following work is about seeing what the risk bounds provided by Statistical Learning Theory can tell us about that risk, and what quantities contribute the most to those bounds.

3 Margin risk bound

Statistical Learning Theory provides us with a bound on the risk of any hypothesis class of real-valued functions. This is the main bound we will be analyzing. Let us take a look at the theorem first, and then give the definition for each of the terms.

Theorem 5.8 (Margin bound for binary classification). *Let H be a set of real-valued functions. Fix $\rho > 0$, then, for any $\delta > 0$, with probability at least $1-\delta$, the following holds for all $h \in H$:*

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathfrak{R}_{\mathrm{m}}(H) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.1}$$

In order to discuss this bound, we will require several definitions, presented in the following subsections.

3.1 Linear hypothesis class

When inspecting the margin risk bound, we will consider H to be the hypothesis class of signed distances to hyperplanes, defined by a unit-norm vector of weights w and an intercept b.

$$H = \left\{x \to w \cdot x + b\right\} \tag{3.2}$$

The confidence margin of a real-valued function h at a point x labeled y is the quantity $yh(x)$. The magnitude of $|h(x)|$ can be viewed as the confidence of a hypothesis h, and when $yh(x) > 0$, the prediction made by that hypothesis is correct.

$$\hat{H} = \left\{(x, y) \to y(w \cdot x + b)\right\} \tag{3.3}$$

3.2 Margin loss function and empirical margin loss

When training the SVM, we will penalize the hypothesis by 1 when it is wrong, and we will also penalize it linearly when it is right with a confidence $|h(x)|$ lower than some parameter ρ. This parameter defines the one-sided margin size of the SVM solution.

Definition 5.5 (Margin loss function). *For any $\rho > 0$, the ρ-margin loss is the function L_ρ : $\mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$ defined for all pairs $y \in \mathbb{R}, y' \in \{-1,+1\}$ by $L_\rho(y,y') = \Phi_\rho(yy')$ with,*

$$\Phi_\rho(t) = \min\left(1, \max\left(0, 1 - \frac{t}{\rho}\right)\right) \tag{3.4}$$

We can then define the empirical margin loss of h on some set S as the average margin loss of $yh(x)$ over all points in S:

Definition 5.6 (Empirical margin loss). *Given a sample $S = (x_1, ..., x_m)$ and a hypothesis h, the empirical margin loss is defined by*

$$\hat{R}_{S,\rho}(h) \leq \frac{1}{m}\sum_{i=1}^{m} 1_{y_i h(x_i) \leq \rho} \tag{3.5}$$

This quantity is greater than the traditional zero-one loss. If a correctly classified point lies within the margin, $1_{y_i h(x_i) \leq \rho} = 0$, but $\hat{R}_{S,\rho}(h) > 0$.

3.3 Rademacher complexity

The Rademacher complexity of a hypothesis class H describes how well it can correlate with a vector of independent, uniformly-distributed random variables taking values in $\{-1,+1\}$.

Definition 3.1 (Empirical Rademacher complexity) . *Let G be a family of real-valued functions, and $S = (z_1, ..., z_m)$ a fixed sample of size m with elements in Z. Then the empirical Rademacher complexity of G with respect to the sample S is defined as:*

$$\hat{\mathfrak{R}}_S(G) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{g \in G} \frac{1}{m}\sum_{i=1}^{m} \sigma_i g(z_i)\right], \tag{3.6}$$

where $\sigma = (\sigma_1, ..., \sigma_m)$, with σ_i being independent uniform random variables taking values in $\{-1,+1\}$. The random variables σ_i are called Rademacher variables.

Definition 3.2 (Rademacher complexity) . *Let D denote the distribution according to which samples are drawn. For any integer $m \geq 1$, the Rademacher complexity of G is the expectation of the empirical Rademacher complexity over all samples of size m drawn according to D:*

$$\mathfrak{R}_{\mathrm{m}}(G) = \mathop{\mathbb{E}}_{S \sim D^m}\left[\hat{\mathfrak{R}}_S(G)\right] \tag{3.7}$$

A more complex class will have a better ability to correlate with a random vector. If H only contains functions mapping x to $[-1,+1]$, then $\mathfrak{R}_{\mathrm{m}}(H) \in [0,1]$. Otherwise it can be greater than 1, or even be infinite.

For example, consider the linear confidences class H, with an unbounded value of the intercept b. If a Rademacher vector has more positive samples than negative ones, we can pick a hyperplane with an arbitrarily large positive intercept, resulting in arbitrarily large correlation with that particular Rademacher vector. If the vector has more negative samples, then we can pick a negative intercept and still get an arbitrarily large correlation.

For this reason, we will be bounding the absolute value of the intercept when we consider our hypothesis classes H and $\hat{H}$. When analyzing an SVM solution, we can bound b by the absolute value of the intercept in that solution.

The inequalities we will be analyzing operate with $\mathfrak{R}_{\mathfrak{m}}(G)$, and we will be taking subsets of size m from the datasets to approximate $\mathfrak{R}_{\mathfrak{m}}(G)$ with $\hat{\mathfrak{R}}_S(G)$ (more on that in section 7).

3.4 Rademacher complexity of linear confidence

Firstly, note that $\mathfrak{R}_{\mathfrak{m}}(H) = \mathfrak{R}_{\mathfrak{m}}(\hat{H})$:

$$\mathfrak{R}_{\mathfrak{m}}(\hat{H}) = \frac{1}{m} \mathop{\mathbb{E}}_{S,\sigma} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i y_i h(x_i) \right] = \frac{1}{m} \mathop{\mathbb{E}}_{S,\sigma} \left[\sup_{h \in H} \sum_{i=1}^{m} \sigma_i h(x_i) \right] = \mathfrak{R}_{\mathfrak{m}}(H) \tag{3.8}$$

Approximating $\hat{\mathfrak{R}}_S(H)$ with respect to a set S is simple. We generate a number of random Rademacher vectors, then find the best-correlating hypothesis h for each one. The average of these correlations will be an approximation of $\hat{\mathfrak{R}}_S(H)$.

For a particular random vector, the best-correlating plane's normal will point towards the center of gravity of the set (accounting for the sign of the Rademacher variables). The intercept will have the maximum absolute value, and will be either positive or negative, depending on the balance of positive / negative labels in the vector.

Given a bound on the intercept $\hat{b}$, the hypothesis $\tilde{h}$ that correlates best with the Rademacher vector $\sigma = (\sigma_1, ..., \sigma_i)$ can be expressed like this:

$$\tilde{h} : x \to x \cdot \frac{\tilde{w}}{\| \tilde{w} \|} + \tilde{b} \tag{3.9}$$

$$\tilde{w}_i = \frac{1}{m} \sum_{i=1}^{m} x_i \sigma_i \tag{3.10}$$

$$\tilde{b} = \hat{b} \times \operatorname{sign} \left[\sum_{i=1}^{m} \sigma_i \right] \tag{3.11}$$

3.5 Rademacher complexity of margin loss

The class of empirical margin loss functions $\left\{ \Phi_\rho(yh(x)) \right\} = \Phi_\rho \circ \hat{H}$ is defined not only by the intercept b, but also by the one-sided margin size ρ, and target labels y_i.

Unfortunately for us, approximating $\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \hat{H})$ with a high accuracy is not easy. Here is our challenge: given a Rademacher vector σ, we need to find the hyperplane (unit-norm w and a bounded b) for which the margin loss function correlates best with σ.

This is not a convex optimization problem due to Φ_ρ being non-convex (as opposed to the traditional hinge loss used in SVMs). Worse still, if ρ is very small in relation to the smallest distance between points in X, it becomes an NP-hard problem, see [2].

The solution for this project is straightforward - use black-box optimization methods for $\mathfrak{R}_m(\Phi_\rho \circ \hat{H})$, look at the resulting data, and try to gather some intuition about it. We start by considering small datasets (200 points with <= 4 features), and look at the dynamics in relation to ρ and b as we increase the size of X and the number of features (more on that in section 5).

It is important to note that our approximations of $\mathfrak{R}_m(\Phi_\rho \circ \hat{H})$ will always be smaller than the true value, because we are approximating a supremum of correlations for each generated Rademacher vector by searching the hypothesis class H. For each Rademacher vector, whatever hypothesis h we find will have a correlation less than or equal to the supremum.

Now, let us return to the margin bound and examine it again. The hypothesis class H in this inequality corresponds to the linear hypothesis class (3.2).

$$R(h) \leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathfrak{R}_m(H) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.1}$$

The LHS of the inequality $R(h)$ is the true risk of a hypothesis h, or the expected proportion of points misclassified by h (in practice, this is the error on the test set).

$$R(h) = \mathbb{E}\left[1_{yh(x)\leq 0}\right] \tag{3.12}$$

$\hat{R}_{S,\rho}(h)$ is the empirical margin loss on our training set, $\mathfrak{R}_m(H)$ is the Rademacher complexity of the linear hypothesis class. The last term in the RHS will be constant after we choose our set S and the required precision δ. For this project, we chose δ to be 0.1, so the bound will hold with probability at least 90%. The general dynamics of the bounds will not depend on this term.

Notice the factor $\frac{2}{\rho}$ in the RHS. If the Rademacher complexity of the linear hypothesis turns out to be significant, then this bound will quickly become uninformative as we look at smaller and smaller margins.

But this is not the only bound we can look at. During the proof of Theorem 5.8, several intermediary bounds are derived. $R(h)$ is being bounded by a chain of expressions that ends in bound 3.1. Let us take a look at these expressions now:

$$R(h) \leq \mathbb{E}\Big(\Phi_\rho(yh(x))\Big) \tag{3.13a}$$

$$\leq \hat{R}_{S,\rho}(h) + \mathbb{E}\Big(\Phi(S)\Big) + \sqrt{\frac{log\frac{1}{\delta}}{2m}} \tag{3.13b}$$

$$\leq \hat{R}_{S,\rho}(h) + 2\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{log\frac{1}{\delta}}{2m}} \tag{3.13c}$$

$$\leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathfrak{R}_{\mathfrak{m}}(H) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.13d}$$

There is one last heuristic we need to define, it appears as $\Phi(S)$ on line 3.13b.

3.6 Outlier Function

Consider any family of functions G mapping Z to $[0,1]$. For any set S, let us denote by $\hat{\mathbb{E}}_S[g]$ the empirical average of $g \in G$ over $S = (z_1,...,z_m)$:

$$\hat{\mathbb{E}}_S[g] = \frac{1}{m}\sum_{i=1}^{m} g(z_i) \tag{3.14}$$

Then, let us define a function $\Phi(S)$ (which we will call "outlier function") as follows:

$$\Phi(S) = \sup_{g \in G}\Big(\mathbb{E}[g] - \hat{\mathbb{E}}_S[g]\Big) \tag{3.15}$$

During the proof of Theorem 3.3, the expectation of $\Phi(S)$ over all possible sets $S = (z_1,...,z_m)$ is upper-bounded by $2\mathfrak{R}_{\mathfrak{m}}(G)$. In our case, $G = \Phi_\rho \circ \tilde{H}$. We already know that approximating $\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H})$ is a non-convex NP-hard problem, and it appears that approximating $\mathbb{E}[\Phi(S)]$ is similar.

The same things we said about $\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H})$ also apply here - we will use black-box optimization on a smaller datasets to get an accurate enough result, and look at general dynamics for larger datasets.

4 Datasets

In this coursework we have worked with several datasets:

- A small (210 points, 7 features) dataset with geometrical properties of kernels belonging to three different varieties of wheat.[3]
- A small (150 points, 4 features) dataset with measurements of iris species.[4]
- A medium-sized (3168 points, 20 features) dataset with acoustic properties of the voice and speech (the classes being "Male" and "Female").[5]
- A large (129487 points, 24 features) dataset of responses to an airline passenger satisfaction survey (the classes being "Satisfaction" and "Neutral or dissatisfaction").[6]

We developed our implementations on smaller datasets, and then examined the bounds on larger datasets. We need a large dataset in order to accurately approximate expectations of heuristics, for example, $\mathbb{E}\left[\Phi(S)\right]$ and $\mathfrak{R}_{\mathfrak{m}}(G) = \mathbb{E}_{S\sim D^m}\left[\hat{\mathfrak{R}}_S(G)\right]$.

5 Experimentals: dependance or $\hat{\mathfrak{R}}_S$ on features and size of the dataset

As already noted, we will be approximating $\hat{\mathfrak{R}}_S$ with black-box optimization. The method we will use is Dual Annealing, with initial temperature of 50000, maximum iterations of 1000, and accept coefficient of -5. These coefficients were picked to strike a balance between finding good correlations and speed of computation to meet the time constraints for this work.

Another method that was considered is Differential Evolution, which worked well on small datasets with 2 features. However, the significant slowdown when including additional features into the dataset made it impractical for our case.

In the following subsections we look at how $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ changes depending on the size of the set and its features. The sets we will be considering are subsets from the "acoustic properties of the voice and speech" dataset mentioned in section 4.

5.1 Dependance on number of features

As we increase the number of meaningful features in the set S, the empirical Rademacher complexity increases as well (see Figure 5.1). This makes intuitive sense - making the points in the dataset more separable gives us more ways to cut the dataset with a hyperplane, thus leading to better correlation with any Rademacher vector.

The complexity on a given set with a given number of features $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ seems to change like $\sqrt{\frac{1}{\rho}}$ (see Figure 5.2). This dependence will also be present in future graphs. If we additionally normalize by $\sqrt{n}$, where n is the number of features, we can still see an increase in complexity (see Figure 5.3). Note this as we continue to the next subsection.

Figure 5.1: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ and $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$; different # of features

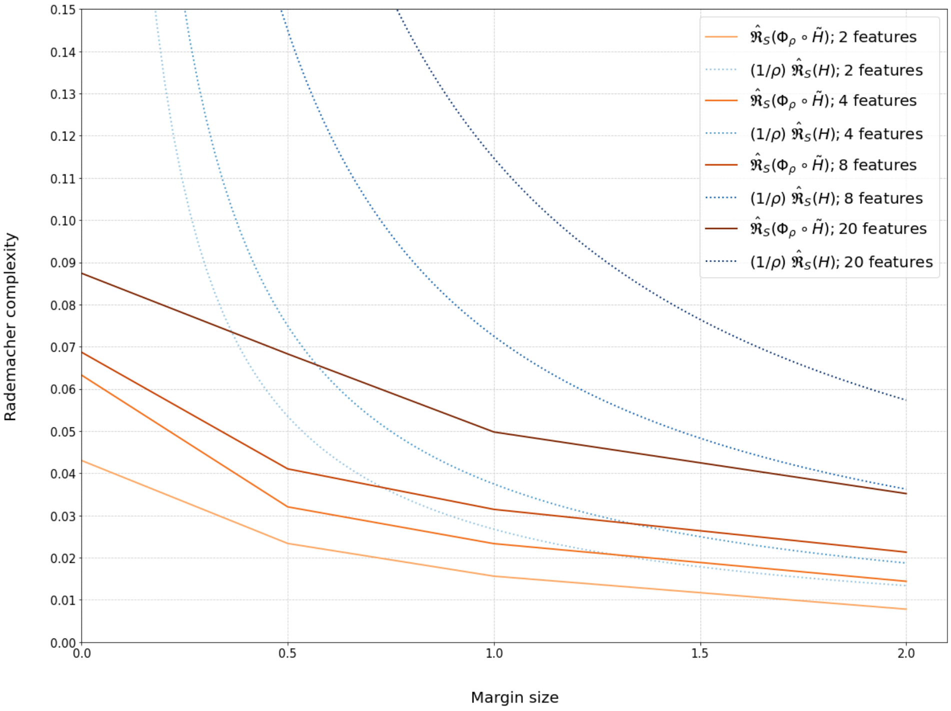


Figure 5.2: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ and $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$; different # of features; normalized by $\sqrt{\frac{1}{\rho}}$

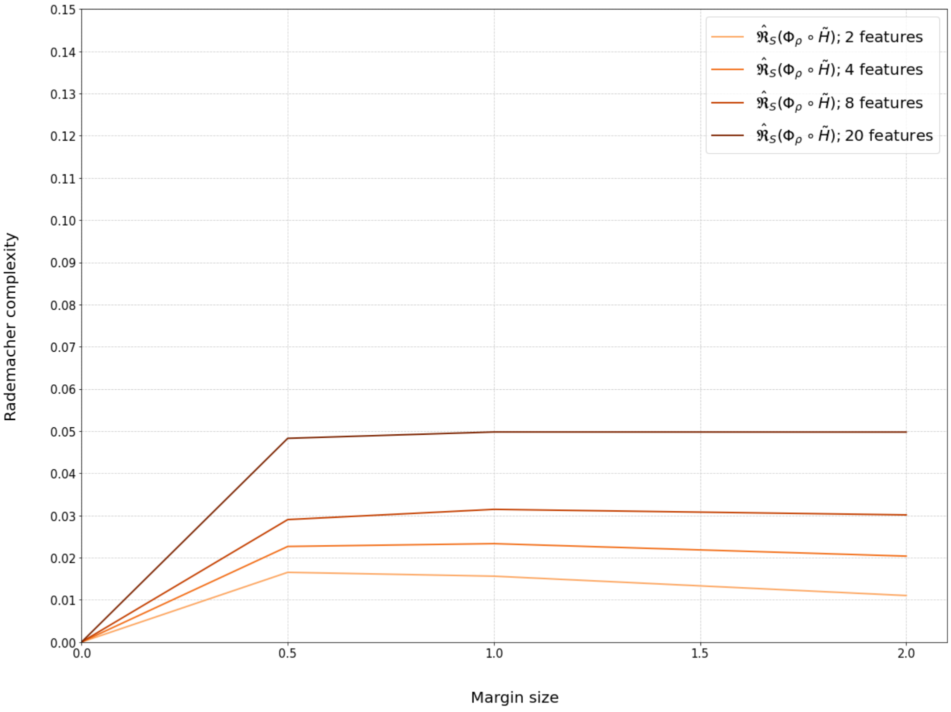
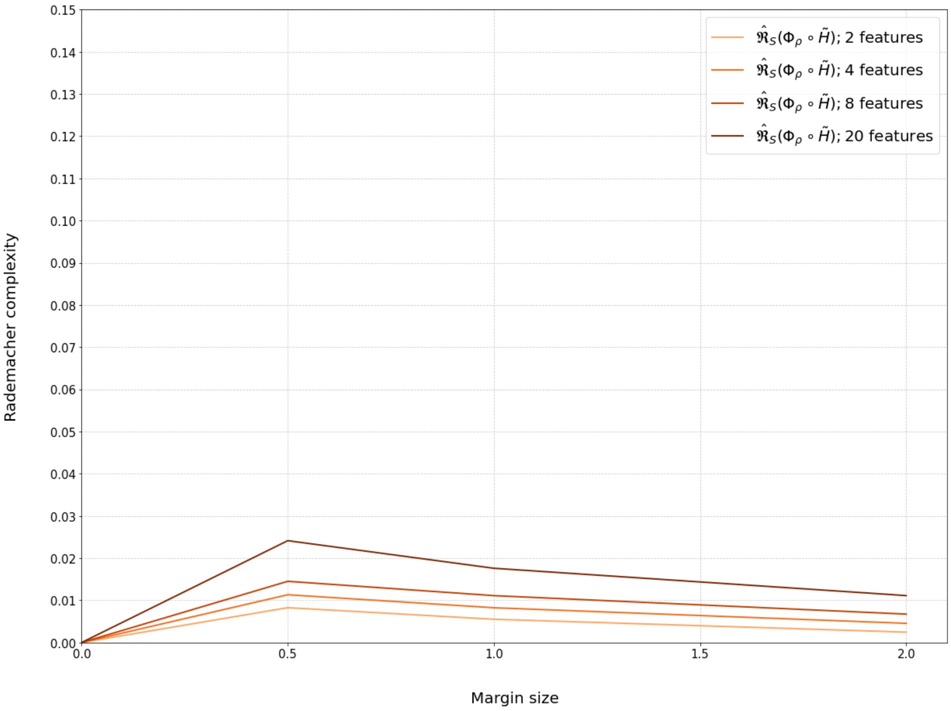


Figure 5.3: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ and $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$; different # of features; normalized by $\sqrt{\frac{n}{\rho}}$



5.2 Adding duplicate features

Adding duplicate features does not give us the same increase in complexity (Figure 5.4). After we normalize by $\sqrt{\frac{n}{\rho}}$ (where n is the number of features), we can notice that the complexity becomes basically constant (Figure 5.5). This suggests a baseline increase of complexity of $\sqrt{n}$, regardless of what the features in the dataset are.

This effect can be explained by how the relative size of the margin to the scale of the dataset changes as we add more features. Distances become longer, while value of ρ stays the same, and so the effective margin size decreases.

Figure 5.4: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ and $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$; different # of duplicated features

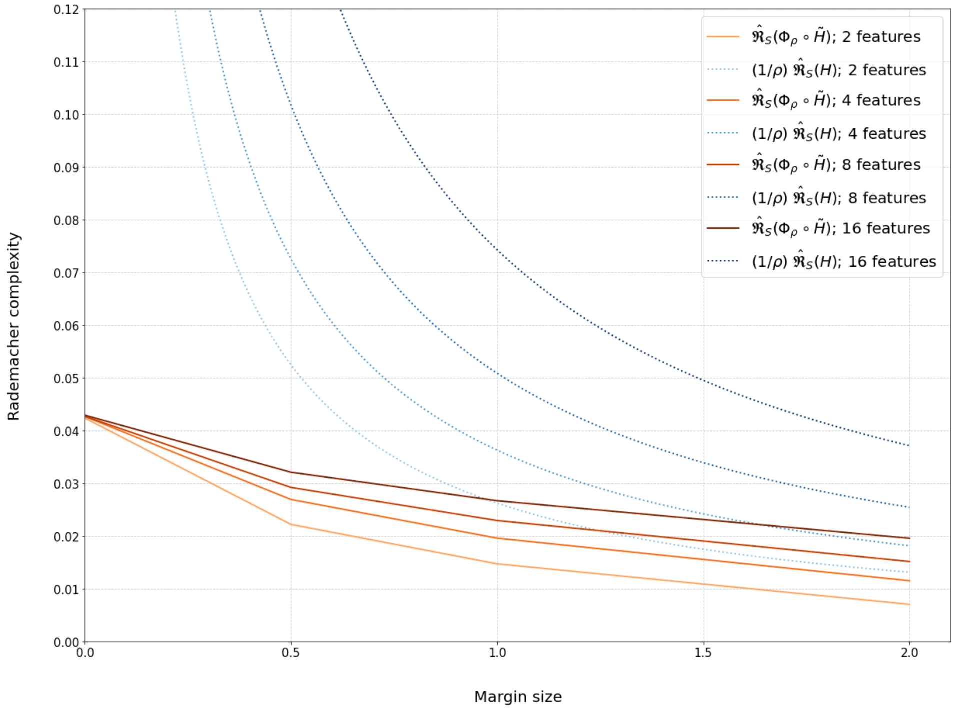
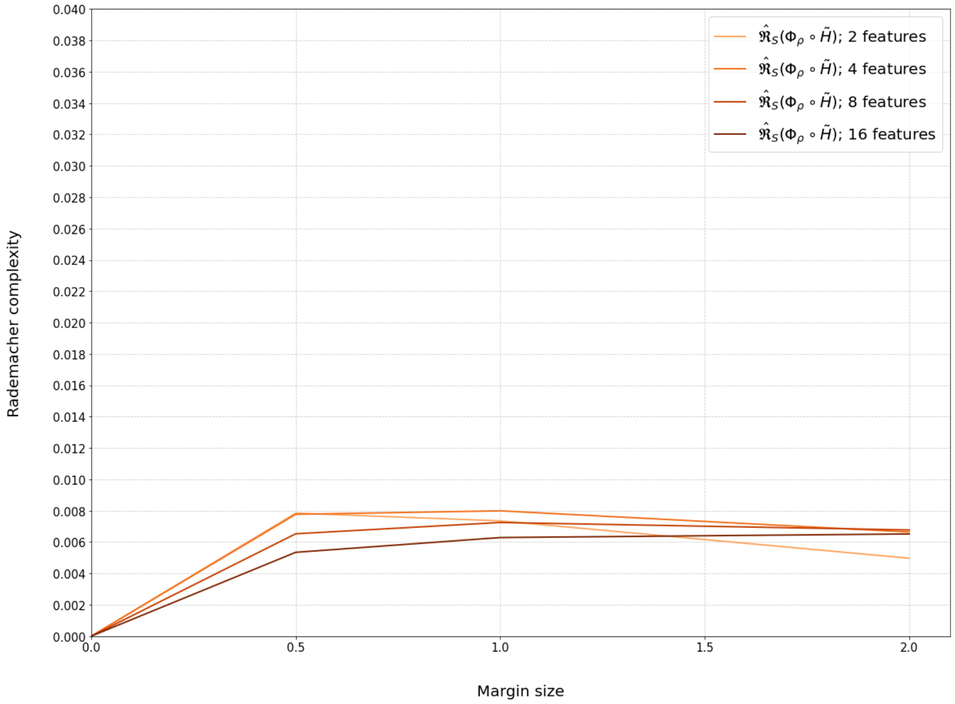


Figure 5.5: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ and $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$; different # of duplicated features; normalized by $\sqrt{\frac{n}{\rho}}$



5.3 Dependance on size of the dataset

Increasing the size of the dataset m will make Rademacher complexity decrease like $\sqrt{\frac{1}{m}}$ (See Figure 5.6, 5.7). This makes intuitive sense - as the dataset becomes more and more dense, it is harder and harder for us to find any significant imbalances or patterns in the Rademacher vector, and so the correlation will get closer to 0.

Figure 5.6: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ and $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$; different # of samples m;

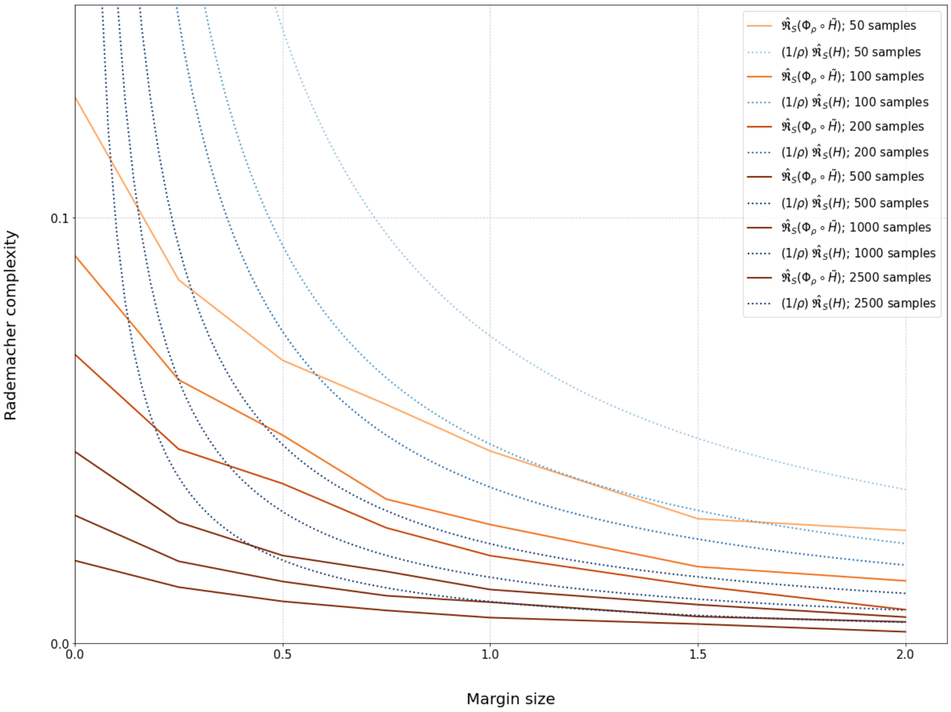
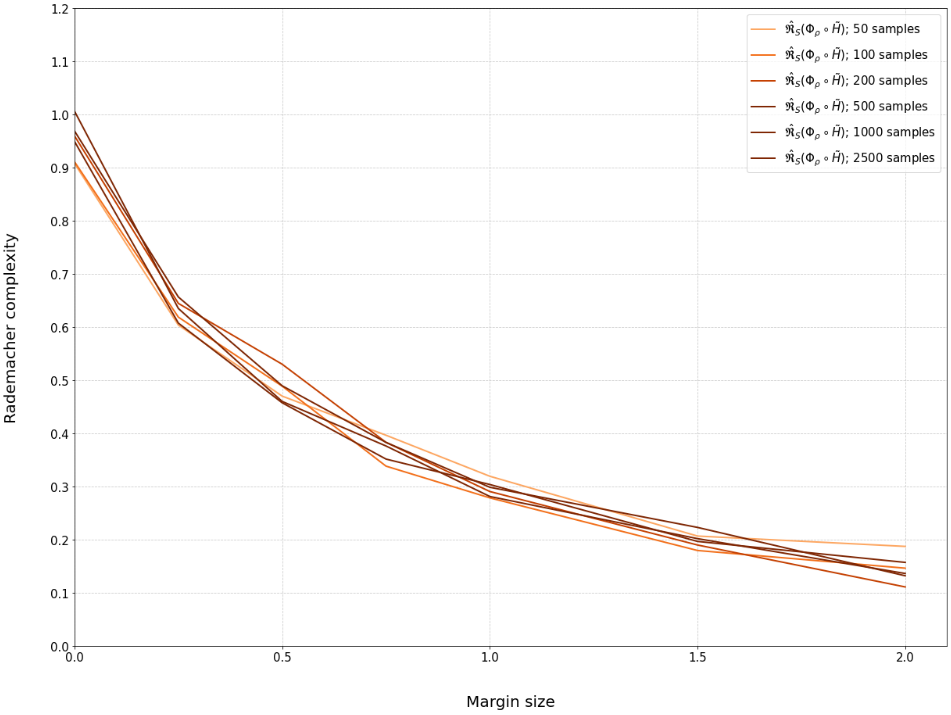


Figure 5.7: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ and $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$; different # of samples m; normalized by $\sqrt{\frac{1}{m}}$



6 Experimentals: dynamics of other bounds

In this section we will look at dynamics of other heuristics that appear in the proof of the margin bound 3.1.

6.1 Upper bound on risk $R(h)$

This subsection is in reference to the bound 3.13a.

$$R(h) \leq \mathbb{E}\Big(\Phi_\rho(yh(x))\Big) \tag{3.13a}$$

It stems from the fact that $R(h) = \mathbb{E}\left[1_{yh(x)\leq 0}\right] \leq \mathbb{E}\left[\Phi_\rho(yh(x))\right]$. The margin loss has an additional penalty to correctly classified points within the margin. Figure 6.1 shows how this difference changes as the margin size ρ increases (In this figure we fix our hyperplane in place and only change the margin size).

6.2 Upper bound on $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$

This section is in reference to the inequality between 3.13c and 3.13d. We have already featured this bound for the "voice" dataset in earlier graphs, so let us look at it for the "airline" dataset while changing the bound on the intercept b (Figure 6.2). In this case we were taking a subset of size 500, with 12 features. What we can notice immediately is how uninformative this bound becomes at smaller margin sizes.

Figure 6.1: $\Phi_\rho(yh(x))$ and Binary loss; changing margin size ρ

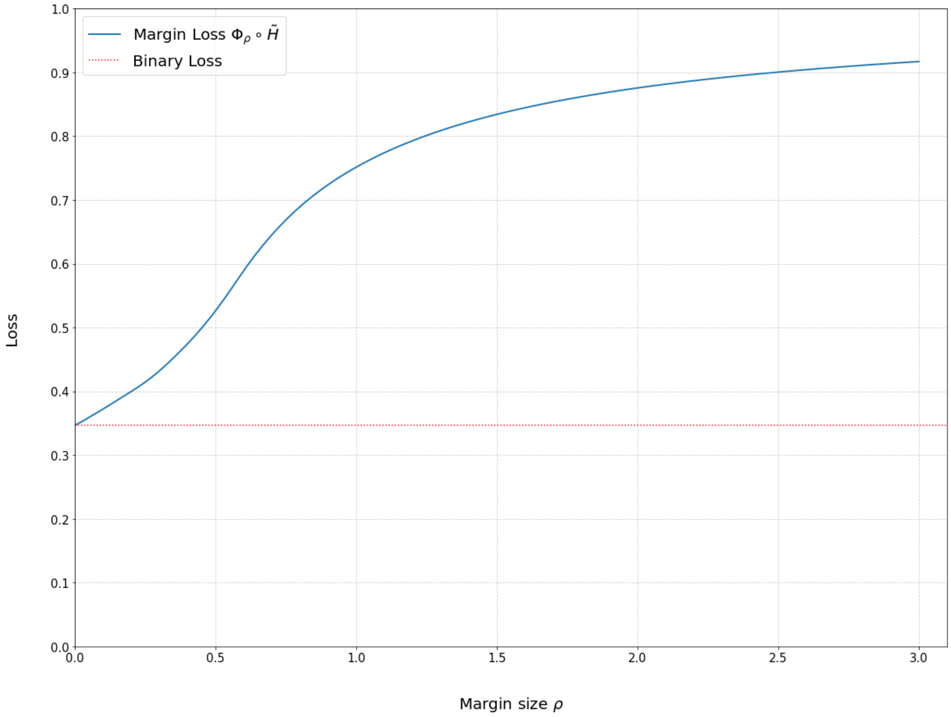
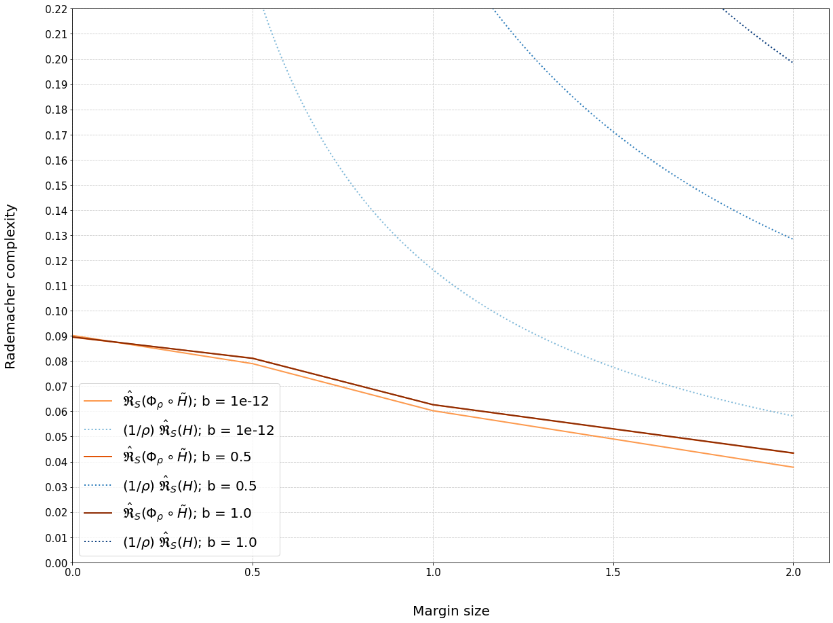


Figure 6.2: $\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ versus $\frac{1}{\rho}\hat{\mathfrak{R}}_S(H)$



6.3 Upper bound on Outlier function $\Phi(S)$

This section is in reference to the inequality between 3.13b and 3.13c. Figure 6.3 shows the difference between $\Phi(S)$ and $2\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$ as we change the margin size ρ, and for two values of bound on intercept b.

As opposed to the bounds mentioned previously, this bound appears to have a more complex dependence on ρ (see Figure 6.4).

Figure 6.3: $\Phi(S)$ versus $2\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$

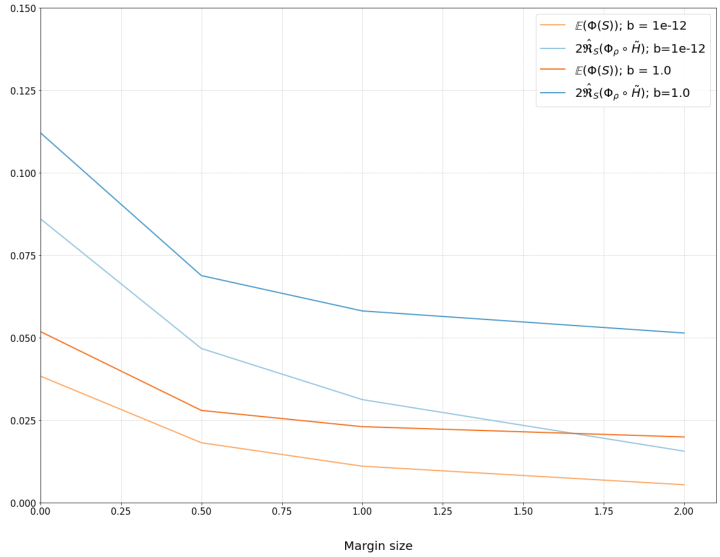
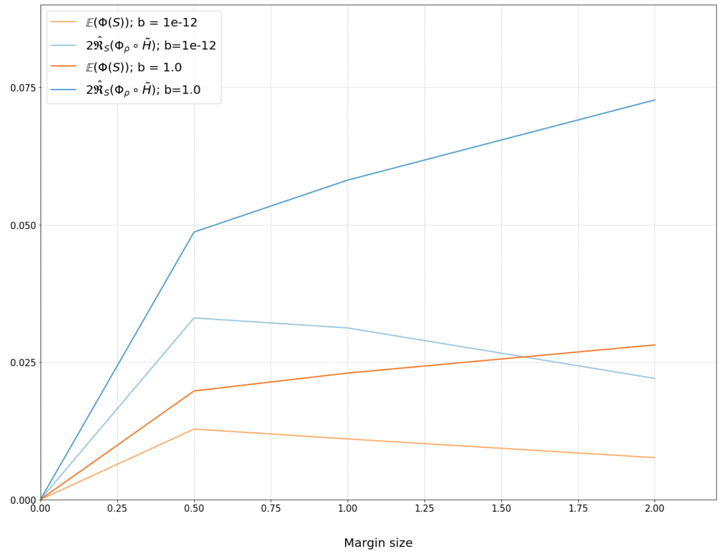


Figure 6.4: $\Phi(S)$ versus $2\hat{\mathfrak{R}}_S(\Phi_\rho \circ \tilde{H})$; normalized by normalized by $\sqrt{\frac{1}{\rho}}$



7 Experimentals: static evaluations

In this section we will put everything we have done together in one place. Firstly, we will compute solutions of standard SVM for different values of parameter C. Each solution will have a margin size ρ and an intercept b, and we will use those values to define our hypothesis classes. After computing every heuristic we need, we will look at the series of inequalities 3.13a - 3.13d at once. Here they are again:

$$R(h) \leq \mathbb{E}\Big(\Phi_\rho(yh(x))\Big) \tag{3.13a}$$

$$\leq \hat{R}_{S,\rho}(h) + \mathbb{E}\Big(\Phi(S)\Big) + \sqrt{\frac{log\frac{1}{\delta}}{2m}} \tag{3.13b}$$

$$\leq \hat{R}_{S,\rho}(h) + 2\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H}) + \sqrt{\frac{log\frac{1}{\delta}}{2m}} \tag{3.13c}$$

$$\leq \hat{R}_{S,\rho}(h) + \frac{2}{\rho}\mathfrak{R}_{\mathfrak{m}}(H) + \sqrt{\frac{\log\frac{1}{\delta}}{2m}} \tag{3.13d}$$

To approximate $\mathfrak{R}_{\mathfrak{m}}$, we take subsets $S = (x_1, ..., x_m)$ from the whole dataset, generate Rademacher vectors for each S, and find the best correlations to them. By Chernoff bound, the resulting value converges to the expected value $\mathbb{E}_S\left[\hat{\mathfrak{R}}_S\right] = \mathfrak{R}_{\mathfrak{m}}$ exponentially fast.

Figure 7.1: Bounds on the "airline" dataset, 4 features, 500 points, $\delta = 0.1$

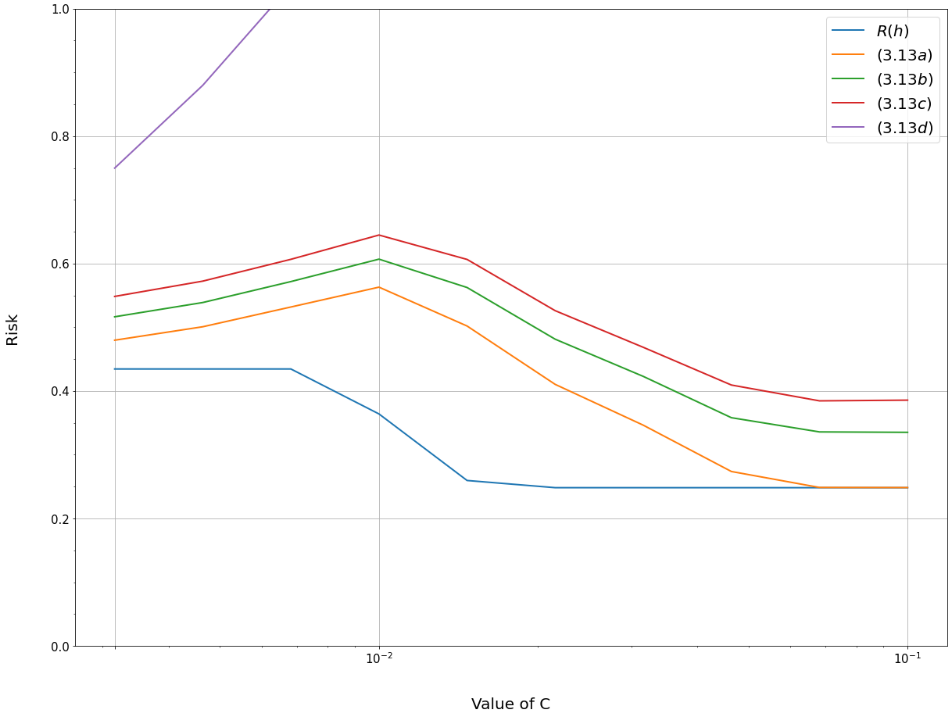


Figure 7.2: Bounds on the "voice" dataset, 8 features, 500 points, $\delta = 0.1$

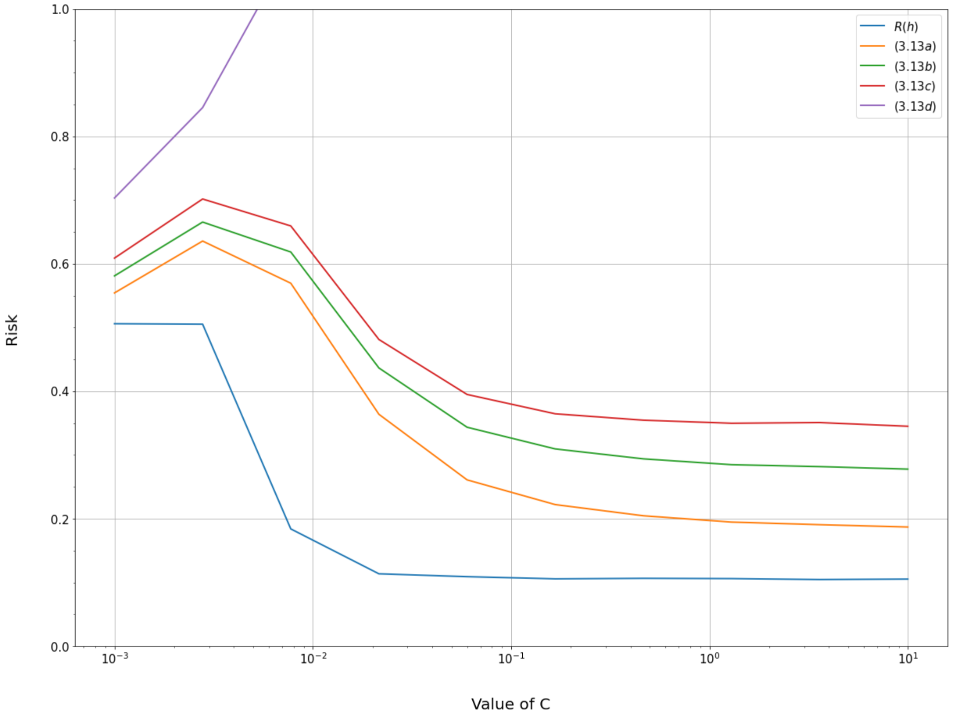


Figure 7.3: Bounds on the "airline" dataset, 4 features, 500 points, $\delta = 0.1$

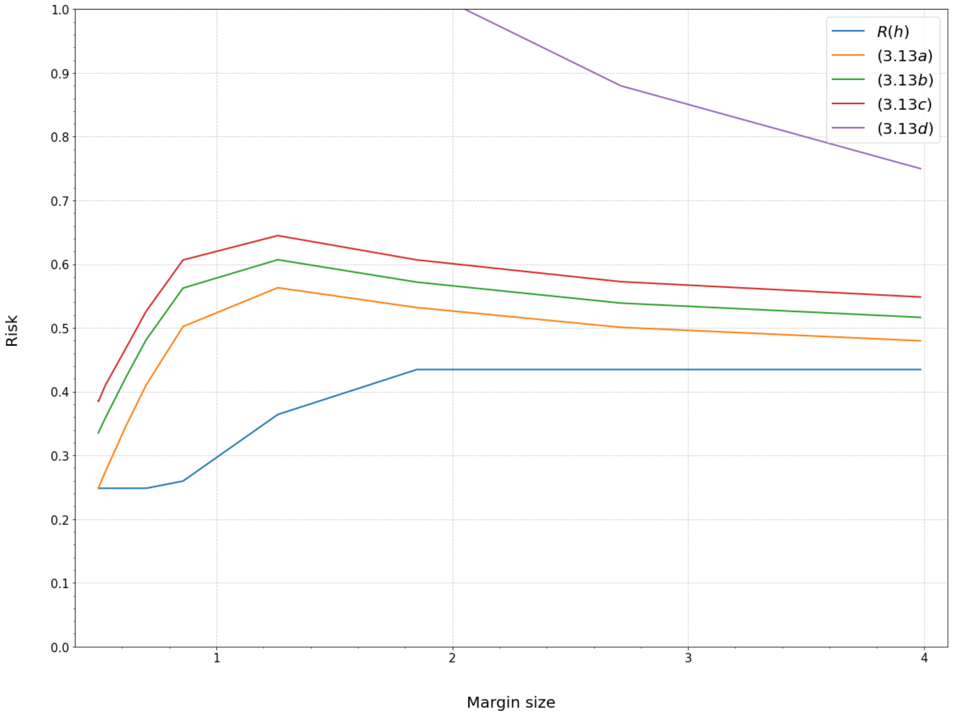
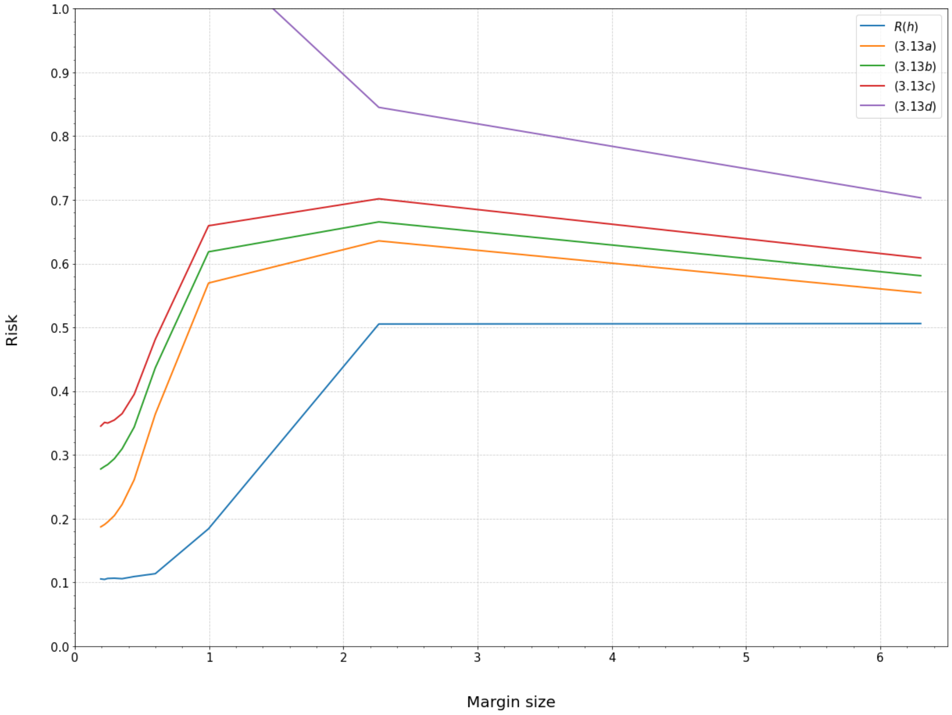


Figure 7.4: Bounds on the "voice" dataset, 8 features, 500 points, $\delta = 0.1$



These figures highlight places where our bound on risk $R(h)$ becomes large. Most noticeably, the final bound 3.13d is uninformative for almost any margin size, but especially at relatively small margin sizes. The other inequalities seem to provide an informative bound on risk $R(h)$.

The final inequality is based on our margin loss function Φ_ρ being ρ-Lipschitz. That is true, however Φ_ρ only has that slope ρ for confidences inside the margin, and it is constant outside of that region. That statement is much too general and results in an uninformative bound for small margin sizes.

Another issue is that as we increase the upper-bound on intercept b, the Rademacher complexity of linear hypothesis $\mathfrak{R}_{\mathfrak{m}}(H)$ grows much faster than $\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H})$. A bigger intercept bound for H, $\tilde{H}$ allows the margin loss more "freedom" to find a better-correlating hyperplane, but for the linear hypothesis it simply increases the correlation.

7.1 Accuracy

All of the quantities for this section were computed with 1350 iterations. When computing the Rademacher complexity, on each iteration we generate a subset of size m from the whole dataset, generate a Rademacher vector, and try to find the best correlation. The achieved correlations display a variance of 7×10^{-7}, which translates to a maximal error of 2.23×10^{-4} with 99% confidence.

However, as we can not guarantee the optimal correlation on every Rademacher vector, the computed Rademacher complexity has an inherent level of inaccuracy. Our goal was to simply have adequate accuracy to see the trends in these bounds. For a dataset with 500 points and ≤ 8 features, the trends we discussed become noticeable even at around 200 iterations.

Tables 7.1 and 7.2 show all computed quantities that were used for the graphs in this section.

7.2 Compute time

$\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H})$ and $\Phi(S)$ were the most time-consuming heuristics to compute, since we were using a black-box optimization method. Both heuristics took around 4 seconds per iteration for a dataset with 500 points and 8 features.

A detailed analysis of these times and how they depend on the data was not made. Such an analysis would be useful when considering another optimization technique for finding $\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H})$ and $\Phi(S)$, which is one of possible directions for future research.

C	ρ	$R(h)$	$\mathbb{E}\big(\Phi_\rho(yh(x))\big)$	$\Phi(S)$	$\mathfrak{R}_\mathfrak{m}(\Phi_\rho \circ \tilde{H})$	$\hat{R}_{S,\rho}(h)$
0.0032	3.9858	0.4346	0.4797	0.0232	0.0276	0.5112
0.0046	2.7155	0.4346	0.5009	0.0263	0.0299	0.4985
0.0068	1.8500	0.4346	0.5318	0.0308	0.0329	0.4588
0.0100	1.2604	0.3641	0.5630	0.0354	0.0366	0.4313
0.0147	0.8587	0.2596	0.5020	0.0400	0.0420	0.3548
0.0215	0.7026	0.2483	0.4104	0.0420	0.0434	0.3580
0.0316	0.6173	0.2483	0.3464	0.0428	0.0442	0.4479
0.0464	0.5288	0.2483	0.2737	0.0448	0.0480	0.5439
0.0681	0.5004	0.2483	0.2487	0.0456	0.0471	0.5701
0.1000	0.5002	0.2483	0.2485	0.0451	0.0478	0.5434

Table 7.1: Computed values for the "airline" dataset

C	ρ	$R(h)$	$\mathbb{E}\big(\Phi_\rho(yh(x))\big)$	$\Phi(S)$	$\mathfrak{R}_\mathfrak{m}(\Phi_\rho \circ \tilde{H})$	$\hat{R}_{S,\rho}(h)$
0.0010	6.3003	0.5060	0.5543	0.0176	0.0228	0.4405
0.0028	2.2643	0.5052	0.6357	0.0218	0.0290	0.2281
0.0077	0.9961	0.1840	0.5696	0.0283	0.0346	0.2511
0.0215	0.5995	0.1136	0.3639	0.0344	0.0395	0.3668
0.0599	0.4423	0.1091	0.2610	0.0379	0.0447	0.3373
0.1668	0.3524	0.1057	0.2222	0.0398	0.0474	0.3337
0.4642	0.2945	0.1064	0.2046	0.0421	0.0514	0.3283
1.2915	0.2452	0.1061	0.1947	0.0443	0.0546	0.3946
3.5938	0.2213	0.1046	0.1907	0.0459	0.0575	0.4533
10.0000	0.1933	0.1053	0.1870	0.0470	0.0571	0.5118

Table 7.2: Computed values for the "voice" dataset

8 CONCLUSIONS

The computed values for the final bound 3.1 show that it is not informative in the case of binary classification with linear-kernel SVM. The underlying inequalities seem to be quite more accurate, but the heuristics required for them are not easy to compute.

We can see two underlying causes for the final bound being too high. Firstly, the use of the ρ-Lipschitz property, although correct, leaves us a coefficient $\frac{1}{\rho}$ that renders the bound uninformative at small margins. Secondly, as we increase the bound on intercept b, the Rademacher complexity of linear hypothesis $\mathfrak{R}_{\mathfrak{m}}(H)$ grows much quicker than $\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H})$.

However, in different contexts, these problems might be less of an issue.

- Since all of the theorems are written for real-valued functions, we can make full use of that fact and try looking at SVM regression instead of SVM classification.
- We can try different kernels for the SVM algorithm.
- We can try bounding the 0-1 loss with a different function that is closer to linear. That would make the ρ-Lipschitz property less of an issue.
- Computing $\mathfrak{R}_{\mathfrak{m}}(\Phi_\rho \circ \tilde{H})$ allows us to get an informative bound, but takes a long time. If we want to improve SVM, then we could try to replace the target of SVM by a better risk bound. But the challenge is to find a risk bound that is both better and still easy to compute.

The graphs and code[7] created during this coursework are already being used in the Statistical Learning Theory course on Coursera.

8.1 FUTURE WORK

In this research, we have come across several other questions that may be of interest. These could be a good starting point for future work:

- Using different kernels in the SVM algorithm.
- Looking at SVM regression instead of classification.
- Trying to find a better black-box method for estimating Rademacher complexity and the outlier function.
- Modifying the objective function of the SVM algorithm (for example, looking at a different power of w) and seeing if the dynamics of the bounds change.
- Modifying the margin loss function such that finding the best correlation to the Rademacher vector is an easier task.

References

[1] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. 2nd ed. Adaptive Computation and Machine Learning. Cambridge, MA: MIT Press, 2018. 504 pp. ISBN: 978-0-262-03940-6.

[2] Søren Frejstrup Maibing and Christian Igel. “Computational Complexity of Linear Large Margin Classification With Ramp Loss”. In: *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*. Ed. by Guy Lebanon and S. V. N. Vishwanathan. Vol. 38. Proceedings of Machine Learning Research. San Diego, California, USA: PMLR, 2015, pp. 259–267. URL: `http://proceedings.mlr.press/v38/frejstrupmaibing15.html`.

[3] M. Charytanowicz et al. *Seeds Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/seeds`.

[4] R.A. Fisher. *Dataset: Iris Species*. URL: `https://archive.ics.uci.edu/ml/datasets/iris`.

[5] Kory Becker. *Dataset: Gender Recognition by Voice*. URL: `https://www.kaggle.com/primaryobjects/voicegender/version/1`.

[6] TJ Klein. *Dataset: Airline Passenger Satisfaction*. URL: `https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction/version/1`.

[7] *All code:* URL: `https://github.com/mikhailivanushko/SLT-bounds`.