Introduction to Statistical Learning Theory

Lecture 4

Quick recap:

We have seen that that VC dimension determines PAC learnability for binary classification:

Theorem (Fundamental Theorem of Statistical Learning)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain X to $\{0, 1\}$ and let the loss function be the $0-1$ loss. The following are equivalent:

1. *$\mathcal{H}$ has uniform convergence.*
2. *The ERM is a PAC learning algorithm for $\mathcal{H}$.*
3. *$\mathcal{H}$ is PAC learnable.*
4. *$\mathcal{H}$ has finite VC dimension.*

We have shown that if $VC(\mathcal{H}) = d$ then we can learn with $\mathcal{M}(\epsilon, \delta) = \mathcal{O}\left(\frac{d\ln(1/\epsilon)+\ln(1/\delta)}{\epsilon^2}\right)$ (and claimed the $\ln(1/\epsilon)$ can be removed). We will show that this bound is tight (up to the $\ln(1/\epsilon)$).

Theorem (Complexity lower bound)

Let $\mathcal{H}$ be a hypothesis class of functions from a domain $\mathcal{X}$ to $\{0, 1\}$ with $VC(\mathcal{H}) > 0$ and let the loss function be the $0-1$ loss. Any PAC learning algorithm has sample complexity $\mathcal{M}(\epsilon, \delta) = \Omega\left(\frac{d+\ln(1/\delta)}{\epsilon^2}\right)$.

We will split the dependence in δ and d, starting with δ:

Lemma (1)

Under the previous conditions, $\mathcal{M}(\epsilon, \delta) \geq 0.5 \log(1/(4\delta))/\epsilon^2$ *for* $\epsilon < 1/\sqrt{2}$.

The idea of the proof is to pick 2 almost identical distributions (depending on ϵ) with different optimal solution, so that in order to differentiate with high probability a large number of samples is needed.

Proof: Choose some $c \in \mathcal{X}$ that $\mathcal{H}$ shatters. For each $b \in \{\pm 1\}$ we will define a distribution $\mathcal{D}_b$ that picks c with probability 1, and b with probability $\frac{1+\epsilon}{2}$. This means that $\mathcal{D}_b((c, y)) = \frac{1+by\epsilon}{2}$. It is also not hard to see that $L_{\mathcal{D}_b}(h) = \frac{1-bh(c)\epsilon}{2}$.

Since $L_{\mathcal{D}_b}(h) = \frac{1-bh(c)\epsilon}{2}$ the optimal hypothesis has $L_{\mathcal{D}_b}(h^*) = \frac{1-\epsilon}{2}$, so if $h(c) \neq b$ then $L_{\mathcal{D}_b}(h) = \frac{1+\epsilon}{2} = L_{\mathcal{D}_b}(h^*) + \epsilon$. This means that h is an ϵ approximation iff $h(c) = b$.

We will use the following notations: As x is irrelevant, we will only look at $Y = (y_1, ..., y_m)$. Also we will write $A(Y)$ for $A(Y)(c)$ (as this is what we care about). Lastly we will define $N_+ = \{Y \in \{\pm 1\}^m : \sum y_i \geq 0\}$ and $N_- = \{\pm 1\}^m \backslash N_+$.

Notice that for $Y \in N_+$, we have $P_+(Y) \geq P_-(Y)$ and the opposite for $Y \in N_-$.
We will now show that optimal algorithm (considering the worst case out of $\mathcal{D}_+$ and $\mathcal{D}_-$) is the ERM.

$$\max_{b\in\{\pm\}} P_b\left(A(Y)\neq b\right) \geq \frac{1}{2}P_+(A(Y)=-1)+\frac{1}{2}P_-(A(Y)=1)$$

$$=\frac{1}{2}\sum_{Y\in N_+} P_+(Y)\mathbb{1}(A(Y)=-1)+\sum_{Y\in N_-} P_+(Y)\mathbb{1}(A(Y)=-1)+$$

$$\frac{1}{2}\sum_{Y\in N_+} P_-(Y)\mathbb{1}(A(Y)=1)+\sum_{Y\in N_-} P_-(Y)\mathbb{1}(A(Y)=1)=$$

$$\frac{1}{2}\sum_{Y\in N_+} P_+(Y)\mathbb{1}(A(Y)=-1)+P_-(Y)\mathbb{1}(A(Y)=1)+$$

$$\frac{1}{2}\sum_{Y\in N_-} P_+(Y)\mathbb{1}(A(Y)=-1)+P_-(Y)\mathbb{1}(A(Y)=1)\geq$$

$$\frac{1}{2}\sum_{Y\in N_+} P_-(Y)\mathbb{1}(A(Y)=-1)+P_-(Y)\mathbb{1}(A(Y)=1)+$$

$$\frac{1}{2}\sum_{Y\in N_+} P_+(Y)\mathbb{1}(A(Y)=-1)+P_+(Y)\mathbb{1}(A(Y)=1)=\frac{1}{2}\left(L_{\mathcal{D}_+}(ERM)+L_{\mathcal{D}_-}(ERM)\right)$$

For the ERM, $L_{\mathcal{D}_+}(ERM) = L_{\mathcal{D}_-}(ERM)$ (up to ties which we can exclude by having uneven m). Both are equal that a binomial $B(m, (1-\epsilon)/2)$ has a value greater then $m/2$. This can be bounded using Slud's inequality:

Theorem (Slud's inequality)

Let $X \sim B(m, (1-\epsilon)/2)$ then

$$P(X \geq m/2) \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp(-m\epsilon^2/(1-\epsilon^2))}\right)$$

So the error probability is greater or equal to $\frac{1}{2}\left(1 - \sqrt{1 - \exp(-m\epsilon^2/(1-\epsilon^2))}\right) \geq \frac{1}{2}\left(1 - \sqrt{1 - \exp(-2m\epsilon^2)}\right)$ using the $\epsilon^2 < 1/2$ assumption. We can conclude that for $m < 0.5\ln(1/(4\delta))/\epsilon^2$

$$\max_b P\left(L_{\mathcal{D}_b}(A(Y)) - \min_h L_{\mathcal{D}_b}(h) \geq \epsilon\right) \geq \frac{1}{2}(1 - \sqrt{1-4\delta}) \geq \delta$$

Where the last inequality is simple algebra (noticing the theorem is trivial for $\delta > 1/4$) . This finishes the proof.

We now need to bound the dependence in $d = VC(\mathcal{H})$

Lemma (2)

Under the previous conditions, $\mathcal{M}(\epsilon, \delta) \geq \frac{d}{8^3\epsilon^2}$ *for* $\epsilon < 1/8\sqrt{2}$.

The proof is similar to the previous proof. Define $\rho = 8\epsilon$. Pick $c_1, ..., c_d$ that $\mathcal{H}$ shatters. for any $b \in \{\pm 1\}^d$ define a distribution $\mathcal{D}_b$ that first picks $x = c_i$ uniformly out of $c_1, ..., c_d$ then picks y with probability $(1 + yb_i\rho)/2$.

The next step is to prove that the ERM is optimal algorithm when considering *worst case*. The proof is very similar to what we did earlier (using independence and the same tricks) but a bit more cumbersome so we will skip it.

For any function f

$$L_{\mathcal{D}_b}(f) = \frac{1+\rho}{2} \cdot \frac{|\{i \in [d] : f(c_i) \neq b_i\}|}{d} + \frac{1-\rho}{2} \cdot \frac{|\{i \in [d] : f(c_i) = b_i\}|}{d}$$

So $L_{\mathcal{D}_b}(f) - \min_h L_{\mathcal{D}_b}(h) = \rho \cdot \frac{|\{i \in [d] : f(c_i) \neq b_i\}|}{d}$.

We will bound $\mathbb{E}_{S \sim \mathcal{D}_b^M}[L_{\mathcal{D}_b}(ERM(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_b}(h)]$ next:

$$\mathbb{E}_S[L_{\mathcal{D}_b}(ERM(S)) - \min_{h \in \mathcal{H}} L_{\mathcal{D}_b}(h)] = \frac{\rho}{d}\mathbb{E}_S[|\{i \in [d] : ERM(c_i) \neq b_i\}|]$$

We can look at the sampling as first sampling the c_i index $K \sim U([d])^m$ and then sampling the labels $y_i \sim b_{K_i}$ (with some abuse of notation).

We define for each $K \in [d]^m$, $n_i(K)$ the number of times the index i appears in K. Then

$$\frac{\rho}{d}\mathbb{E}_S[|\{i \in [d] : ERM(c_i) \neq b_i\}|] = \frac{\rho}{d}\sum_{i=1}^{d}\mathbb{E}_K\mathbb{E}_{y_j \sim b_{K_j}}[\mathbb{1}(ERM(S)(c_i) \neq b_i]$$

$$\overset{1}{\geq} \frac{\rho}{2d}\sum_{i=1}^{d}\mathbb{E}_K(1-\sqrt{1-\exp(-2\rho^2 n_i(K))}) \overset{2}{\geq} \frac{\rho}{2d}\sum_{i=1}^{d}\mathbb{E}_K(1-\sqrt{2\rho^2 n_i(K)})$$

$$\overset{3}{\geq} \frac{\rho}{2d}\sum_{i=1}^{d}\left(1-\sqrt{2\rho^2\mathbb{E}_K[n_i(K)]}\right) = \frac{\rho}{2d}\sum_{i=1}^{d}\left(1-\sqrt{2\rho^2 m/d}\right)$$

$$= \frac{\rho}{2}\left(1-\sqrt{2\rho^2 m/d}\right)$$

Where (1) is Slut's inequality as before (using $\rho^2 < 1/2$), (2) if from the inequality $1 - e^{-a} \leq a$ and (3) is Jensen's inequality.

In summery we have shown so far that for every algorithm A, there exists a distribution such that
$\mathbb{E}_S[L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)] \geq \frac{\rho}{2}\left(1 - \sqrt{2\rho^2 m/d}\right) \geq \frac{\rho}{4}$
for $m < \frac{d}{8^3\epsilon^2} = \frac{d}{8\rho^2}$.

To finish we will use a version of the Markov inequality
$P(X > a) \geq \mathbb{E}[X] - a$, for $X \in [0,1]$, $a \in (0,1)$. Define
$\Delta = \frac{1}{\rho}\left(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)\right)$ and notice that $\Delta \in [0,1]$.

$$P(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h) > \epsilon) = P(\Delta > \epsilon/\rho)$$

finishing the proof of the lemma. With both lemmas, the theorem is straightforward.

In summery we have shown so far that for every algorithm A, there exists a distribution such that
$\mathbb{E}_S[L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)] \geq \frac{\rho}{2}\left(1 - \sqrt{2\rho^2 m/d}\right) \geq \frac{\rho}{4}$
for $m < \frac{d}{8^3\epsilon^2} = \frac{d}{8\rho^2}$.

To finish we will use a version of the Markov inequality
$P(X > a) \geq \mathbb{E}[X] - a$, for $X \in [0,1]$, $a \in (0,1)$. Define
$\Delta = \frac{1}{\rho}\left(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)\right)$ and notice that $\Delta \in [0,1]$.

$$P(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h) > \epsilon) = P(\Delta > \epsilon/\rho) \geq \mathbb{E}[\Delta] - \frac{\epsilon}{\rho}$$

finishing the proof of the lemma. With both lemmas, the theorem is straightforward.

In summery we have shown so far that for every algorithm A, there exists a distribution such that
$\mathbb{E}_S[L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)] \geq \frac{\rho}{2}\left(1 - \sqrt{2\rho^2 m/d}\right) \geq \frac{\rho}{4}$
for $m < \frac{d}{8^3\epsilon^2} = \frac{d}{8\rho^2}$.

To finish we will use a version of the Markov inequality
$P(X > a) \geq \mathbb{E}[X] - a$, for $X \in [0,1]$, $a \in (0,1)$. Define
$\Delta = \frac{1}{\rho}\left(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h)\right)$ and notice that $\Delta \in [0,1]$.

$$P(L_{\mathcal{D}_b}(A(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}_b}(h) > \epsilon) = P(\Delta > \epsilon/\rho) \geq \mathbb{E}[\Delta] - \frac{\epsilon}{\rho} \geq \frac{1}{4} - \frac{\epsilon}{\rho} = \frac{1}{8}$$

finishing the proof of the lemma. With both lemmas, the theorem is straightforward.

We have seen that learning is possible with $\mathcal{M}(\epsilon,\delta) = \mathcal{O}\left(\frac{d\ln(1/\epsilon)+\ln(1/\delta)}{\epsilon^2}\right)$ using the ERM algorithm, and that $\mathcal{M}(\epsilon,\delta) = \Omega\left(\frac{d+\ln(1/\delta)}{\epsilon^2}\right)$ for any learning algorithm.

We have seen (and it can be extended) that the ERM is optimal when it comes to minimizing the worst case scenario.

It is important to note, that under further assumptions (such as smoothness, etc.) other algorithms may perform much better.

So far we have studied learnability via uniform convergence in binary classification. We will now show a more general was to bound uniform convergence - Rademacher complexity.

First a small notation change - define $z = (x, y)$ and $l(h, z) = l(h(x), y)$. This allows us to work in a more general setting with the same notation.

Another notation for simplicity: Define $\mathcal{F} = l \circ \mathcal{H}$, so for $f \in \mathcal{F}$- $L_{\mathcal{D}}(f) = \mathbb{E}_{\mathcal{D}}[f(z)]$ and $L_S(f) = \frac{1}{m} \sum_{i=1}^{m} f(z_i)$.

We are interested in bounding $\sup_{h \in \mathcal{H}} (L_{\mathcal{D}}(h) - L_S(h))$. As we have seen before, a good proxy for $L_{\mathcal{D}}(h)$ is $L_{\tilde{S}}(h)$ the loss on some second test sample. As we only have S we can split it into two equal size disjoint sets, S_1 and S_2.

Definition

$\sup_{h\in\mathcal{H}}(L_{S_2}(h)-L_{S_1}(h)) = \frac{2}{m}\sup_{f\in\mathcal{F}}\left(\sum_{z_i\in S_2} f(z_i) - \sum_{z_j\in S_1} f(z_j)\right) =$

$\frac{2}{m}\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i f(z_i)\right)$. where $\sigma_i \in \{\pm 1\}$ indicates if z_i is in S_1 or S_2. If we randomize σ_i we get the Rademacher complexity.

Definition (General Rademacher Complexity)

For $A \subset \mathbb{R}^m$ define $R(A) = \frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}\left[\sup_{a\in A}\sum_{i=1}^{m}\sigma_i a_i\right]$

Definition (Empirical Rademacher Complexity)

Define $\mathcal{F}\circ S = \{(f(z_1),...,f(z_m)) : f\in\mathcal{F}\} \subset \mathbb{R}^m$ the empirical Rademacher complexity is defined as

$R(\mathcal{F}\circ S) = \frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i f(z_i)\right]$

Definition (Rademacher Complexity)

The Rademacher complexity of $\mathcal{F}$ is the expected empirical Rademacher complexity, $\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) = \mathbb{E}_{S\sim\mathcal{D}^m}\left[R(\mathcal{F}\circ S)\right]$

The following lemma gives a nice intuition of the Rademacher complexity when considering binary classification

Lemma

Let $\mathcal{H} : \mathcal{X} \to \{\pm 1\}$, $S_X = \{x_1, ..., x_m\} \subset \mathcal{X}$. Define $err(\mathcal{H})$ as the expected sample error of the ERM algorithm on random labels, then $err(\mathcal{H}) = 1/2\left(1 - R(\mathcal{H}\circ S_X)\right)$.

Proof.

Let σ be any labeling on S_x. Then

$$L_{S_X,\sigma}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq \sigma_i\}$$

This means that
$L_{S_X,\sigma}(ERM) = \min_{h\in\mathcal{H}} \frac{1}{2}(1 - \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)) = \frac{1}{2} - \frac{1}{2m}\max_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)$. Take expectation with regard to $\sigma_i \sim \{\pm 1\}^m$ and you get the Rademacher complexity. □

Proof.

Let σ be any labeling on S_x. Then

$$L_{S_X,\sigma}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq \sigma_i\} = \frac{1}{m}\sum_{i=1}^{m} \frac{1-\sigma_i h(x_i)}{2}$$

This means that
$L_{S_X,\sigma}(ERM) = \min_{h\in\mathcal{H}} \frac{1}{2}(1 - \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)) = \frac{1}{2} - \frac{1}{2m}\max_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)$. Take expectation with regard to $\sigma_i \sim \{\pm 1\}^m$ and you get the Rademacher complexity. □

Proof.

Let σ be any labeling on S_x. Then

$$L_{S_X,\sigma}(h) = \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}\{h(x_i) \neq \sigma_i\} = \frac{1}{m}\sum_{i=1}^{m} \frac{1-\sigma_i h(x_i)}{2}$$

$$= \frac{1}{2} - \frac{1}{2m}\sum_{i=1}^{m} \sigma_i h(x_i)$$

This means that
$L_{S_X,\sigma}(ERM) = \min_{h\in\mathcal{H}} \frac{1}{2}(1 - \frac{1}{m}\sum_{i=1}^{m} \sigma_i h(x_i)) = \frac{1}{2} - \frac{1}{2m}\max_{h\in\mathcal{H}} \sum_{i=1}^{m} \sigma_i h(x_i)$. Take expectation with regard to $\sigma_i \sim \{\pm 1\}^m$ and you get the Rademacher complexity. □

Lemma

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\sup_{f\in\mathcal{F}}(L_{\mathcal{D}}(f) - L_S(f))] \leq 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m)$$

Proof: Let $\tilde{S} \sim \mathcal{D}^m$ be another sample, then

$$L_{\mathcal{D}}(f) - L_S(f) = \mathbb{E}_{\tilde{S}}[L_{\tilde{S}}(f)] - L_S(f) = \mathbb{E}_{\tilde{S}}[L_{\tilde{S}}(f) - L_S(f)] \tag{1}$$

therefore:

$$\sup_{f\in\mathcal{F}}(L_{\mathcal{D}}(f) - L_S(f)) = \sup_{f\in\mathcal{F}} \mathbb{E}_{\tilde{S}}[L_{\tilde{S}}(f) - L_S(f)] \tag{2}$$

$$\leq \mathbb{E}_{\tilde{S}}[\sup_{f\in\mathcal{F}}(L_{\tilde{S}}(f) - L_S(f))] \tag{3}$$

Taking expectation with regard to S we get

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\sup_{f\in\mathcal{F}}(L_{\mathcal{D}}(f)-L_S(f))] \leq \mathbb{E}_{S,\tilde{S}}[\sup_{f\in\mathcal{F}}(L_{\tilde{S}}(f)-L_S(f))] \tag{4}$$

$$= \frac{1}{m}\mathbb{E}_{S,\tilde{S}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}(f(\tilde{z}_i)-f(z_i))\right] \tag{5}$$

$$= \frac{1}{m}\mathbb{E}_{S,\tilde{S}}\left[\sup_{f\in\mathcal{F}}\sum_{i\neq j}(f(\tilde{z}_i)-f(z_i))+f(\tilde{z}_j)-f(z_j)\right] \tag{6}$$

$$= \frac{1}{m}\mathbb{E}_{S,\tilde{S}}\left[\sup_{f\in\mathcal{F}}\sum_{i\neq j}(f(\tilde{z}_i)-f(z_i))+f(z_j)-f(\tilde{z}_j)\right] \tag{7}$$

$$= \frac{1}{m}\mathbb{E}_{S,\tilde{S},\sigma_j}\left[\sup_{f\in\mathcal{F}}\sum_{i\neq j}(f(\tilde{z}_i)-f(z_i))+\sigma_j(f(z_j)-f(\tilde{z}_j))\right] \tag{8}$$

This is true since z_i and $\tilde{z}_i$ are drawn from the same distribution. We can do this for all $1 \leq j \leq m$ and get

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\sup_{f\in\mathcal{F}}(L_{\mathcal{D}}(f) - L_S(f))] \leq \frac{1}{m}\mathbb{E}_{S,\tilde{S},\sigma}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{m}\sigma_i(f(\tilde{z}_i) - f(z_i))\right] \tag{9}$$

$$\leq \frac{1}{m}\mathbb{E}_{S,\tilde{S},\sigma}\left[\sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}\sigma_i f(\tilde{z}_i)\right) + \sup_{f\in\mathcal{F}}\left(\sum_{i=1}^{m}-\sigma_i f(z_i)\right)\right] \tag{10}$$

$$= 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) \quad \square \tag{11}$$

In order to turn our expectation bound to a high-probability bound, we need a concentration of measure theorem. We will use Mcdiarmid's inequality.

Theorem (McDiarmid's Inequality)

Let V be some set and $f : V^m \to \mathbb{R}$ be a function such that for some $c > 0$ and all $x_1, ..., x_m, x_i' \in V$ we have

$$|f(x_1, ..., x_m) - f(x_1, ..., x_{i-1}, x_i', x_{i+1}, ..., x_m)| \leq c \tag{12}$$

If $X_1, ..., X_m$ are independent r.v. taking values in V, then with probability greater or equal to $1 - \delta$ we have

$$|f(X_1, ..., X_m) - \mathbb{E}[f(X_1, ..., X_m)]| \leq c\sqrt{\ln\left(\frac{2}{\delta}\right)\frac{m}{2}} \tag{13}$$

We can now state and prove the main theorem -

Theorem

If for all z and $h \in \mathcal{H}$ we have $|l(h,z)| \leq c$. Then with probability at least $1-\delta$, for all $h \in \mathcal{H}$:

1. $|L_{\mathcal{D}}(h) - L_S(h)| \leq 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) + c\sqrt{\frac{2\ln(2/\delta)}{m}}$
2. $|L_{\mathcal{D}}(h) - L_S(h)| \leq 2R(\mathcal{F} \circ S) + 3c\sqrt{\frac{2\ln(4/\delta)}{m}}$
3. $L_{\mathcal{D}}(ERM) - L_{\mathcal{D}}(h^*) \leq 2R(\mathcal{F} \circ S) + 5c\sqrt{\frac{2\ln(8/\delta)}{m}}$

Notice that the last two inequalities only use the empirical sample, and can (up to computational complexity issues) be calculated for a given instance.

Proof:
We have $\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h))$ satisfies the bounded diffrences condition eq. 12 with constant $2c/m$. Using the expectation bound of lemma 10 and the McDiarmid's inequality we have with probability $\geq 1-\delta$

$$|\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h))| \leq \mathbb{E}_S[\sup_{h\in\mathcal{H}}(L_{\mathcal{D}}(h) - L_S(h))] + c\sqrt{\frac{2\ln(2/\delta)}{m}} \quad (14)$$

$$\leq 2\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) + c\sqrt{\frac{2\ln(2/\delta)}{m}}. \quad (15)$$

To prove the second inequality we note that $\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m)$ satisfies the bounded difference condition with the same constant, so with probability $\geq 1-\delta/2$, we have $\mathcal{R}_{\mathcal{D}}(\mathcal{F}, m) \leq R(\mathcal{F}\circ S) + c\sqrt{\frac{2\ln(4/\delta)}{m}}$. This and the union bound finish the proof of part 2.

The last part uses the 2nd inequality, the Hoeffding inequality to bound $L_S(h^*) - L_{\mathcal{D}}(h^*)$ and the union bound. It is left as an exercise.

We will prove some useful theorem for bounding the Rademacher complexity.

Lemma

For any $A \in \mathbb{R}^m$, scalar $c > 0$ and $v \in \mathbb{R}^m$ we have $R(cA + v) = cR(A)$.

Proof.

$$R(cA+v) = \frac{1}{m}\mathbb{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^{m}\sigma_i(ca_i+v_i)\right] =$$

$$\frac{1}{m}\mathbb{E}_\sigma\left[c\sup_{a\in A}\left(\sum_{i=1}^{m}\sigma_i a_i\right)+\sum_{i=1}^{m}\sigma_i v_i\right] = cR(A)+\frac{1}{m}\sum_{i=1}^{m}v_i\mathbb{E}_\sigma\left[\sigma_i\right] = cR(A)$$

Lemma

For any $A \in \mathbb{R}^m$, $R(conv(A)) = R(A)$, when $conv(A)$ is the convex hall of A.

Proof.

Define $\Delta^n = \{\lambda \in \mathbb{R}^n : \forall i : \lambda_i \geq 0, \sum_{i=1}^n \lambda_i = 1\}$. The convex hall is defined as $conv(A) = \{\sum_{i=1}^n \lambda_i a^{(i)} : \forall i a^{(i)} \in A, \lambda \in \Delta^n\}$. The key observation is that for every vector a we have $\sup_{\lambda \in \Delta^n} \sum \lambda_i x_i = \max_i x_i$.

$$mR(conv(A)) = \mathbb{E}_\sigma \left[\sup_{\lambda \in \Delta^n} \sup_{a^{(1)},\ldots,a^{(n)} \in A} \sum_{i=1}^m \sigma_i \sum_{j=1}^n \lambda_j a_i^{(j)} \right]$$

$$= \mathbb{E}_\sigma \left[\sup_{\lambda \in \Delta^n} \sum_{j=1}^n \lambda_j \sup_{a^{(j)}} \sum_{i=1}^m \sigma_i a_i^{(j)} \right] = \mathbb{E}_\sigma \left[\sup_{a \in A} \sum_{i=1}^m \sigma_i a_i \right] = R(A)$$

Lemma (Massart Lemma)

If $A = \{a_1, ..., a_M\} \in \mathbb{R}^m$ *is a finite set and* $\bar{a} = \frac{1}{N}\sum_{i=1}^{N} a_i$, *then*

$$R(A) \leq \max_{a \in A} ||a - \bar{a}|| \frac{\sqrt{2\ln(N)}}{m} \tag{16}$$

Immediate corollary - if l is the zero one loss and $\mathcal{H}$ has VC dimension d then $R(l \circ \mathcal{H} \circ S) \leq \sqrt{\frac{2d\ln(em/d)}{m}}$.

Proof: Using lemma 13, we can assume $\bar{a} = 0$ and recall that $R(\lambda A) = \lambda R(A)$ for $\lambda > 0$.

$$mR(\lambda A) = \mathbb{E}_\sigma \left[\max_{b \in \lambda A} \langle \sigma, b \rangle \right] = \mathbb{E}_\sigma \left[\ln \left(\max_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right) \right]$$
$$\leq \mathbb{E}_\sigma \left[\ln \left(\sum_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right) \right] \overset{1}{\leq} \ln \left(\mathbb{E}_\sigma \left[\sum_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right] \right)$$
$$\overset{2}{=} \ln \left(\sum_{b \in \lambda A} \prod \mathbb{E}_{\sigma_i} \left[e^{\sigma_i \cdot b_i} \right] \right) \overset{3}{\leq} \ln \left(\sum_{b \in \lambda A} \prod e^{b_i^2/2} \right)$$

Where (1) is the Jensen inequality, (2) is from independence, and (3) is from a technical inequality $\frac{e^a + e^{-a}}{2} \leq e^{a^2/2}$ we will prove shortly. We now have

$$\lambda mR(A) = mR(\lambda A) \leq \ln \left(\sum_{b \in \lambda A} e^{||b||^2/2} \right)$$

$$mR(\lambda A) = \mathbb{E}_\sigma \left[\max_{b \in \lambda A} \langle \sigma, b \rangle \right] = \mathbb{E}_\sigma \left[\ln \left(\max_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right) \right]$$

$$\leq \mathbb{E}_\sigma \left[\ln \left(\sum_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right) \right] \overset{1}{\leq} \ln \left(\mathbb{E}_\sigma \left[\sum_{b \in \lambda A} e^{\langle \sigma, b \rangle} \right] \right)$$

$$\overset{2}{=} \ln \left(\sum_{b \in \lambda A} \prod \mathbb{E}_{\sigma_i} \left[e^{\sigma_i \cdot b_i} \right] \right) \overset{3}{\leq} \ln \left(\sum_{b \in \lambda A} \prod e^{b_i^2/2} \right)$$

Where (1) is the Jensen inequality, (2) is from independence, and (3) is from a technical inequality $\frac{e^a + e^{-a}}{2} \leq e^{a^2/2}$ we will prove shortly. We now have

$$\lambda mR(A) = mR(\lambda A) \leq \ln \left(\sum_{b \in \lambda A} e^{||b||^2/2} \right) \leq \ln(|A|) + \max_{a \in A} \lambda^2 ||a||^2/2$$

We have $R(A) \leq \frac{\ln(|A|)+\lambda^2 \max_{a\in A}(||a||^2/2)}{\lambda m}$. Setting $\lambda = \sqrt{2\ln(|A|)/\max_{a\in A}||a||^2}$ concludes the proof.

We still need to prove $\frac{e^a+e^{-a}}{2} \leq e^{a^2/2}$. Using the Tylor series $e^a = \sum_{i=0}^{\infty} \frac{a^n}{n!}$ so $\frac{e^a+e^{-a}}{2} = \sum_{i=0}^{\infty} \frac{a^{2n}}{(2n)!}$. On the other hand $e^{a^2/2} = \sum_{i=0}^{\infty} \frac{a^{2n}}{2^n n!}$. Observing that $(2n)! \geq 2^n n!$ finishes the proof. □

Lemma (Contraction Lemma)

For all $i \in [m]$, let $\phi_i : \mathbb{R} \to \mathbb{R}$ be a $\rho-$Lipschitz function. For all $a \in \mathbb{R}^m$, define $\phi(a) = (\phi_1(a_1), ..., \phi_m(a_m))$. Then $R(\phi \circ A) \leq \rho R(A)$.

Using lemma 13 we can assume $\rho = 1$. Moreover if we define $A_i = \{(a_1, ..., a_{i-1}, \phi(a_i), a_{i+1}, ..., a_m) : a \in A\}$, it is enough to show $R(A_1) \leq R(A)$.

$$mR(A_1) = \mathbb{E}_\sigma \left[\sup_{a \in A_1} \sum_{i=1}^{m} \sigma_i a_i \right] = \mathbb{E}_\sigma \left[\sup_{a \in A} \sigma_1 \phi(a_1) + \sum_{i=2}^{m} \sigma_i a_i \right] \tag{17}$$

$$= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_2, ..., \sigma_m} \left[\sup_{a \in A} \left(\phi(a_1) + \sum_{i=2}^{m} \sigma_i a_i \right) + \sup_{a \in A} \left(-\phi(a_1) + \sum_{i=2}^{m} \sigma_i a_i \right) \right] \tag{18}$$

$$= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_2, ..., \sigma_m} \left[\sup_{a, a' \in A} \left(\phi(a_1) - \phi(a'_1) + \sum_{i=2}^{m} \sigma_i (a_i + a'_i) \right) \right] \tag{19}$$

$$\leq \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_2,\ldots,\sigma_m} \left[\sup_{a,a' \in A} \left(|a_1 - a_1'| + \sum_{i=2}^{m} \sigma_i (a_i + a_i') \right) \right] \tag{20}$$

$$= \frac{1}{2} \mathop{\mathbb{E}}_{\sigma_2,\ldots,\sigma_m} \left[\sup_{a,a' \in A} \left(a_1 - a_1' + \sum_{i=2}^{m} \sigma_i (a_i + a_i') \right) \right] \tag{21}$$

Where the last equality is from the fact we can switch a and a'. If we look at steps $17 - 19$ using $\phi = Id$ we see that the last line is equal to $R(A)$ finishing the proof $\square$.