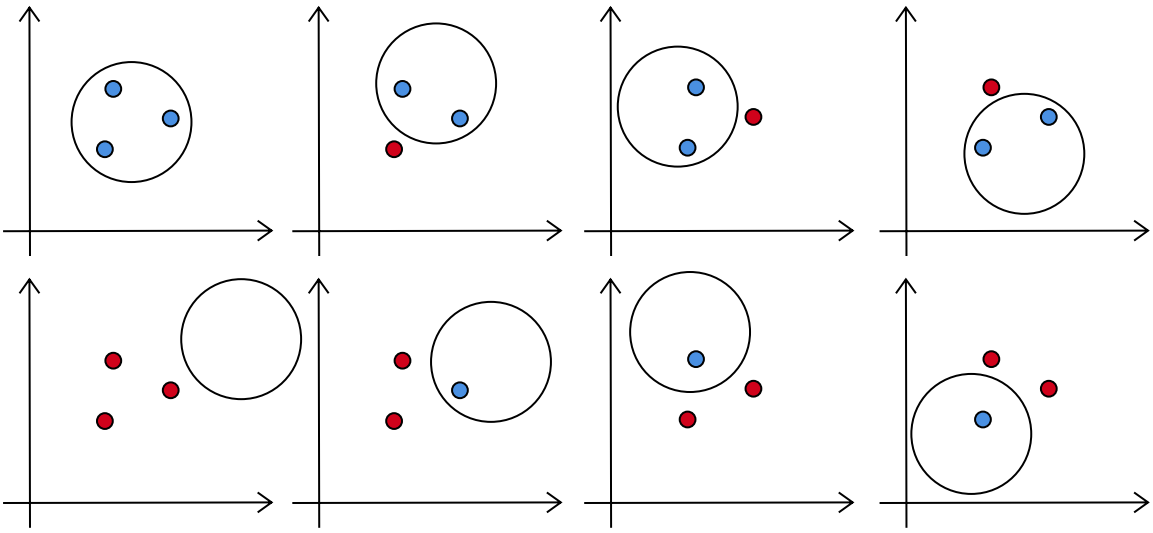
VC-Dimension of Circles

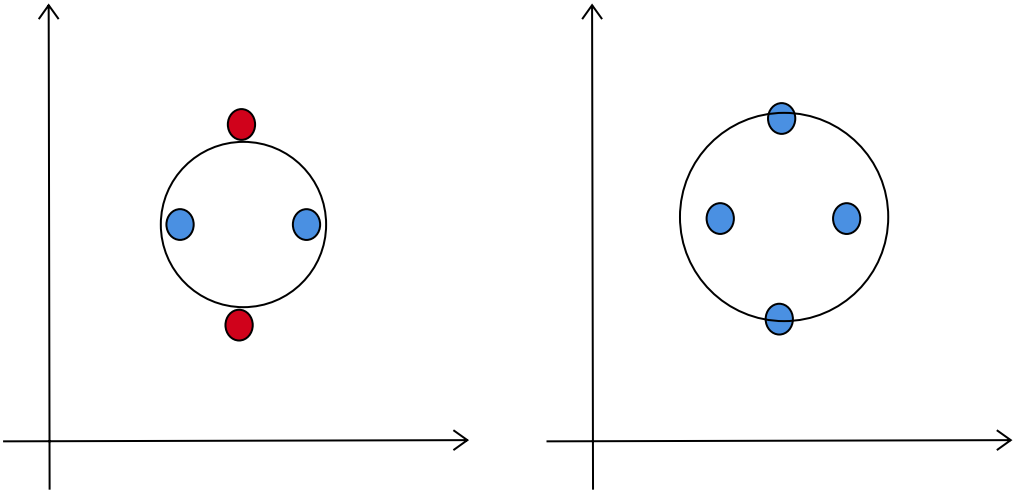
For the first part of the question, we need to determine the VC-Dimension of the class $\mathcal{A}$ of all circles in R^2, without any restrictions on the radius.

$$\mathcal{A} = \{C_{c,r} : \boldsymbol{c} \in \mathbb{R}^2, r \geq 0\}$$

Clearly, one point can always be shattered by a circle, since it's either in or out of it. The same holds for two points. Three points can also be shattered, seen in the following:



However, four points in a plane can not be shattered by a circle. The reason is, that there are two possible arrangements of four points, either as a tetragon or as a triangle with a point in the middle. In both cases, it is impossible to get all possible combinations of pairs of two points:



In the case of the tetragon, there will always be a combination of two points which are opposite of each other, with maximum distance among all points. These points can only be combined in a circle which has a diameter that is bigger or equal the distance between these

two. By defintion, this circle will include the other points with lower distance.
Since the VC dimension is the maximum number of points which can be shattered, the VC dimension of $\mathcal{A}$ is 3.

Now for the special case of all circles with radios one:

$$A_1 = \{C_{c,1} : c \in \mathbb{R}^2\}$$

Again, one and two points can be shattered without a problem. Further, three points can also be shattered, as long their arranged in a triangle that is inscribed in the unit circle. Further the points can't lie on the circle, but have to lie inside. It will be possible to shatter these three points as long as the sides of the triangle are sufficiently short. Again, for four points, the same issue arises as in the general case. So the VC dimension for unit circles in the plane is 3.

Shatter Coefficient of Half Spaces in R^2

Half Spaces through the Origin

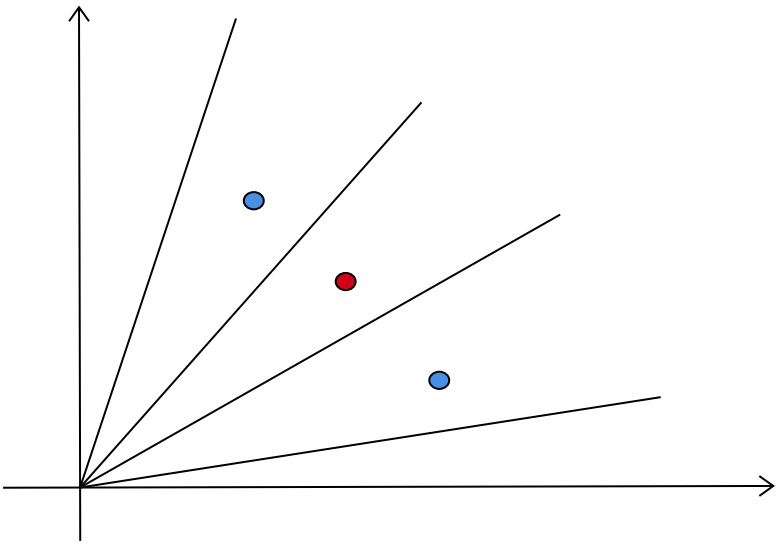
First, we can determine the n-th shatter coefficient for the class of all half spaces containing the origin:

$$\mathcal{A}_0 = \{H_{a,b,0} : a, b \in \mathbb{R}\}$$

where

$$H_{a,b,c} = \{(x, y) \in \mathbb{R}^2 : ax + by \geq c\}$$

In order to derive a general form of the shatter coefficient of $\mathcal{A}$ we need to establish a pattern of the shatter coefficient depending on the number of points. Clearly, one point can be either in or out of the set. Two points can be shattered in 4 sets. So up to $n = 2$ the shatter coefficient is $2n$. In the case of three points, we can't shatter them anymore:



More specifically, there is no set of points in which can shatter them into all possible sets containing two points, like seen above. So in this case the shatter coefficient will be 6, which is again $2n$. Any further point we add can either be in or out of a given set generated by a half space through the origin. So we can conclude, that the shatter coefficient of $\mathcal{A}_0$ in R^2 is

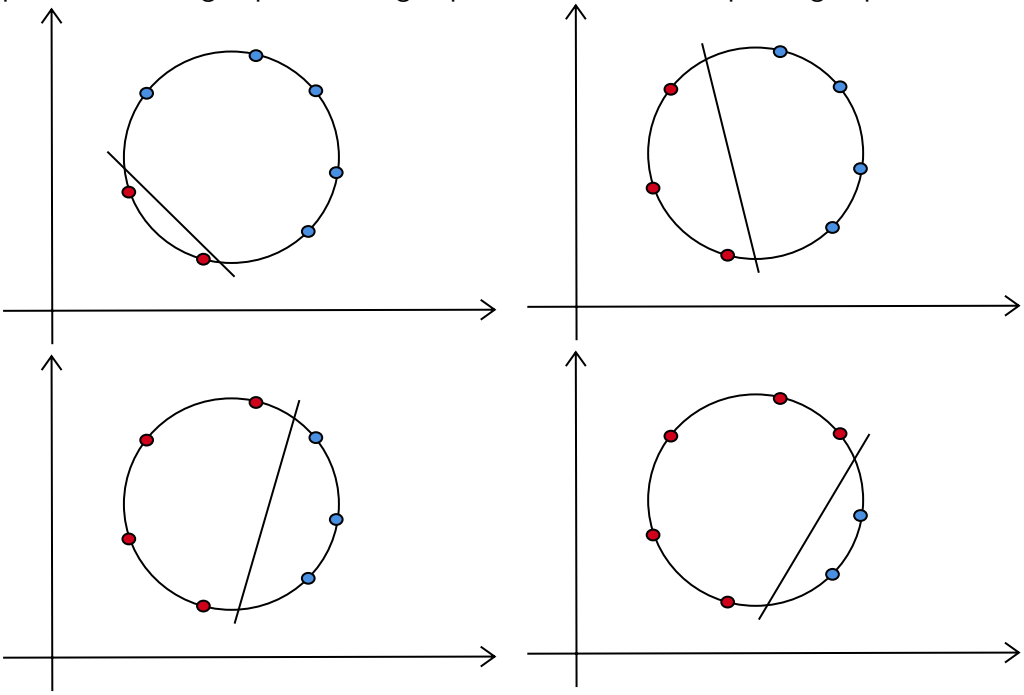
$$S_{\mathcal{A}_0}(n) = 2n$$

General Half Spaces

For establishing the nth-Shatter coefficient for general half spaces $\mathcal{A} = \{H_{a,b,c} : a, b, c \in \mathbb{R}\}$, we observe the following. Arranging all points on a circle in R^2,

- we can always generate sets containing all or none of the points using half-spaces, so the shatter coefficient will be at least two
- it will always be possible to generate sets containing only one point, so the shatter coefficient will be at least 2 + n
- it will be impossible to generate a set containing points opposite of each other without capturing adjacent points
- we can always group points together which are connected through the circle, as long as there is no point inbetween

The last observation leads us to the following result: There will always be a way for each point to be in a group of two, a group of three and so on, up to a group of $n - 1$.



Consequently, the shatter coefficient for the group of all half-spaces will be:

$$S_{\mathcal{A}}(n) = 2 + n(n-1)$$

Rademacher Averages

First Structural Result

$$\begin{aligned} R_n(A \cup B) &= \mathbf{E} \sup_{x \in A \cup B} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i x_i \right| \\ &\leq \mathbf{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right| + \mathbf{E} \sup_{b \in B} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i b_i \right| \\ &= R_n(A) + R_n(B) \end{aligned}$$

Second Structural Result

Note, that c is a constant.

$$\begin{aligned} R_n(cA) &= \mathbf{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i c \cdot a_i \right| \\ &= \mathbf{E} \sup_{a \in A} |c| \cdot \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right| \\ &= |c| \cdot \mathbf{E} \sup_{a \in A} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i a_i \right| \\ &= |c| \cdot R_n(A) \end{aligned}$$

Third Structural Result

$$\begin{aligned} R_n(A \oplus B) &= \mathbb{E} \sup_{a \in \mathcal{A}, b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(a_i + b_i\right) \right| \\ &\leq \mathbb{E} \sup_{a \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i a_i \right| + \mathrm{E} \sup_{b \in \mathcal{B}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| \\ &= R_n(\mathcal{A}) + R_n(\mathcal{B}) \end{aligned}$$

The second step uses the triangle inequality applied to the supremum.

Fourth Structural Result

For the last result we can make use of the previous ones.
First we note that the absolute convex hull of $\mathcal{A}$ contains the set itself.

$$\mathcal{A} \subset \text{absconv}(\mathcal{A})$$

therefore $R_n(\mathcal{A}) \leq R_n(\text{ absconv } \mathcal{A})$
Where $\text{absconv}(A) = \left\{ \sum_{j=1}^{N} c_j a^{(j)} : N \in \mathbb{N}, \sum_{j=1}^{N} |c_j| \leq 1, a^{(j)} \in A \right\}$

Further, for any value of c_j we have

$$R_n\left(c_1\mathcal{A} + \ldots + c_N\mathcal{A}\right) \leq \sum_{i=1}^{N} |c_i|\, R_n(\mathcal{A}) \leq R_n(\mathcal{A})$$

where $\sum_{j=1}^{N} |c_j| = 1$. Since the absolute convex hull is the union of all sets of the form

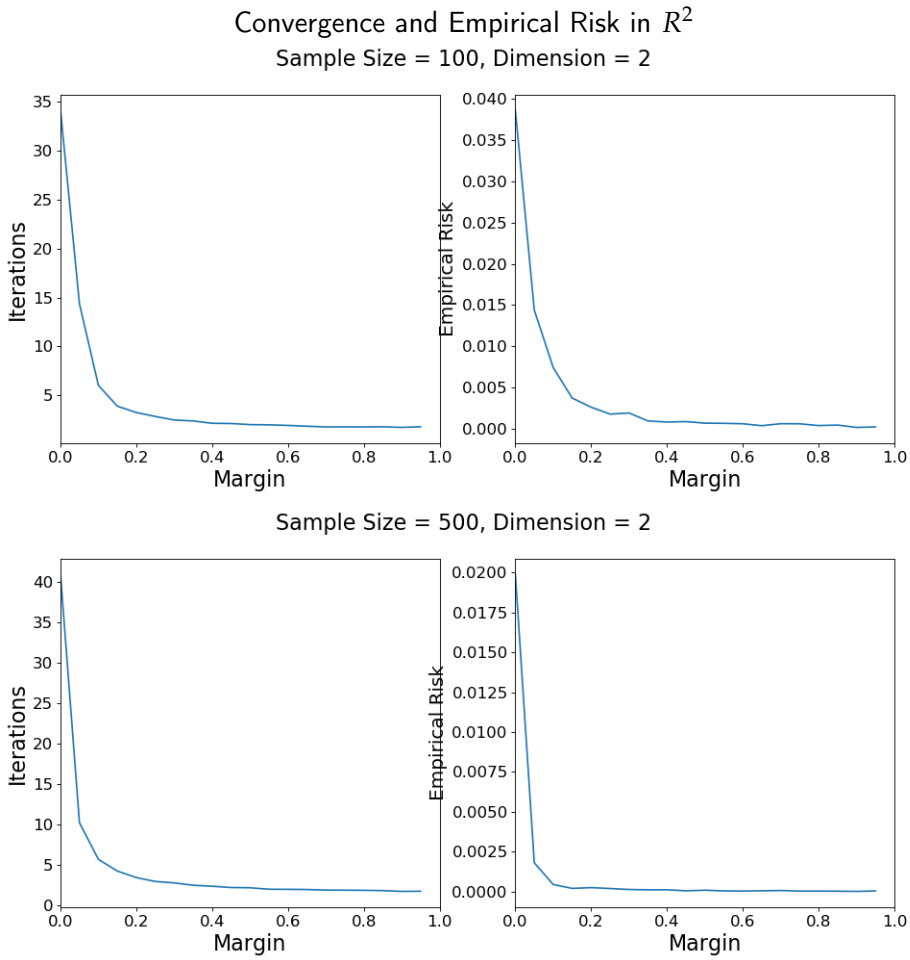
$$c_1\mathcal{A} + \ldots + c_N\mathcal{A} \equiv \{c_1a_1 + \ldots + c_Na_N : a_1, \ldots, a_N \in \mathcal{A}\}$$

Consequently, $R_n(\mathcal{A}) = R_n(\text{absconv}(\mathcal{A}))$.

The Perceptron Algorithm

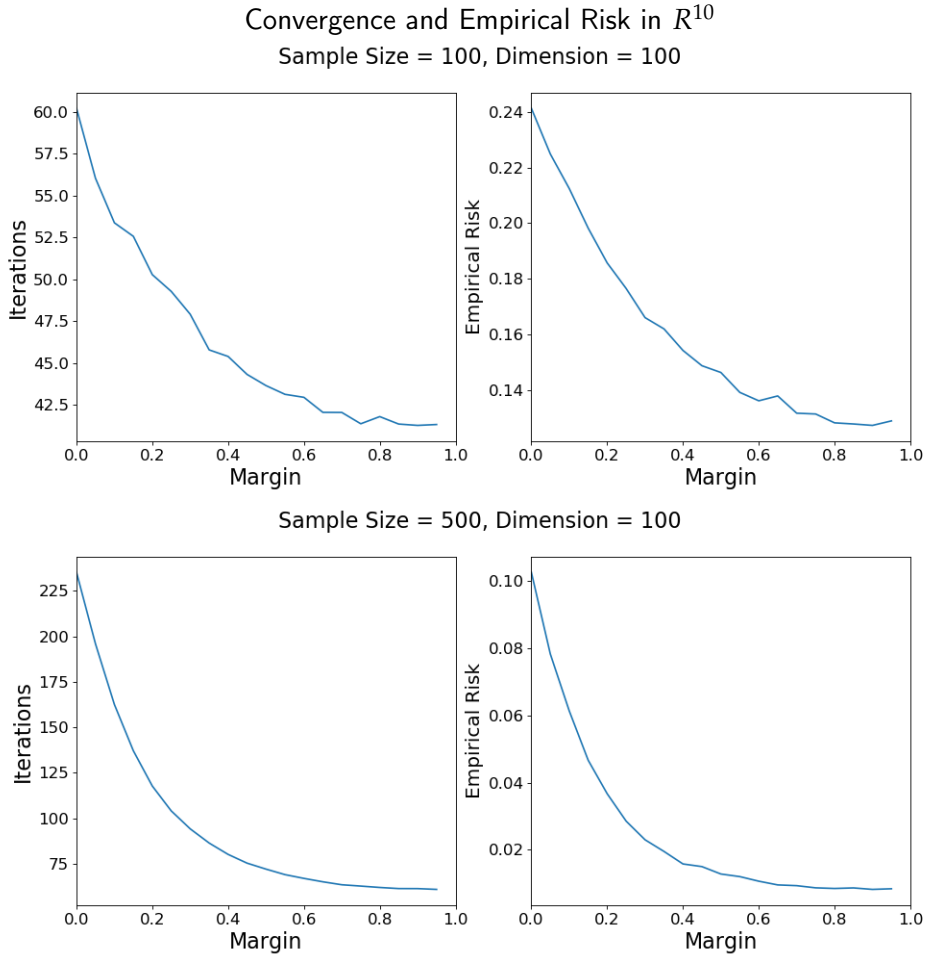
Linearly Seperable Data

As we noted in class, the perceptron algorithm is guaranteed to converge with linearly seperable data. Further, we derived that the algorithm will run for a maximum of $\frac{R^2}{\gamma^2}$ rounds where $R = \max ||x||$ and γ denotes the margin.



The simulations show, that the algorithm converges quickly in R^2 even for margins below 0.2. Further, the estimated probability of error is close to zero, once the margin exceeds 0.1. The

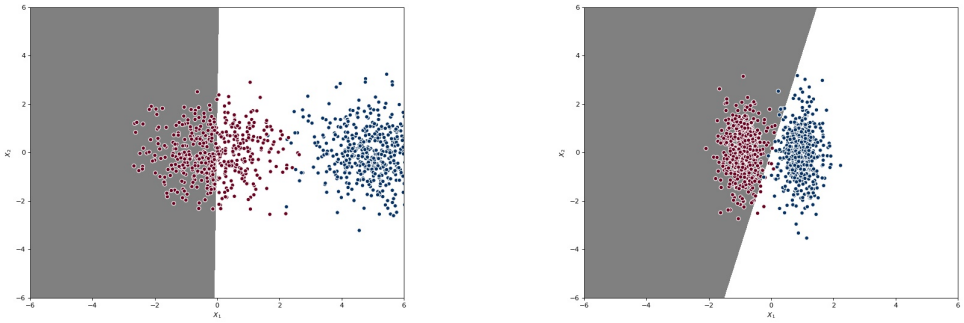
sample size doesn't influence the number of iterations. Larger samples improve the empirical risk for lower margins.
For higher dimensional spaces we have a similar picture, however with a higher number of iterations and empirical risk.



Again, as the margin increases the number of iterations falls. However, the overall number of iterations in R^{10} is much higher than in R^2, for the same margins. Why is that so? Since we keep γ constant, we need to find bounds for the maximum norm R. In the worst case scenario we have one observation which has all entries either 1 or -1 (depending on the class). In that case the maximum norm would be $\sqrt{\sum_{i=1}^{d} x_i^2} = \sqrt{d}$. So the upper bound for the number of iterations will be d/γ. We can clearly see in the simulations, that the perceptron stays below this bound, even for larger margins. The sample size seems to play a role for larger samples, probably since the chances of having higher R values increases. Finally, the probability of error converges to zero, but slower than in the two dimensional case.

Non-Seperable Case

For the non-seperable case we need to include a stopping condition. The condition used in my simulation is a maximum number of iterations of 2000 or the in-sample accuracy doesn't change more than by 10^{-5}. Further, without centering and re-scaling the input data the perceptron algorithm will perform badly, as shown in the following plot:

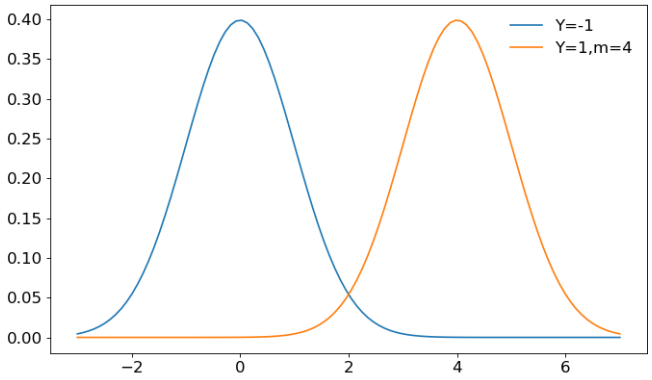


(a) (b)

Figure 1: (a) No Scaling (b) Scaling

Bayes Risk

The Bayes classifier g^* predicts 1 if $\eta(x) \geq 0.5$ and -1 else, with $\eta(x) = f_{Y=1|X}$. We can visualize the error region of the Bayes classifier graphically:



The error region is indicated by the overlap of the two conditional density functions. The expected loss is:

$$E\left[l(g,y)\right] = \mathbb{1}_{g=1}P(Y=-1|X<m/2) + \mathbb{1}_{g=-1}P(Y=1|X>m/2)$$

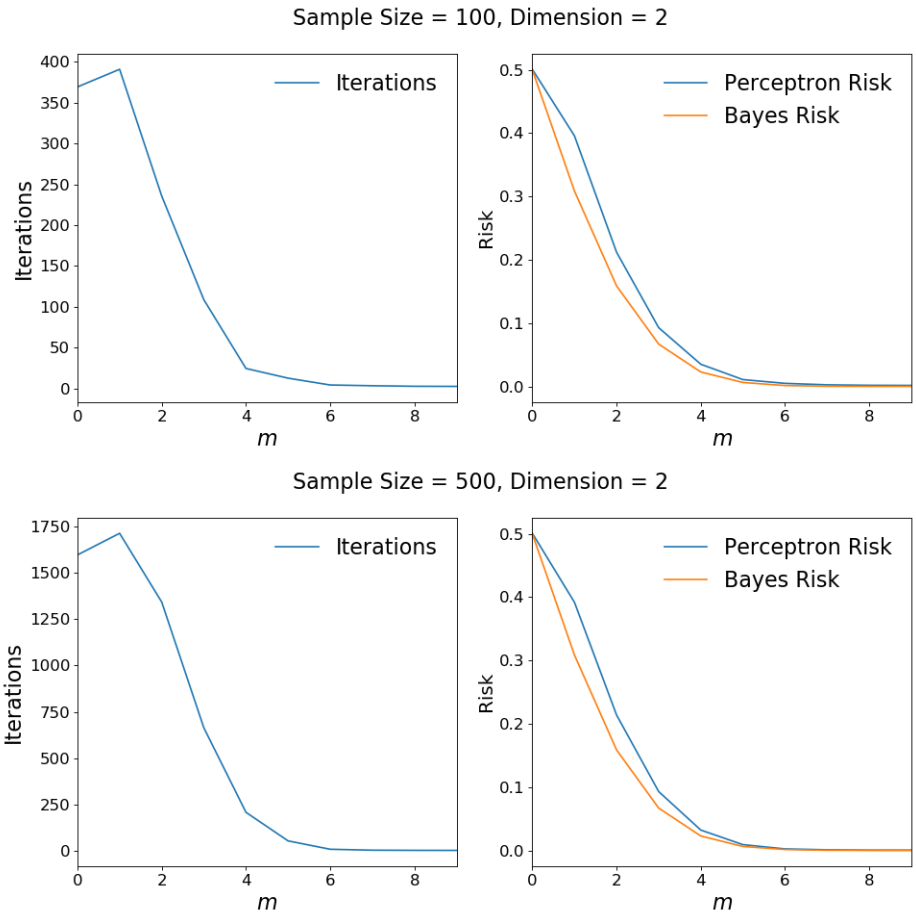
Which leads to a Bayes risk of

$$R(g^*) = \Phi\left(m/2\right)$$

with Φ being the CDF of the normal distribution with mean m and variance one.

Results

Once the data is scaled, the empirical risk convergec towards zero with larger seperation between the two clusters (indicated by m).The number of iterations for low seperation is considerably higher than the iterations in the seperable case. However, as the distance between the clusters increases, the sample data becomes linearly seperable and the algorithm converges quickly.



For larger samples, the algorithm reaches its stopping condition less quickly, while the empirical risk seems similar. Overall, the empirical risk converges towards zero with larger values of m. However, without scaling, the empirical risk will converge towards a value of 0.25.

We can find similar results for higher-dimensional data. Larger sample sizes have a positive influence on the risk, but also lead to a higher number of iterations.