Recap

Our framework is the following

- $\mathcal{X}$ – input space; (as earlier, *Feature space*)
- $\mathcal{Y}$ – output space (as earlier, *Set of class labels*)
- $\mathcal{H}$ – hypothesis space (*family of classifiers*)

  Each $h \in \mathcal{H}$ is a function, $h : \mathcal{X} \to \mathcal{A}$,
  where $\mathcal{A}$ is called *action space*.
- Training data: $\{(X_i, y_i),\ i = 1, \cdots, n\}$
  drawn *iid* according to some distribution P_{xy} on $\mathcal{X} \times \mathcal{Y}$.

Recap

- Loss function: $L : \mathcal{Y} \times \mathcal{A} \to \Re^+$.
- The risk function, $R : \mathcal{H} \to \Re^+$, is given by

$$R(h) = E[L(y, h(X))] = \int L(y, h(X)) dP_{xy}$$

  We assume that L is bounded so that the expectation always exists.
- Let $h^* = \arg\min_{h \in \mathcal{H}} R(h)$
- We define the goal of learning as finding h^*, the global minimizer of risk.

Recap

- However, we cannot directly minimize $R(\cdot)$.
- The **empirical risk function**, $\hat{R}_n : \mathcal{H} \to \Re^+$, is defined by

$$\hat{R}_n(h) = \frac{1}{n} \sum_{i=1}^{n} L(y_i, h(X_i))$$

- Let

$$\hat{h}_n^* = \arg\min_{h \in \mathcal{H}} \hat{R}_n(h)$$

- An algorithm learns $\hat{h}_n^*$ by minimizing empirical risk.

Recap

- Our objective is to find h^*, minimizer of risk $R(\cdot)$.
- Since we do not know R, we minimize $\hat{R}_n$ instead, and thus find $\hat{h}_n^*$.
- Hence we are interested in the question

$$R(\hat{h}_n^*) \to R(h^*)?$$

  where the convergence is in probability.
- This is the issue of consistency of empirical risk minimization.

Recap

- For empirical risk minimization to be effective, we need $R(\hat{h}_n^*)$ to converge in probability to $R(h^*)$:

$$\text{Prob}\left[|R(\hat{h}_n^*) - R(h^*)| > \epsilon\right] \leq \delta, \ \forall n \geq N(\epsilon, \delta)$$

- This will happen if $\hat{R}_n(h)$ converges to $R(h)$ uniformly over $\mathcal{H}$:

$$\text{Prob}\left[\sup_{h \in \mathcal{H}} |\hat{R}_n(h) - R(h)| > \epsilon\right] \leq \delta, \ \forall n \geq N(\epsilon, \delta)$$

- This uniform convergence is necessary and sufficient for consistency of empirical risk minimization.

Recap

- The uniform convergence holds if $\mathcal{H}$ is finite. $\mathcal{H} = \{h_1, \cdots, h_M\}$.
- For calculating the number of examples needed, we used the Hoeffding inequality given by

$$\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \mu\right| > \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

- Using Hoeffding inequality, we showed that

$$\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| > \epsilon\right] \leq 2M\exp(-2n\epsilon^2)$$

- For the case where $\mathcal{H}$ is infinite, our idea is the folowing. .
- Given n training examples, as far as empirical risk is concerned, only finitely many (at most 2^n) functions from $\mathcal{H}$ can be distinguished.
- Hence we can try to employ the argument we used for finite $\mathcal{H}$ case to tackle the general case.

- Suppose we had $2n$ examples and $\hat{R}_n$ and $\hat{R}'_n$ are n-sample estimates from the two halfs of the examples.
- Then it can be shown that

$$\text{Prob}\left[\sup_{h\in\mathcal{H}} |R(h) - \hat{R}_n(h)| > \epsilon\right] \leq$$

$$2\,\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - \hat{R}'_n(h)| > \frac{\epsilon}{2}\right]$$

  (Showing the above is non-trivial)
- By using Hoeffding Inequality we showed

$$\text{Prob}\left[|\hat{R}_n(h) - \hat{R}'_n(h)| > \frac{\epsilon}{2}\right] \leq 4\,\exp\left(-\frac{n\epsilon^2}{8}\right)$$

- The probability that we want to bound is

$$\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - \hat{R}'_n(h)| > \frac{\epsilon}{2}\right]$$

  where the supremum is needed to be taken only over fiitely many functions.
- Let S_{2n} denote the sample of $2n$ examples.
- Then the number of functions that we need to consider can be written as $M(\mathcal{H}, 2n, S_{2n})$.
- It depends on the family $\mathcal{H}$, the number of samples, $2n$ and also on the specific set of examples we have, S_{2n}.

- Let

$$\Pi(\mathcal{H}, m) = \max_{S_m} M(\mathcal{H}, m, S_m)$$

denote the maximum nuber of functions to consider if we have m examples.

- Now we can get a bound on the probability of interest as

$$\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| > \epsilon\right]$$

$$\leq 2\,\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - \hat{R}'_n(h)| > \frac{\epsilon}{2}\right]$$

$$\leq 8\,\exp\left(-\frac{n\epsilon^2}{8}\right)\,\Pi(\mathcal{H}, 2n)$$

- Thus, we finally get a bound that we want as

$$\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| > \epsilon\right]$$

$$\leq 8\, \exp\left(-\frac{n\epsilon^2}{8} + \ln\left(\Pi(\mathcal{H}, 2n)\right)\right)$$

- Whether or not this bound is useful depends on how $\ln(\Pi(\mathcal{H}, m))$ grows with m.
- If the rate of growth of $\ln(\Pi(\mathcal{H}, m))$ is linear in m (or that of $\Pi(\mathcal{H}, m)$ is exponential), then the bound is not useful. Otherwise, it is.

- $\Pi(\mathcal{H}, m)$ is the maximum number of distinguihable functions in $\mathcal{H}$ based on a sample of m points.
- Its maximum possible value is 2^m.
- If for all m, it is 2^m then the bound is not useful.
- The hope is that as m increases, the number of distinguishable functions does not grow exponentially.

VC Dimension of $\mathcal{H}$

- We define the VC dimension of $\mathcal{H}$ as

$$d_{VC}(\mathcal{H}) = \max \{n : \Pi(\mathcal{H}, n) = 2^n\}$$

- We have $\Pi(\mathcal{H}, n) = 2^n$ only till $n \leq d_{VC}(\mathcal{H})$; after that it would be less.
- Note that there may be $\mathcal{H}$ for which $d_{VC}(\mathcal{H})$ may be infinite.

- Suppose our hypothesis space is such that $d_{VC}(\mathcal{H}) = d < \infty$.
- Then we have the following interesting result.
- **Sauer's Lemma**: Let $d_{VC}(\mathcal{H}) = d < \infty$. Then, for all integers m,

$$\Pi(\mathcal{H}, m) = \sum_{i=0}^{d} \left(\begin{array}{c} m \\ i \end{array} \right)$$

Can be proved using induction on m and d.

- **corollary**: Let $d_{VC}(\mathcal{H}) = d < \infty$. Then, for all $m > d$

$$\Pi(\mathcal{H}, m) \leq \left(\frac{em}{d}\right)^d$$

- Note that this means

$$\ln(\Pi(\mathcal{H}, m)) \leq d\left(\ln\left(\frac{m}{d}\right) + 1\right)$$

Proof of Corollary

We have

$$
\begin{aligned}
\Pi(\mathcal{H}, m) &\leq \sum_{i=0}^{d} \binom{m}{i} \\
&\leq \sum_{i=0}^{d} \binom{m}{i} \left(\frac{m}{d}\right)^{d-i} \quad \text{since } m \geq d,\ d \geq i \\
&= \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{d} \binom{m}{i} \left(\frac{d}{m}\right)^{i} 1^{m-i} \\
&\leq \left(\frac{m}{d}\right)^{d} \sum_{i=0}^{m} \binom{m}{i} \left(\frac{d}{m}\right)^{i} 1^{m-i} \\
&= \left(\frac{m}{d}\right)^{d} \left(1 + \frac{d}{m}\right)^{m} \leq \left(\frac{em}{d}\right)^{d}
\end{aligned}
$$

- Let $G_{\mathcal{H}}(m) = \ln(\Pi(\mathcal{H}, m))$.
- Then for any $\mathcal{H}$, we have

$$G_{\mathcal{H}}(m) = \begin{cases} m \ \ln 2 & \text{for} \quad m \leq d_{VC}(\mathcal{H}) \\ d_{VC}(\mathcal{H}) \left(\ln \frac{m}{d_{VC}(\mathcal{H})} + 1 \right) & \text{for} \quad m > d_{VC}(\mathcal{H}) \end{cases}$$

- Thus, if $d_{VC}(\mathcal{H}) < \infty$, then we have a proper bound and consistency of ERM is assured.

- Recall that $\Pi(\mathcal{H}, m)$ is the maximum number of distinguishable functions based on (all possible sets of) m *iid* examples.
- We have that $\Pi(\mathcal{H}, m) = 2^m$ only as long as $m \leq d_{VC}(\mathcal{H})$.
- After that, the growth is linear and hence we can bound the generalization error.
- We can also show that ERM is not consistent if $d_{VC}(\mathcal{H}) = \infty$.

- Let us sum-up the whole argument.
- For empirical risk minimization to be effective, we need $R(\hat{h}_n^*)$ to converge in probability to $R(h^*)$.
- This will happen if $\hat{R}_n(h)$ converges to $R(h)$ uniformly over $\mathcal{H}$. ($\mathcal{H}$ is the family of classifiers over which we are minimizing empirical risk).
- The needed uniform convergence holds if $\mathcal{H}$ has finite VC-dimension.

- Hence, now the question is how do we find VC-dimension of $\mathcal{H}$.
- Before discussing this, we first examine the bound we can get on generalization error, using VC-dimension.
- This allows us to appreciate how VC-dimension of $\mathcal{H}$ captures some notion of the complexity of the learning problem.

- Using the VC-dimension of $\mathcal{H}$, we can bound the generalization error as follows.
- The bound we have is

$$\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| > \epsilon\right]$$

$$\leq 8\ \exp\left(-\frac{n\epsilon^2}{8}\ +\ \ln\left(\Pi(\mathcal{H}, 2n)\right)\right)$$

- Take

$$\epsilon_0 = \sqrt{\frac{8}{n}\left(\ln(\Pi(\mathcal{H}, 2n)) + \ln\frac{8}{\delta}\right)}$$

- Then

$$8\ \exp\left(-\frac{n\epsilon_0^2}{8}\ +\ \ln\left(\Pi(\mathcal{H}, 2n)\right)\right)\ <\ \delta$$

- Then we know that

$$\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| \leq \epsilon_0\right] \geq (1-\delta)$$

- Thus, with probability greater than $(1-\delta)$, for all $h \in \mathcal{H}$,

$$R(h) \leq \hat{R}_n(h) \; + \; \sqrt{\frac{8}{n}\left(\ln(\Pi(\mathcal{H}, 2n)) + \ln\frac{8}{\delta}\right)}$$

- If $d_{VC}(\mathcal{H}) < \infty$, then for sufficiently large m,

$$\ln(\Pi(\mathcal{H}, m)) = d_{VC}(\mathcal{H}) \left(\ln \frac{m}{d_{VC}(\mathcal{H})} + 1 \right)$$

- This gives us

$$R(h) \leq \hat{R}_n(h) + \sqrt{\frac{8}{n} \left(d_{VC}(\mathcal{H}) \left(\ln \frac{2n}{d_{VC}(\mathcal{H})} + 1 \right) + \ln \frac{8}{\delta} \right)}$$

- The term in square-root above goes to zero if $n >> d_{VC}(\mathcal{H})$.

- We note that the actual bound we got is very loose.
- So, numerically the bound may not be very useful.
- However, the form of the bound is very interesting.

- We have, for any h, with probability $> (1 - \delta)$,

$$R(h) \leq \hat{R}_n(h) \; + \; \Omega(\mathcal{H}, n)$$

  where Ω is a 'complexity' term.
- The true risk of h depends both on the 'data error' and the 'model complexity'.
- The complexity term goes to zero as $d_{VC}(\mathcal{H})/n$.
- We need large number of examples to believe the data error.
- 'Large' depends on complexity of $\mathcal{H}$, namely, its VC-dimension

- We minimize empirical risk over $\mathcal{H}$, a chosen family of classifiers.
- If VC-dimension of $\mathcal{H}$ is infinite, then empirical risk minimization is not effective.
- When $d_{VC}(\mathcal{H}) < \infty$, then $\hat{R}_n(h)$ converges to $R(h)$ uniformly over $\mathcal{H}$.
- This ensures that with large number of examples, minimizer of empirical risk would have low true risk also.

- Further, the true risk can be bounded above by empirical risk plus a complexity term that goes to zero as $d_{VC}(\mathcal{H})/n$.
- The higher the VC-dimension the higher is the number of examples needed.
- In this sense, $d_{VC}(\mathcal{H})$ tells you the complexity of learning with $\mathcal{H}$.
- Next, we discuss how to obtain VC-dimension for a family of 2-class classifiers.

- Recall that $\Pi(\mathcal{H}, m)$ is the maximum number of distinguishable functions from $\mathcal{H}$ based on all possible samples of m points from $\mathcal{X}$.
- This number grows as 2^m only till $m \leq d_{VC}(\mathcal{H})$.
- Given any m points in $\mathcal{X}$, there are 2^m ways of labelling them with 0 or 1.
- If there is a set of m points where every such labelling is achieved by some function or the other in $\mathcal{H}$, then for this m, $\Pi(\mathcal{H}, m) = 2^m$

- Now we can redefine VC-dimension of $\mathcal{H}$ using this idea.
- Recall that we are considering 2-class classification problems.
- Hence, every $h \in \mathcal{H}$ is a function, $h : \mathcal{X} \to \{0, 1\}$.

- A set $A \subset \mathcal{X}$ is said to be **shattered** by $\mathcal{H}$ if
  - for every $B \subset A$, there is a $h \in \mathcal{H}$ such that

$$h(x) = 1 \ \forall x \in B \quad \text{and} \quad h(x) = 0 \ \forall x \in (A - B)$$

- Choosing an arbitrary $B \subset A$ is like labelling an arbitray subset of points of A as Class-1 and others Class-0
- If A is an m-point subset of $\mathcal{X}$ that is shattered then for every one of the 2^m possible labellings of points in A, there is a function in $\mathcal{H}$ to realize that labelling.
- Hence we know that $\Pi(\mathcal{H}, m) = 2^m$ (if there is an m-point set that is shattered).

- VC-dimension of $\mathcal{H}$ is the cardinality of the largest shattered subset of $\mathcal{X}$.
- Note that if we have a m-point subset of $\mathcal{X}$ that is shattered by $\mathcal{H}$ then, for every $m' < m$, there is a m'-point subset of $\mathcal{X}$ that is shattered.
- If for every integer m, there is a m-point subset of $\mathcal{X}$ that is shattered by $\mathcal{H}$, then, VC-dimension of $\mathcal{H}$ is infinity.

- If we find at least one m-point subset of $\mathcal{X}$ that is shattered then we can conclude

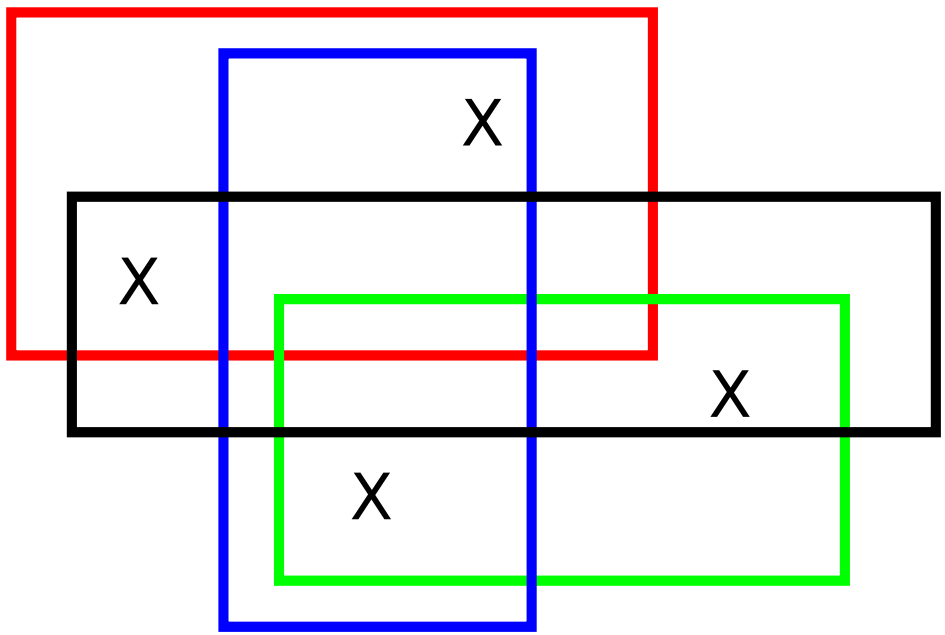
$$d_{VC}(\mathcal{H}) \geq m$$

- Note that there may be other m-point sets that are not shattered.
- To show that $d_{VC}(\mathcal{H}) < m$ we have to show that **no** m-point set is shattered.
- Let us look at some examples of VC-dimension calculation.

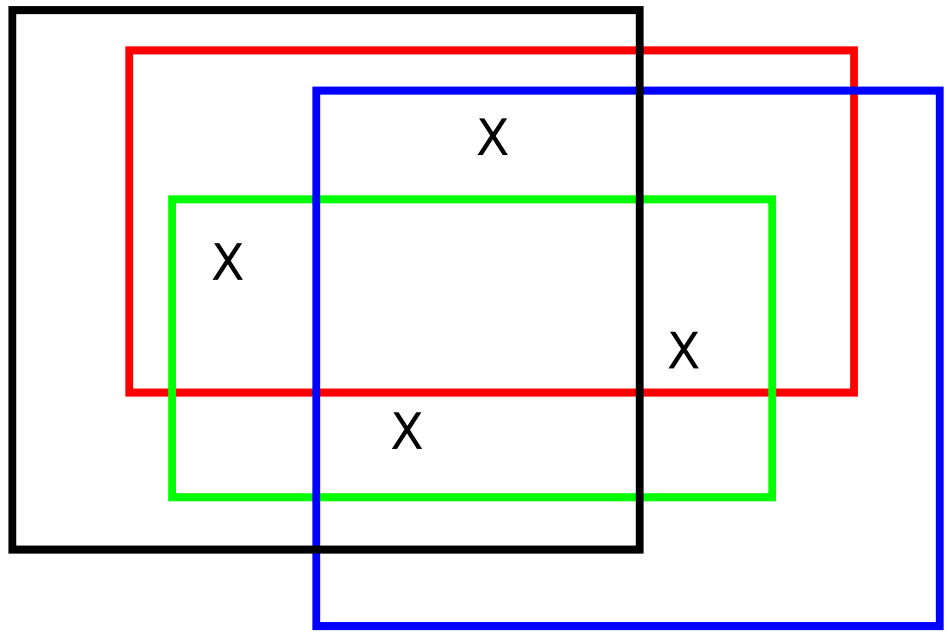
- Let us consider our old example.
- $\mathcal{X} = \Re^2$ and $\mathcal{H}$ is the set of axis parallel rectangles.
- We will show that $d_{VC}(\mathcal{H}) = 4$.
- For this, we have to
  - exhibit at least one 4-point subset of $\Re^2$ that is shattered by the family of axis parallel rectangles; and
  - show that no 5-point set can be shattered.

- Let us say each h here takes value 1 on points inside and on the axis parallel rectangle and it take 0 on points outside.
- To show that a specific 4-point set is shattered we need to essentially show:
  - (i) given any two of the four points there is an axis parallel rectangle that contains those two points but not the other two;
  - (ii) given any three of the four points there is an axis parallel rectangle that contains those three points but not the remaining one.

- Here is a four point set that is shattered.
- We can pick any two of the four points as shown. (We show only 4 out of six possibilities)

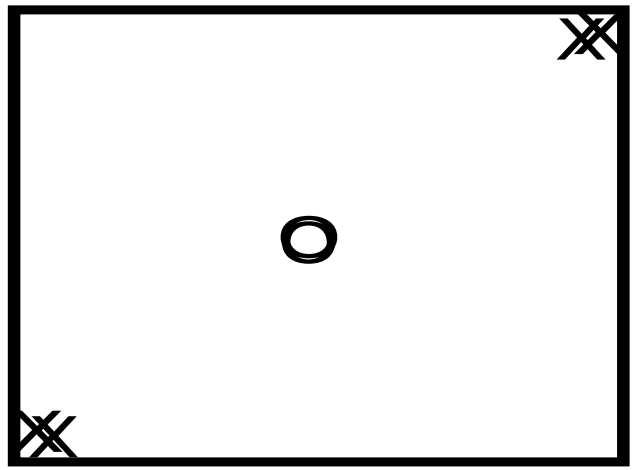


- Similarly, we can pick any three of the four points.



- Though we found one 4-point set that is shattered, there may be other sets that are not shattered.

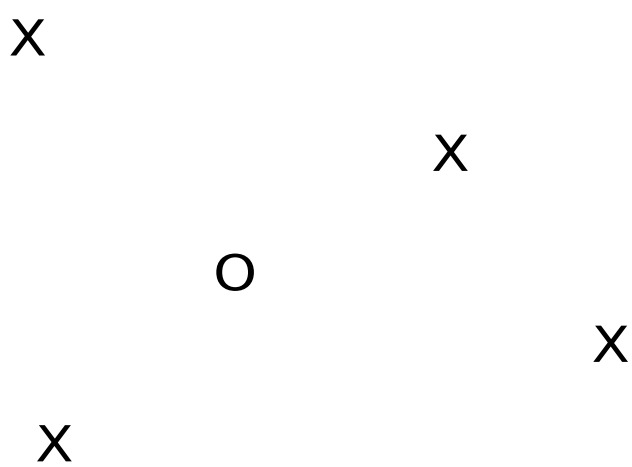
- For example, any 4-point set that contains the following three points can not be shattered.



- In an m point set, if any three are colinear then it can not be shattered.

- Now to show that VC-dimension is 4, we need to show that no 5-point set can be shattered.
- Consider any 5-point set. (All points are distinct).
- Find the max and min along x and y coordinates of these points.
- There has to be a point on or in the interior of the rectangle formed by these max and min values.
- If we label the interior point as negative while the rest as positive, then that labelling cannot be realized.

- Here is an example of a 5-point set with a labelling that can not be realized.



- We have shown that the VC-dimension of the family of axis-parallel rectangles is four.
- We need four parameters to represent this family.
- Every axis-parallel rectangle can be specified by, e.g., the coordinates of bottom left and top right corners.
- Such relationship between VC dimension and the number of parameters needed, often holds.
- But that is not necessarily true always.

- Since the VC-dimension is finite, empirical risk minimization is consistent here.
- If our algorithm returns a classifier that is a global minimizer of empirical risk, then given enough examples, its true risk is also close to the global minimum.
- That is why this class was PAC-learnable.
- In PAC sense (where there is no noise), any algorithm that returns a classifier which classifies all examples correctly is good enough.

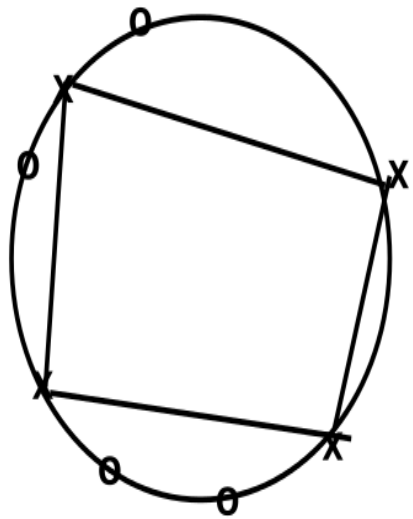
- Now let us consider the extreme example we saw:
- $\mathcal{X} = \Re^2$ and $\mathcal{H} = 2^{\mathcal{X}}$.
- Now for any finite $A \subset \mathcal{X}$, given any $B \subset A$, there is a $h \in \mathcal{H}$ such that $h(x) = 1$ if and only if $x \in B$. (This is the set B itself!)
- Hence, every finite subset is shattered here.
- The VC-dimension is infinite.
- We know that ERM is not consistent here and hence uniform convergence does not hold.
- Thus, in general, when VC-dimension is infinite, ERM is not consistent.

- We can construct other examples of families with infinite VC-dimension.
- Essentially, if the family of classifiers is too 'flexible' then VC-dimension is infinite.
- For example, take $\mathcal{H}$ to be the family of all convex polygons over $\Re^2$.
- Once again, the family is too 'flexible'.
- We can show that VC-dimension is infinite.

- For this we have to show that for every m, we can find a m-point subset of $\Re^2$ that is shattered by $\mathcal{H}$.
- Take the m points on a circle.
- Now, given any labelling of these points by 1 and 0, we have to show that there is a convex polygon such that points labelled 0 are outside it and points labelled 1 are inside or on the polygon.

- Take any arbitrary labelling of the points.
- Now we draw a convex polygon as follows.
- Start with any point labelled 1 as a vertex of the polygon. Then join that point to the next point in the set that is labelled 1.
- Since all points are on a circle, the next point can be defined as the next one in the clockwise direction.
- We continue this process till we reach the starting point.
- Now we have a convex polygon whose verices are points labelled 1 and points labelled 0 are outside it.

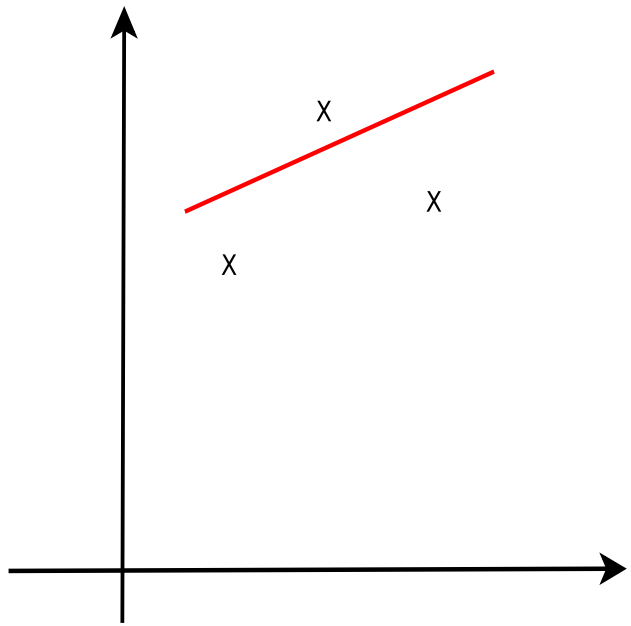
- Here is an example of such a construction



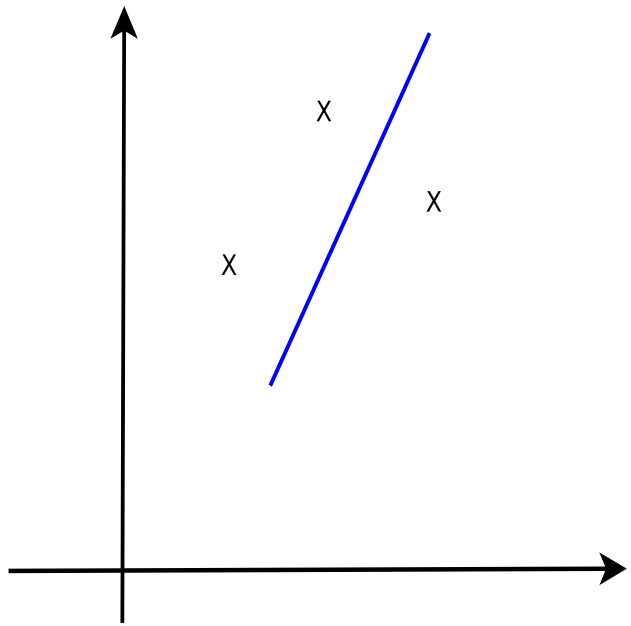
- As we have seen, linear classifiers are an important special case.
- Hence we want VC-dimension of hyperplane classifiers in $\Re^d$.
- For this family the VC-dimension is $d + 1$.
- We illustrate it in $\Re^2$.

- We want to show that VC-dimension of hyperplanes in $\Re^2$ is 3.
- For this we have to show at least one 3-point subset that is shattered.
- Also we have to show that no 4-point set is shattered.

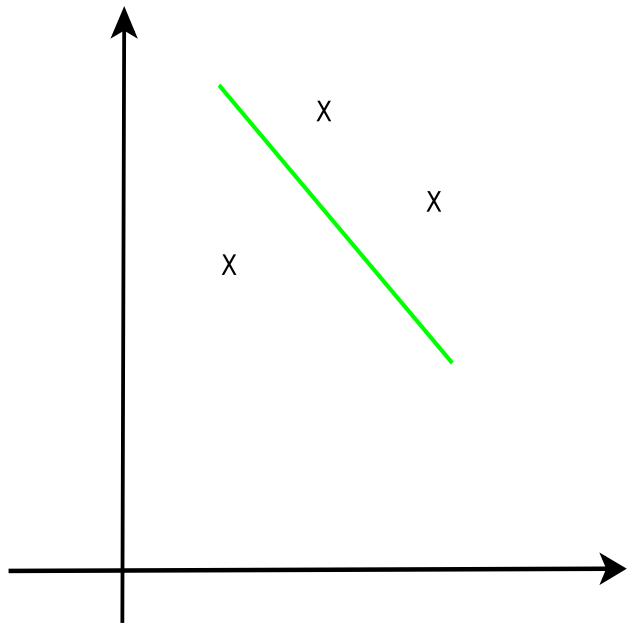
- Here is a 3-point set that is shattered.



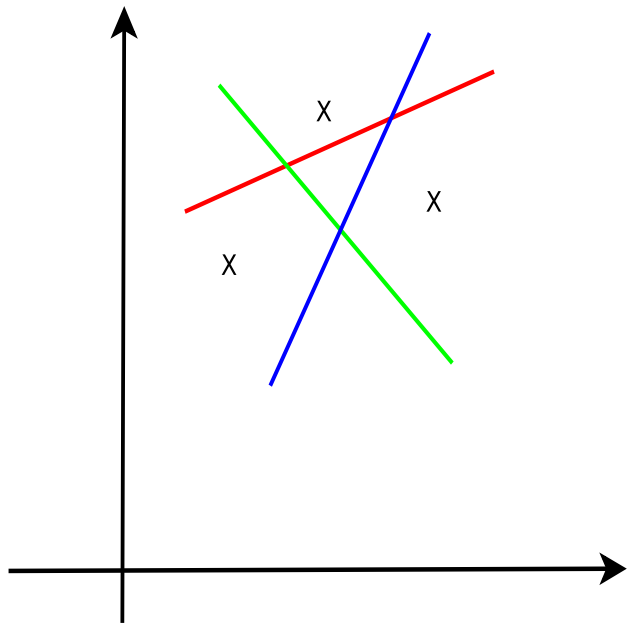
- Here is a 3-point set that is shattered.



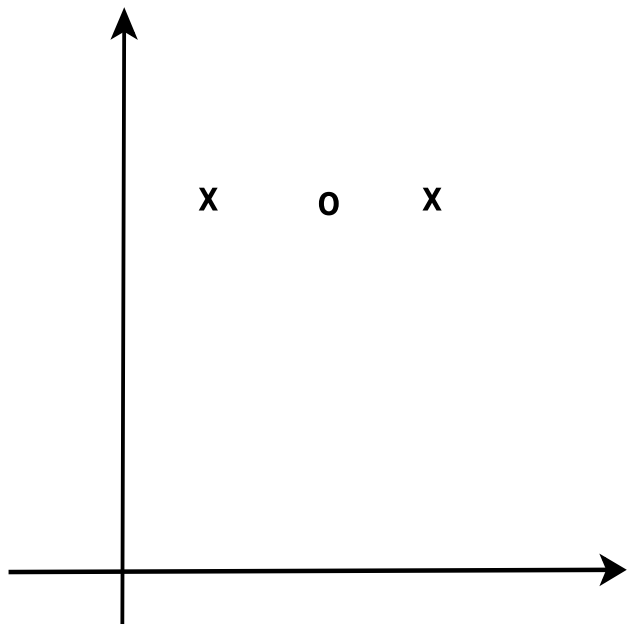
- Here is a 3-point set that is shattered.



- Here is a 3-point set that is shattered.

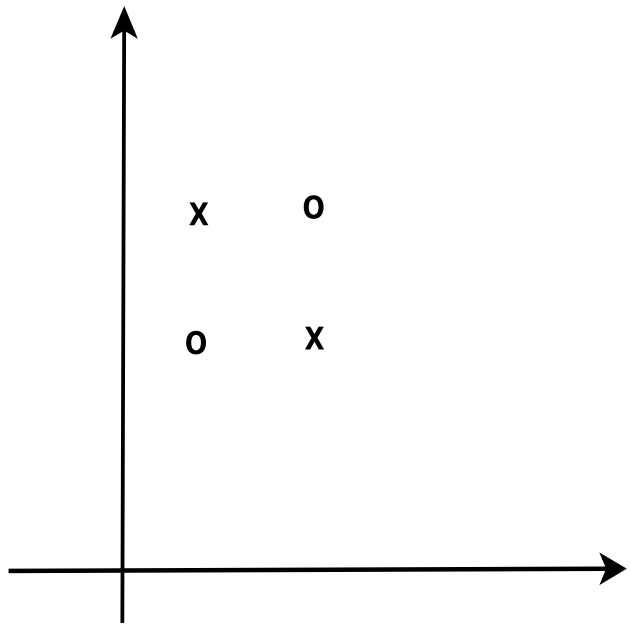


- Note that if the three points are collinear then the set is not shattered.



- Now, to show VC-dimension is 3, we have to show that no 4-point set is shattered.
- If any three of the four points are collinear, then the set is not shattered.
- Suppose no three are collinear. Then the four points form a quadrilateral.
- If we label one pair of opposite vertices by 1 and the other pair by 0, no hyperplane can realize this classification.

- Here is such a generic set.



VC Dimension of Hyperplanes in $\Re^d$ is $d+1$

- Consider hyperplanes in $\Re^d$.
- To show that VC-dimension here is $d+1$ we need to show that there is a set of $d+1$ points that is shattered and that no set of $d+2$ points can be shattered.
- To show a set of m points is shattered by family of hyperplanes, we need to show the following.

  Given any division of the set into two sets of m_1 and $m_2 = m - m_1$ points, we need to show that the two sets of points are linearly separable.

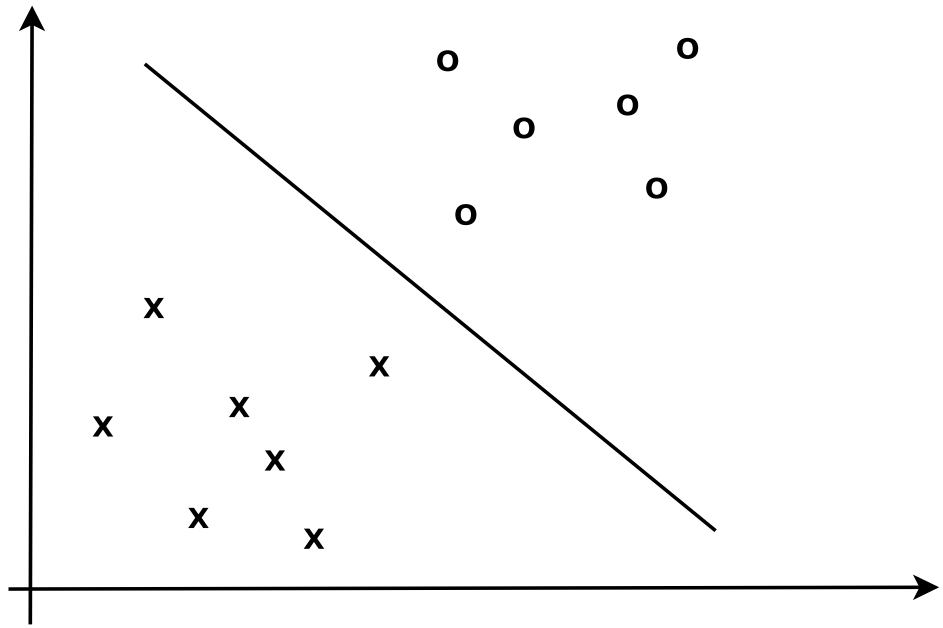
- For the proof we need the notion of a convex hull.
- Given $S = \{x_1, \cdots, x_m\}$, the convex hull of S is

$$\text{Conv}(S) = \left\{ x \,:\, x = \sum_{i=1}^{m} \alpha_i x_i,\ \alpha_i \geq 0,\ \sum_{i=1}^{m} \alpha_i = 1 \right\}$$

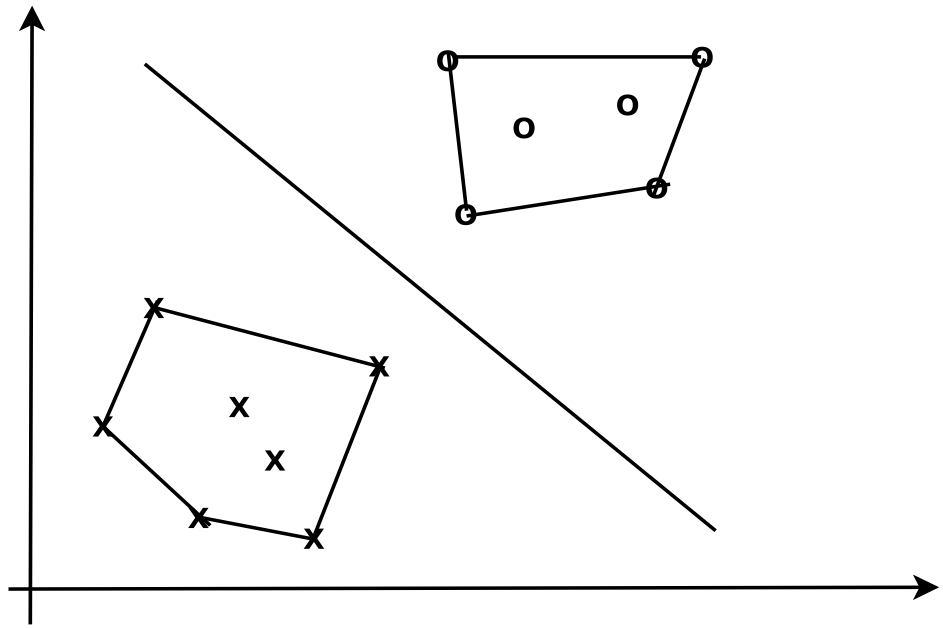
- Convex hull of a set contains all points that can be written as convex combination of points in S.

- Let $S_1 = \{x_1, \cdots, x_{m_1}\}$ and $S_2 = \{y_1, \cdots, y_{m_2}\}$ be two sets of points in $\Re^d$.
- They are linearly separable if and only if $\text{Conv}(S_1) \cap \text{Conv}(S_2) = \phi$.
- Two sets of points in $\Re^d$ are linearly separable if and only if their convex hulls do not intersect.
- This is easy to show.

- Here is a simple illustration of this.



- Here is a simple illustration of this.



- **Theorem**: Given m points in $\Re^d$. Take one of them as origin. The set of m points is shattered if and only if remaining $m-1$ points are linearly independent.
- For the first part we need to show:
  linearly independent $\Rightarrow$ shattered.
- Let $S = \{0, x_1, \cdots, x_{m-1}\}$ be the set.
(Here, 0 is the origin or zero vector in $\Re^d$).
- We are given that the $(m-1)$ points are linearly independent:
  we cannot have $\sum \gamma_i x_i = 0$ unless all γ_i are zero.

- Suppose the set is not shattered.
- That means, there is a way of dividing S into two sets S_1 and S_2 such that they are not linearly separable.
- Hence, $\text{Conv}(S_1) \cap \text{Conv}(S_2) \neq \phi$. This implies

$$\exists \alpha_i, \beta_i \geq 0, \ \sum \alpha_i = \sum \beta_i = 1, \ s.t. \ \sum_{x_i \in S_1} \alpha_i x_i = \sum_{x_i \in S_2} \beta_i x_i$$

- This means, there exist γ_i such that

$$\sum_{i=1}^{m-1} \gamma_i x_i = 0, \quad \text{where} \quad \gamma_i = \begin{cases} \alpha_i & \text{if } x_i \in S_1 \\ -\beta_i & \text{if } x_i \in S_2 \end{cases}$$

- This contradicts the fact the the $(m-1)$ points are linearly independent.
- This completes the proof for:
  linearly independent $\Rightarrow$ shattered.

- This means, there exist γ_i such that

$$\sum_{i=1}^{m-1} \gamma_i x_i = 0$$

- This contradicts the fact the the $(m-1)$ points are linearly independent.
- This completes the proof for

  linearly independent $\Rightarrow$ shattered.

- Now we have to show:
  shattered $\Rightarrow$ linearly independent.
- Like earlier, we use the contra positive form.
- That is, we show:
  not linearly independent $\Rightarrow$ not shattered.
- Now we are given that there are scalars $\alpha_i,\ i = 1, \cdots, m-1$ such that

$$\sum_{i=1}^{m-1} \alpha_i x_i = 0$$

- In a linear combination, the scalars can take any positive or negative value.
- First we prove for the simpler special case where all α_i are of the same sign.
- Then we consider the general case where some α_i are positive while others are negative.

- When all α_i are of same sign, we have

$$\sum_{i=1}^{m-1} |\alpha_i| x_i = 0$$

- Let

$$\gamma_i = \frac{|\alpha_i|}{\sum_j |\alpha_j|}$$

.

- Note that $\gamma_i \geq 0$ and $\sum_{i=1}^{m-1} \gamma_i = 1$.

- Now we have
$$\sum_{i=1}^{m-1} \gamma_i x_i = 0$$
- This means that the zero vector is in the convex hull of the rest of the points.
- If we take $S_1 = \{x_1, \cdots, x_{m-1}\}$ and $S_2 = \{0\}$, then, convex hulls of S_1 and S_2 intersect and hence we can not linearly separate them.
- Hence S is not shattered.

- Now we consider the general case.
- Let $I_1 = \{i \ : \ \alpha_i \geq 0\}$ and $I_2 = \{i \ : \ \alpha_i < 0\}$.
- Define $\beta_i = \alpha_i, \ \forall i \in I_1$ and $\gamma_i = -\alpha_i, \ \forall i \in I_2$.
- Note that $\beta_i, \gamma_i \geq 0$.
- Now, what we have is

$$\sum_{i \in I_1} \beta_i x_i \ = \ \sum_{j \in I_2} \gamma_j x_j$$

- Let $\sum_{i \in I_1} \beta_i = Z$ and $\sum_{j \in I_2} \gamma_j = Z'$.
- Without loss of generality, assume $Z \geq Z'$.
- Now we can rewrite the earlier equation as

$$\sum_{i \in I_1} \frac{\beta_i}{Z} x_i = \sum_{j \in I_2} \frac{\gamma_j}{Z} x_j = \sum_{j \in I_2} \frac{\gamma_j}{Z} x_j + \frac{Z - Z'}{Z} 0$$

- Note that

$$\sum_{i \in I_1} \frac{\beta_i}{Z} = 1 \quad \text{and} \quad \sum_{j \in I_2} \frac{\gamma_j}{Z} + \frac{Z - Z'}{Z} = 1$$

- Let $S_1 = \{x_i \ : \ i \in I_1\}$ and $S_2 = \{x_i \ : \ i \in I_2\} \cup \{0\}$.
- Then convex hulls of S_1 and S_2 intersect and hence S_1 and S_2 are not linearly separable.
- This shows S is not shattered.
- This completes proof of the theorem

- Shattering of a set of points by hyperplanes does not depend on coordinate origin.
- In $\Re^d$ we can have d linearly independent points. These along with origin gives us a set of $d+1$ points that is shattered.
- Given any set of $d+2$ points, if we take one of them as origin, the rest of $d+1$ points would be linearly dependent.
- Hence no set of $d+2$ points is shattered.
- Hence VC dimension of hyperplanes in $\Re^d$ is $d+1$.
- It takes $d+1$ parameters to represent the family of hyperplanes in $\Re^d$.

Summary

- We want minimizer of empirical risk to have risk that is close to global minimum of risk.
- Empirical risk minimization is consistent if $R(\hat{h}_n^*) \xrightarrow{P} R(h^*)$.
- This holds iff $\hat{R}_n(h)$ converges to $R(h)$ uniformly over $\mathcal{H}$.
- The uniform convergence holds if and only if the VC-dimension of $\mathcal{H}$ is finite.
- The VC-dimension also gives us an idea of the complexity of the family $\mathcal{H}$.
- If the VC-dimension is high then we need correspondingly larger number of examples to have confidence that low empirical risk means low true risk.

- We have considered only the case of 2-class classifiers.
- That is, we considered the case with $h : \mathcal{X} \to \{0, 1\}$.
- Only for this case we bounded the generalization error and defined VC-dimension.
- The risk minimization framework, as we saw, is more general.
- All this can be extended to family of real-valued functions over $\mathcal{X}$ also.