Recap – The PAC learning framework

A Learning problem is defined by giving:

(i) $\mathcal{X}$ – input space; (*feature space*, often $\Re^d$)

(ii) $\mathcal{Y} = \{0,\ 1\}$ – output space (*set of class labels*)

(iii) $\mathcal{C} \subset 2^{\mathcal{X}}$ – concept space (*family of classifiers*)
Each $C \in \mathcal{C}$ can also be viewed as a function
$C : \mathcal{X} \to \{0,\ 1\}$, with $C(X) = 1$ iff $X \in C$.

(iv) $S = \{(X_i, y_i),\ i = 1, \cdots, n\}$ – the set of examples, where X_i are drawn *iid* according to some distribution P_x on $\mathcal{X}$ and $y_i = C^*(X_i)$ for some $C^* \in \mathcal{C}$. C^* is called target concept.

Recap

- The learning algorithm knows $\mathcal{X}, \mathcal{Y}, \mathcal{C}$ but it does not know C^*.
- It also does not know P_x.
- Given n examples, the learning algorithm searches over $\mathcal{C}$ and outputs a concept C_n.

Recap

- We define **error** of C_n by

$$\begin{aligned} \mathrm{err}(C_n) &= P_x(C_n \Delta C^*) \\ &= \mathrm{Prob}[\{X \in \mathcal{X} : C_n(X) \neq C^*(X)\}] \end{aligned}$$

- The $\mathrm{err}(C_n)$ is the probability that on a random sample, drawn according to P_x, the classification of C_n and C^* differ.

Recap

- We say a learning algorithm **Probably Approximately Correctly** (PAC) learns a concept class $\mathcal{C}$ if given any $\epsilon, \delta > 0$, $\exists N(\epsilon, \delta) < \infty$ such that

$$\text{Prob}[\text{err}(C_n) > \epsilon] < \delta$$

  for all $n > N(\epsilon, \delta)$ and for any distribution P_x and any C^*.
- The probability above is with respect to the distribution of n-tuples of *iid* samples drawn according to P_x on $\mathcal{X}$.
- The P_x is arbitrary. But, for testing and training the distribution is same – 'fair' to the algorithm.

Recap

- PAC learnability deals with ideal learning situations.
- We can generalize it.

Recap – Risk Minimization framework

In our new framework we are given

- $\mathcal{X}$ – 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 $\mathcal{X} \times \mathcal{Y}$.

Some Comments

- We have replaced $\mathcal{C}$ with $\mathcal{H}$.
- If we take $\mathcal{A} = \mathcal{Y}$ then it is same as earlier.
- But the freedom in choosing $\mathcal{A}$ allows for taking care of many situations.

- Now we draw examples from $\mathcal{X} \times \mathcal{Y}$ according to P_{xy}. This allows for 'noise' in the training data.
- For example, when class conditional densities overlap, same X can come from different classses with different probabilities.
- We can always factorize $P_{xy} = P_x P_{y|x}$. In the earlier PAC framework, $P_{y|x}$ is a degenerate distribution.

- As before, the learning machine outputs a hypothesis, $h_n \in \mathcal{H}$, given the training data consisting of n examples.
- However, now there is no notion of a target concept/hypothesis.
- There may be no $h \in \mathcal{H}$ which is consistent with all examples.
- Hence we use the idea of loss functions to define the goal of learning.

Recap – Loss function

- Loss function: $L : \mathcal{Y} \times \mathcal{A} \to \Re^{+}$.
- $L(y, h(X))$ is the 'loss' suffered by $h \in \mathcal{H}$ on a (random) sample (X, y).
- By convention we assume that the loss function is non-negative.

Recap – Risk Function

- Define the **risk** function, $R : \mathcal{H} \to \Re^+$, by

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

- Risk is expectation of loss where expectation is with respect to P_{xy}.
- We want to find h with low risk.

Recap – Risk Minimization

- Let
$$h^* = \arg \min_{h \in \mathcal{H}} R(h)$$
- We define the goal of learning as finding h^*, the global minimizer of risk.
- Risk minimization is a very general strategy adopted by most machine learning algorithms.
- Note that we may not have any knowledge of P_{xy}.
- Minimization of $R(\cdot)$ directly is not feasible.

Recap – Empirical Risk function

- Define the **empirical risk function**, $\hat{R}_n : \mathcal{H} \to \Re^+$, by

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

  This is the sample mean estimator of risk obtained from n *iid* samples.
- Let $\hat{h}_n^*$ be the global minimizer of empirical risk, $\hat{R}_n$.

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

Recap – Empirical Risk Minimization

- Given any h we can calculate $\hat{R}_n(h)$.
- Hence, we can (in principle) find $\hat{h}_n^*$ by optimization methods.
- Approximating h^* by $\hat{h}_n^*$ is the basic idea of empirical risk minimization strategy.
- Used in most ML algorithms.

- Is $\hat{h}_n^*$ a good approximator of h^*, the minimizer of true risk (for large n)?
- This is the question of **consistency of empirical risk minimization**.
- Thus, we can say a learning problem has two parts.
  - The optimization part: find $\hat{h}_n^*$, the minimizer of $\hat{R}_n$.
  - The statistical part: Is $\hat{h}_n^*$ a good approximator of h^*.

- Note that the loss function is chosen by us; it is part of the specification of the learning problem.
- The loss function is intended to capture how we would like to evaluate performance of the classifier and hence the goal of learning.
- We look at a few loss functions in the 2-class case.

The 0–1 loss function

- Let $\mathcal{Y} = \{0,\ 1\}$ and $\mathcal{A} = \mathcal{Y}$.
- Now, the 0–1 loss function is defined by

$$L(y, h(X)) = I_{[y \neq h(X)]}$$

where $I_{[A]}$ denotes indicator of event A.

- The 0-1 loss function is

$$L(y, h(X)) = I_{[y \neq h(X)]}$$

- Risk is expectation of loss.
- Hence, $R(h) = \text{Prob}[y \neq h(X)]$;
  the risk is probability of misclassification.
- So, h^* minimizes probability of misclassification.
  (Bayes classifier)

- Here we assumed that the learning algorithm searches over a class of binary-valued functions on $\mathcal{X}$.
- We can extend this to, e.g., discriminant function learning.
- We take $\mathcal{Y} = \{+1, -1\}$ and $\mathcal{A} = \Re$ (now $h(X)$ is a discriminant function).
- We can define the 0-1 loss now as

$$L(y, h(X)) = I_{[y \neq sgn(h(X))]}$$

- Having any fixed misclassification costs is essentially same as 0–1 loss.
- Even if we take $\mathcal{A} = \Re$, the 0–1 loss compares only sign of $h(x)$ with y. The magnitude of $h(x)$ has no effect on the loss.
- Here, we can not trade 'good' performance on some data with 'bad' performance on others.
- This makes 0–1 loss function more robust to noise in classification labels.

- While 0–1 loss is an intuitively appealing performance measure, minimizing empirical risk here is hard.
- The 0–1 loss function is non-differentiable which makes the empirical risk function also non-differentiable.
- Hence many other loss functions are often used in Machine Learning.

Squared error loss

- The squared error loss function is defined by

$$L(y, h(X)) = (y - h(X))^2$$

- As is easy to see, the linear least squares method is empirical risk minimization with squared error loss function.
- Here we can take $\mathcal{Y}$ as $\{+1, \ -1\}$ and $\mathcal{A} = \Re$ so that each h is a discriminant function.
- As we know, we can use this for regression problems also and then we take $\mathcal{Y} = \Re$.

- Another interesting scenario here is to take $\mathcal{Y} = \{0,\ 1\}$ and $\mathcal{A} = [0,\ 1]$.
- Then each h can be interpreted as a posterior probability (of class-1) function.
- As we know, the minimizer of expectation of squared error loss (the risk here) is the posterior probability function.
- So, risk minimization would now look for a function in $\mathcal{H}$ that is a good approximation for the posterior probability function.

- The empirical risk minimization under squared error loss is a convex optimization problem for linear models (when h is linear in its parameters).
- The squared error loss is extensively used in many learning algorithms.

soft margin loss or hinge loss

- Take $\mathcal{Y} = \{+1,\ -1\}$ and $\mathcal{A} = \Re$. The loss function is given by

$$L(y, h(X)) = \max(0,\ 1 - yh(X))$$

- Here, if $yh(X) > 0$ then classification is correct and if $yh(X) \geq 1$, loss is zero.
- This also results in convex optimization for empirical risk minimization.

Margin Losses

- All three losses we mentioned can be written as function of $yh(X)$ by taking $\mathcal{Y} = \{-1,\ +1\}$.
- The 0–1 loss :

$$L(y, h(X)) = \mathsf{sign}(-yh(X))$$

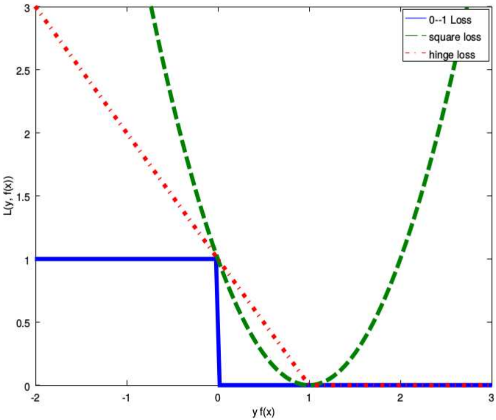
- The squared error loss:

$$L(y,\ h(X)) = (y - h(X))^2 = (1 - yh(X))^2$$

- The hinge loss (used in SVM):

$$L(y, h(X)) = \max(0,\ 1 - yh(X))$$

Plot of 2-class loss functions



- We can think of the other losses as convex approximations of 0–1 loss.

- As we saw, there are many different loss functions one can think of.
- Many of them also make the empirical risk minimization problem efficiently solvable.
- We consider many such algorithms in this course.
- Now, let us get back to the statistical question that we started with.

Consistency of Empirical Risk Minimization

- Our objective is to find h^*, minimizer of risk $R(\cdot)$.
- We minimize the empirical risk, $\hat{R}_n$, and thus find $\hat{h}_n^*$.
- We want h^* and $\hat{h}_n^*$ to be 'close'.
- More precisely we are interested in the question: Does

$$\forall \delta > 0,\ \text{Prob}[|R(\hat{h}_n^*) - R(h^*)| > \delta] \to 0,\ \text{as } n \to \infty?$$

- Same as asking whether $R(\hat{h}_n^*)$ converges in probability to $R(h^*)$

- What is the intuitive reason for using empirical risk minimization?
- Sample mean is a good estimator and hence, with large n, $\hat{R}_n(h)$ converges to $R(h)$, for any $h \in \mathcal{H}$.
- This is (weak) law of large numbers.
- But this does not necessarily mean $R(\hat{h}_n^*)$ converges to $R(h^*)$.
- Let us consider a specific scenario to appreciate this.

- We take $\mathcal{A} = \mathcal{Y} = \{0,\ 1\}$. We use 0–1 loss.
- Suppose the examples are drawn according to P_x on $\mathcal{X}$ and classified according to a $\tilde{h} \in \mathcal{H}$.
- That is, $P_{xy} = P_x P_{y|x}$ and $P_{y|x}$ is a degenerate distribution.
- Now the global minimum of risk $R(\tilde{h}) = 0$.
- We are in the earlier PAC learning framework

- Now, under 0–1 loss, the global minimum of empirical risk is also zero.
- For any n, there may be many h (other than $\tilde{h}$) with $\hat{R}_n(h) = 0$.
- Hence our optimization algorithm can only use some general rule to output one such hypothesis.

- Consider $h_1 : \mathcal{X} \to \mathcal{Y}$ with $h_1(X_i) = y_i,\ (X_i, y_i) \in S$ and $h_1(X) = 1$ for all other X
- Then $\hat{R}_n(h_1) = 0$! It is a global minimizer of empirical risk. But it is obvious that h_1 is not a good classifier.
- Such h_1 may or may not be there in $\mathcal{H}$.
- But, e.g., if we take $\mathcal{H}$ to be all possible classifiers, such h_1 would be in it.
- This is same as the example we considered earlier.
- Thus, here, $R(\hat{h}_n^*)$ will not converge to $R(h^*)$.
- Note that the law of large numbers still implies that $\hat{R}_n(h)$ converges to $R(h)$, $\forall h$.

- If functions like h_1 are in our $\mathcal{H}$ then empirical risk minimization (ERM) may not yield good classifiers.
- If $\mathcal{H}$ contains all possible functions, then this is certainly the case as we saw in our example.
- Functions like h_1 could be non-smooth and hence one possible way is to impose some smoothness conditions on the learnt function (e.g., regularization).
- Issue of consistency depends on $\mathcal{H}$, the class of functions over which we minimize empirical risk.
- Hence, the question is: for what $\mathcal{H}$ is empirical risk minimization consistent.

Consistency of Empirical Risk Minimization

- We would like the algorithm to satisfy: $\forall \epsilon, \delta > 0$, $\exists N < \infty$, such that

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

- In addition, we would also like to have

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

  We would like to (approximately) know the true risk of the learnt classifier.

- For what kind of $\mathcal{H}$ do these hold?

- As we already saw, the law of large numbers (that $\hat{R}_n(h) \to R(h),\ \forall h$) is not enough.
- As it turns out, what we need is that the convergence under law of large numbers be **uniform** over $\mathcal{H}$.
- Such uniform convergence is necessary and sufficient for consistency of empirical risk minimization.

- Law of large numbers says that sample mean converges to expectation of the random variable.
- Given any h, $\forall \epsilon, \delta > 0$, $\exists N < \infty$ such that

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

- The N that exists can depend on ϵ, δ and **also on** h.
- The convergence is said to be uniform if the N depends only on ϵ, δ and not on h.
- That is, for a given ϵ, δ the same $N(\epsilon, \delta)$ works for all $h \in \mathcal{H}$.

- To sum up, $\hat{R}_n(h)$ converges (in probability) to $R(h)$ uniformly over $\mathcal{H}$ if $\forall \epsilon, \delta > 0$, $\exists N(\epsilon, \delta) < \infty$ such that

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

- It is easy to show that uniform convergence is sufficient for consistency of empirical risk minimization.

We have

$$\begin{aligned} R(\hat{h}_n^*) - R(h^*) &= [R(\hat{h}_n^*) - \hat{R}_n(\hat{h}_n^*)] + \\ &\quad [\hat{R}_n(\hat{h}_n^*) - \hat{R}_n(h^*)] + [\hat{R}_n(h^*) - R(h^*)] \\ &\leq [R(\hat{h}_n^*) - \hat{R}_n(\hat{h}_n^*)] + [\hat{R}_n(h^*) - R(h^*)] \end{aligned}$$

($\hat{R}_n(\hat{h}_n^*) - \hat{R}_n(h^*)) \leq 0$ because $\hat{h}_n^*$ is minimizer of $\hat{R}_n$)

- Also, since h^* is minimizer of R,

$$(R(\hat{h}_n^*) - R(h^*)) \geq 0.$$

- Hence

$$0 \leq R(\hat{h}_n^*) - R(h^*) \leq [R(\hat{h}_n^*) - \hat{R}_n(\hat{h}_n^*)] + [\hat{R}_n(h^*) - R(h^*)]$$

- Hence we have

$$|R(\hat{h}_n^*) - R(h^*)| \leq |R(\hat{h}_n^*) - \hat{R}_n(\hat{h}_n^*)| + |\hat{R}_n(h^*) - R(h^*)|$$

- Because of uniform convergence, we can make both terms on the RHS less that $\epsilon/2$, with a high probability, for large n and hence can make the LHS less that ϵ with a large probability.
- This shows consistency of ERM.
- Since arguments like this are needed many times here, let us argue the above more precisely.

- Because of uniform convergence,
  $\forall \epsilon, \delta > 0$, $\exists N(\epsilon, \delta) < \infty$, s.t. $\forall n \geq N(\epsilon, \delta)$,

$$\text{Prob}\left[|R(\hat{h}_n^*) - \hat{R}_n(\hat{h}_n^*)| > \frac{\epsilon}{2}\right] \leq \frac{\delta}{2}, \quad \text{and}$$

$$\text{Prob}\left[|\hat{R}_n(h^*) - R(h^*)| > \frac{\epsilon}{2}\right] \leq \frac{\delta}{2}$$

- Using this, we have to show

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

- Define events A, B, C by

$$A = [|R(\hat{h}_n^*) - R(h^*)| \le \epsilon], \; B = [|R(\hat{h}_n^*) - \hat{R}_n(\hat{h}_n^*)| \le \frac{\epsilon}{2}],$$

$$C = [|\hat{R}_n(h^*) - R(h^*)| \le \frac{\epsilon}{2}]$$

- Since
$|R(\hat{h}_n^*) - R(h^*)| \le |R(\hat{h}_n^*) - \hat{R}_n(\hat{h}_n^*)| + |\hat{R}_n(h^*) - R(h^*)|,$

  we have $\quad A \supset (B \cap C)$ and hence $A^c \subset (B^c \cup C^c)$

- This gives us

$$\text{Prob}[A^c] \leq \text{Prob}[B^c \cup C^c] \leq \text{Prob}[B^c] + \text{Prob}[C^c]$$

- By uniform convergence, probability of both B^c and C^c are less than $\delta/2$. Hene,

$$\text{Prob}[A^c] = \text{Prob}[|R(\hat{h}_n^*) - R(h^*)| > \epsilon] \leq \delta$$

- Thus uniform convergence is sufficient for consistency of empirical risk minimization.

Consistency of Empirical Risk Minimization

- For consistency,the algorithm should satisfy: $\forall \epsilon, \delta > 0$, $\exists N < \infty$, such that

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

- We have shown this.
- In addition, we wanted

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

- We can show this also (using the uniform convergence)

- We have

$$|\hat{R}_n(\hat{h}_n^*) - R(h^*)| \leq |\hat{R}_n(\hat{h}_n^*) - R(\hat{h}_n^*)| + |R(\hat{h}_n^*) - R(h^*)|$$

  by triangular inequality.
- By uniform convergence, for sufficiently large n, we can make both terms on the RHS smaller that $\epsilon/2$ with a large probability. Hence we can make the LHS smaller than ϵ with a large probability (for large n).
- This gives us the result we want.

- Thus convergence of $\hat{R}_n(h)$ to $R(h)$ **uniformly** over $\mathcal{H}$ is sufficient for consistency of empirical risk minimization.
- This uniform convergence is also necessary for the consistency.

- The next question is, given a $\mathcal{H}$, how do we know whether the needed uniform convergence holds.
- We need some useful characterization of family of functions for which this uniform convergence holds.
- That is what we do next.
- We consider only family of binary-valued functions on $\mathcal{X}$. (That is, we are considering 2-class problems with $\mathcal{Y} = \mathcal{A} = \{0,\ 1\}$).
- We also assume that $L(y, h(X)) \in [0,\ 1]$.

- First we note that if $\mathcal{H}$ is finite then the uniform convergence always holds.
- Suppose $\mathcal{H} = \{h_1, h_2, \cdots, h_M\}$.
- By law of large numbers, for any h_i, given any $\epsilon, \delta > 0$, there would be a $N_i(\epsilon, \delta)$ such that

$$\text{Prob}[|\hat{R}_n(h_i) - R(h_i)| > \epsilon] \leq \delta, \ \forall n > N_i(\epsilon, \delta)$$

- Take $N(\epsilon, \delta) = \max_i N_i(\epsilon, \delta)$.
- This N would work for all h_i and hence we have uniform convergence.

- Let us actually calculate the bound on the examples needed, namely, N.
- For this we try to bound Prob$[|\hat{R}_n(h_i) - R(h_i)| > \epsilon]$ with a function of n.
- If we want to use, e.g., Chebyshev inequality for this, we need moments of random variables $L(y, h_i(X))$. But we may not have such information.
- Hence we would use some distribution independent bounds for this.

- Let Z_i be *iid* random variables taking values in $[a,\ b]$, with mean μ. Then the two sided Hoeffding inequality is

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

- This gives a distribution independent bound and hence we can use this.

- Take $Z_i = L(y_i, h(X_i))$. Then Z_i are *iid* random variables taking values in $[0,\ 1]$.
- Then $\frac{1}{n} \sum Z_i = \hat{R}_n(h)$ and $EZ_i = R(h)$.
- Hence, for any h, we have

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

- Recall that $\mathcal{H} = \{h_1, \cdots, h_M\}$.
- Define the events

$$C_\epsilon^i = \left[|\hat{R}_n(h_i) - R(h_i)| > \epsilon \right], \; i = 1, \cdots, M$$

- We have just seen that

$$\text{Prob}(C_\epsilon^i) \le 2\exp(-2n\epsilon^2), \; \forall i$$

Now we have

$$\begin{aligned}\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| > \epsilon\right] &= \text{Prob}\left(C_\epsilon^1 \cup \cdots \cup C_\epsilon^M\right) \\ &\leq \sum_{i=1}^{M} \text{Prob}(C_\epsilon^i) \\ &\leq 2M\exp(-2n\epsilon^2)\end{aligned}$$

- Now we can find how large n should be so that the bound on the RHS is less than δ

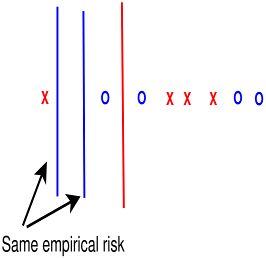
$$2M\exp(-2n\epsilon^2) \leq \delta \quad \text{or} \quad n \geq \frac{1}{2\epsilon^2}\ln\left(\frac{2M}{\delta}\right)$$

- One situation where we can take $\mathcal{H}$ to be finite is when we have Boolean features.
- Suppose we have d Boolean features. Then $\mathcal{X}$ is set of all d-bit Boolean numbers and $2^{\mathcal{X}}$ is a finite set.
- In such cases, we know that ERM is always consistent. However, taking $\mathcal{H} = 2^{\mathcal{X}}$ would still not be nice.
- Here, $\mathcal{X}$ itself is finite with 2^d elements. Hence What is important is the sample complexity.

- For specific finite $\mathcal{H}$ we can get better bounds.
- There are classes of Boolean functions that can be learnt efficiently.
- But, for us, the reason for doing the finite $\mathcal{H}$ case is that it gives us ideas on how to tackle the general case.

- Now let $\mathcal{H}$ be arbitrary.
- Given any h, the value of $\hat{R}_n(h)$ is calculated based on n iid samples.
- Given $h, h' \in \mathcal{H}$, if $h(X_i) = h'(X_i),\ i = 1, \cdots, n$, then, $\hat{R}_n(h) = \hat{R}_n(h')$.

- Consider $\mathcal{X} = \Re$.
- Consider a threshold based classifier.
- That is, let $h_\theta(x) = \text{sign}(x - \theta)$.



- Now let $\mathcal{H}$ be arbitrary.
- Given any h, the value of $\hat{R}_n(h)$ is calculated based on n iid samples.
- Given $h, h' \in \mathcal{H}$, if $h(X_i) = h'(X_i),\ i = 1, \cdots, n$, then, $\hat{R}_n(h) = \hat{R}_n(h')$.
- Since each h is a binary valued function, on the n traning samples, X_i, there are only 2^n tuples of distinct values any function can take.
- Hence based on the values of $\hat{R}_n(h)$ we can only distinguish finitely many functions from $\mathcal{H}$.

- We can sum-up this insight as follows.
- 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 may be able to employ the argument we used for finite $\mathcal{H}$ case to tackle the general case.

- Recall that in the finite case we had

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

- Using our insight we may be able to use this but then M would be a function of n. So, this depends on how the number of distinguishable functions grow with n, the number of examples.
- If it grows as 2^n it would not help.
- Next, we explore this intuitive idea in a more precise fashion.

- Suppose we have $2n$ examples.
- Given any $h \in \mathcal{H}$, we can get an n-sample estimate of $R(h)$ using either the first half or the second half of the examples.
- Let

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

$$\hat{R}'_n(h) = \frac{1}{n} \sum_{i=n+1}^{2n} L(y_i, h(X_i))$$

- Since the examples are *iid*, we can expect that the accuracy of the two estimates, $\hat{R}_n(h)$ and $\hat{R}'_n(h)$ would be about the same for all h.
- Thus, for any h, if $\hat{R}_n(h)$ and $\hat{R}'_n(h)$ differ by a large amount then we can expect that the estimates would differ from the true value, $R(h)$, also by a large amount.

- It is possible to formalize such intuition and show 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)
- This allows us to use the procedure that we adopted for finite $\mathcal{H}$ case to bound the LHS in the inequality above.

- If we can bound

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

  then we can bound the probability we want.
- In the above probability, we need to consider only finitely many h for the supremum.

- First consider any one $h \in \mathcal{H}$.
- Let $Z_i = L(y_i, h(X_i))$.
- By definition of Z_i,

$$|\hat{R}_n(h) - \hat{R}'_n(h)| = \left| \frac{1}{n} \sum_{i=1}^{n} Z_i \ - \ \frac{1}{n} \sum_{i=n+1}^{2n} Z_i \right|$$

- Now, by triangular inequality, we have

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{1}{n}\sum_{i=n+1}^{2n} Z_i\right| \leq$$

$$\left|\frac{1}{n}\sum_{i=1}^{n} Z_i - EZ\right| + \left|EZ - \frac{1}{n}\sum_{i=n+1}^{2n} Z_i\right|$$

- Since examples are *iid*, both terms on the RHS above have the same distribution.

- By same arguments we used earlier, we get

$$\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i \;-\; \frac{1}{n}\sum_{i=n+1}^{2n} Z_i\right| > \frac{\epsilon}{2}\right] \leq$$

$$2\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i \;-\; EZ\right| > \frac{\epsilon}{4}\right]$$

- Now we can use the Hoeffding bound to bound the probability on the RHS in the above inequality.

- The Hoeffding bound gives us

$$\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i \; - \; EZ\right| > \frac{\epsilon}{4}\right] \leq 2\exp\left(-\frac{n\epsilon^2}{8}\right)$$

- Hence we get

$$\text{Prob}\left[\left|\frac{1}{n}\sum_{i=1}^{n} Z_i \; - \; \frac{1}{n}\sum_{i=n+1}^{2n} Z_i\right| > \frac{\epsilon}{2}\right] \leq 4\exp\left(-\frac{n\epsilon^2}{8}\right)$$

- Recall that $Z_i = L(y_i, h(X_i)$ for a specific h.
- Hence $\frac{1}{n}\sum_{i=1}^{n} Z_i = \hat{R}_n(h)$ and $\frac{1}{n}\sum_{i=n+1}^{2n} Z_i = \hat{R}'_n(h)$.

- What we have shown so far is

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

- Since the bound is independent of h, the same bound holds for any h.
- Hence if we want to take supremum over M functions in the LHS above, then we get a multiplicative factor of M on the RHS.

- The probability that we want to bound is

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

- We also know that, given a sample of $2n$ data, we need consider only finitely many h while dealing with the term $|\hat{R}_n(h) - \hat{R}'_n(h)|$.
- Hence the supremum need to be taken over only fintely many h.

- However, the catch is that, the actual number of such h depends on the random sample of examples we have and hence this number is a random variable.
- 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}.

- $M(\mathcal{H}, 2n, S_{2n})$, the number of distinguishable functions, is random because it is a function of S_{2n}.
- For a given S_{2n}, it is just a number.
- Hence we have

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

$$4\, M(\mathcal{H}, 2n, S_{2n})\; \exp\left(-\frac{n\epsilon^2}{8}\right)$$

- Let A be an event, I_A its indicator function and X any random variable.
- Then, by properties of conditional expectation

$$\begin{aligned}\mathsf{Prob}[A] = E[I_A] &= E[\,E[I_A \mid X]\,] \\ &= \int E[I_A \mid X]\, dP(X) \\ &= \int \mathsf{Prob}[A|X] dP(X)\end{aligned}$$

- We can use this idea as follows.

- We get

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

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

- Recall that we have a bound on the probability inside the integral in the RHS above.

- We can use this bound to get

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

$$\int 4\, M(\mathcal{H}, 2n, S_{2n})\, \exp\left(-\frac{n\epsilon^2}{8}\right)\, dP(S_{2n})$$

- The integral on the RHS above is $4\ \exp\left(-\frac{n\epsilon^2}{8}\right)\ EM(\mathcal{H}, 2n, S_{2n})$.

- We do not know $EM(\mathcal{H}, 2n, S_{2n})$. But we can approximate it as

$$EM(\mathcal{H}, 2n, S_{2n}) \leq \max_{S_{2n}} M(\mathcal{H}, 2n, S_{2n})$$

- 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 use all this and get a bound on the probability of interest as

$$
\begin{aligned}
\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - R(h)| > \epsilon\right] \\
&\leq 2\,\text{Prob}\left[\sup_{h\in\mathcal{H}} |\hat{R}_n(h) - \hat{R}'_n(h)| > \frac{\epsilon}{2}\right] \\
&\leq 8\ \exp\left(-\frac{n\epsilon^2}{8}\right)\ \Pi(\mathcal{H}, 2n)
\end{aligned}
$$

- 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 is linear in m, 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

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

- If $d_{VC}(\mathcal{H}) = d$, then only till $n = d$ we have $\Pi(\mathcal{H}, n) = 2^n$; 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,

$$\Pi(\mathcal{H}, m) = \sum_{i=0}^{d} \binom{m}{i}$$

  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

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

- Let $G_{\mathcal{H}}(m) = \ln(\Pi(\mathcal{H}, m))$.
- Then for any $\mathcal{H}$, with $d_{VC}(\mathcal{H}) \leq \infty$, 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.