10-806 Foundations of Machine Learning and Data Science

Maria-Florina Balcan Lecture 8: October 5, 2015

1 Rademacher Complexity

1.1 Motivation

The ultimate goal of passive supervised machine learning is to find a hypothesis function based on a set of examples that has small error with respect to some target function. This goal is independent of most aspects of the learning setting—that is, the same goal applies to both classification and regression problems in both the realizable and agnostic cases—so it would be nice to have a general way of dealing with this type of problem.

Often, our attempts to get a handle on the sufficient conditions for learning (most notably in the PAC model) led us to proving results known as uniform convergence bounds. These bounds stated that the empirical errors of concepts from a given concept class $\mathcal{H}$ converge uniformly to their true errors. In other words, we bounded the difference between the training error and generalization error for all functions in $\mathcal{H}$.

Through our discussions of VC theory we have seen that we can improve generalization by controlling the complexity of the concept class $\mathcal{H}$ from which we are choosing a hypothesis. We saw that the shatter coefficient and VC dimension were useful measures of complexity because we could bound the performance of a learning algorithm in terms of these quantities and the amount of data we have.

The bounds we derived based on VC dimension were distribution independent. In some sense, distribution independence is a nice property because it guarantees the bounds to hold for any data distribution. On the other hand, the bounds may not be tight for some specific distributions that are more benign than the worst case. Furthermore, the concepts used in defining VC dimension apply only to binary classification, but we are often interested in generalization bounds for multiclass classification and regression as well.

Rademacher complexity is a more modern notion of complexity that is distribution dependent and defined for any class real-valued functions (not only discrete-valued functions).

1.2 Definitions

Given a space Z and a fixed distribution $D|_Z$, let $S = \{z_1, \ldots, z_m\}$ be a set of examples drawn i.i.d. from $D|_Z$. Furthermore, let $\mathcal{F}$ be a class of functions $f : Z \to \mathbb{R}$.

Definition. The *empirical Rademacher complexity* of $\mathcal{F}$ is defined as

$$\hat{R}_m(\mathcal{F}) \;=\; \mathsf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right) \right]$$

where $\sigma_1, \ldots, \sigma_m$ are independent random variables uniformly chosen from $\{-1, 1\}$. We will refer to such random variables as *Rademacher variables*.

Definition. The *Rademacher complexity* of $\mathcal{F}$ is defined as

$$R_m(\mathcal{F}) \;=\; \mathsf{E}_D[\hat{R}_m(\mathcal{F})]$$

Intuitively, the supremum measures, for a given set S and Rademacher vector σ, the maximum correlation between $f(z_i)$ and σ_i over all $f \in \mathcal{F}$. Taking the expectation over σ, we can then say that the empirical Rademacher complexity of $\mathcal{F}$ measures the ability of functions from $\mathcal{F}$ (when applied to a fixed set S) to fit random noise. The Rademacher complexity of $\mathcal{F}$ then measures the expected noise-fitting-ability of $\mathcal{F}$ over all data sets $S \in Z^m$ that could be drawn according to the distribution $D|_Z$.

We note that Rademacher complexity can be defined even more generally on sets $A \subseteq \mathbb{R}^m$ by making the supremum over $a \in A$ (instead of $f \in \mathcal{F}$) and replacing each $f(z_i)$ with a_i. Taking $A = \mathcal{F}(S) = \{f(z) | f \in \mathcal{F}, z \in S\}$ recovers the definition above. It will sometimes be convenient to use the more general definition.

1.3 A General Sample Complexity Result

1.3.1 A useful tail inequality

In deriving generalization bounds using Rademacher complexity, we will make use of the following concentration bound. The bound, also known as the bounded differences inequality, can be very useful in other applications as well.

Theorem 1 (McDiarmid Inequality)**.** *Let $x_1, \ldots, x_n$ be independent random variables taking on values in a set A and let $c_1, \ldots, c_n$ be positive real constants. If $\varphi : A^n \to \mathbb{R}$ satisfies*

$$\sup_{x_1,\ldots,x_n,x_i' \in A} |\varphi(x_1, \ldots, x_i, \ldots, x_n) - \varphi(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i,$$

for $1 \leq i \leq n$, then

$$\Pr[\varphi(x_1, \ldots, x_n) - \mathsf{E}[\varphi(x_1, \ldots, x_n)] \geq \epsilon] \ \leq \ e^{-2\epsilon^2 / \sum_{i=1}^n c_i^2}.$$

1.3.2 Rademacher-based uniform convergence

We now show a uniform convergence result for any class of (bounded) real-valued functions. We bound the expectation of each function in terms of the empirical average of the function, the Rademacher complexity of the class, and an error term depending on the confidence parameter and sample size. We denote the empirical average over a sample S as $\hat{\mathsf{E}}_S[f(z)] = \frac{1}{|S|} \sum_{z \in S} f(z)$.

Theorem 2. *Fix distribution $D|_Z$ and parameter $\delta \in (0, 1)$. If $\mathcal{F} \subseteq \{f : Z \to [a, a+1]\}$ and $S = \{z_1, \ldots, z_n\}$ is drawn i.i.d. from $D|_Z$ then with probability $\geq 1 - \delta$ over the draw of S, for every function $f \in \mathcal{F}$,*

$$\mathsf{E}_D[f(z)] \ \leq \ \hat{\mathsf{E}}_S[f(z)] + 2R_m(\mathcal{F}) + \sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{1}$$

In addition, with probability $\geq 1 - \delta$, for every function $f \in \mathcal{F}$,

$$\mathsf{E}_D[f(z)] \ \leq \ \hat{\mathsf{E}}_S[f(z)] + 2\hat{R}_m(\mathcal{F}) + 3\sqrt{\frac{\ln(2/\delta)}{2m}}. \tag{2}$$

Proof. For a fixed function f, the definition of supremum leads us directly to

$$\mathsf{E}_D[f(z)] \;\leq\; \hat{\mathsf{E}}_S[f(z)] + \sup_{h\in\mathcal{F}} \Big(\mathsf{E}_D[h(z)] - \hat{\mathsf{E}}_S[h(z)]\Big)$$

which is already in a form similar to the statement we wish to prove. The supremum term is a random variable that depends on the draw of S. Denoting

$$\varphi(S) = \sup_{h\in\mathcal{F}} \Big(\mathsf{E}_D[h(z)] - \hat{\mathsf{E}}_S[h(z)]\Big), \tag{3}$$

we would like to bound φ with high probability in terms of its expectation, and we will do this by using the McDiarmid Inequality. Later, we will bound its expectation in terms of the Rademacher complexity of $\mathcal{F}$, and this will complete the proof.

In order to satisfy the conditions of Theorem 1, we need the following lemma.

Lemma 1. *The function φ defined in (3) satisfies*

$$\sup_{z_1,\ldots,z_n,z_i'\in Z} |\varphi(z_1,\ldots,z_i,\ldots,z_n) - \varphi(z_1,\ldots,z_i',\ldots,z_n)| \leq \frac{1}{m}$$

A formal proof of this lemma is left to the appendix. An intuitive justification follows from the fact that the image of every function $h \in \mathcal{F}$ is a subset of $[a, a+1]$, so the maximum change in the value of $h(z)$ is 1. This change is occurring within an empirical average, so its effect on the value of $\varphi(S)$ is scaled down by a factor of $1/m$.

Now that we have determined that φ satisfies the proper conditions, we can apply the McDiarmid Inequality to find

$$\Pr[\varphi(S) - \mathsf{E}[\varphi(S)] \;\geq\; t] \;\leq\; e^{-2t^2/\sum_{i=1}^m \frac{1}{m^2}} \;=\; e^{-2t^2 m}.$$

Setting the above probability to be less than δ and solving for t, we find that this probability is less than δ if and only if $t \geq \sqrt{\frac{\ln(1/\delta)}{2m}}$. Since

$$\Pr[\varphi(S) \leq \mathsf{E}[\varphi(S)] + t] \;=\; 1 - \Pr[\varphi(S) - \mathsf{E}[\varphi(S)] \geq t],$$

we have determined that with probability at least $1-\delta$,

$$\mathsf{E}_D[f(z)] \;\leq\; \hat{\mathsf{E}}_S[f(z)] + \mathsf{E}_S\left[\sup_{h\in\mathcal{F}} \Big(\mathsf{E}_D[h(z)] - \hat{\mathsf{E}}_S[h(z)]\Big)\right] + \sqrt{\frac{\ln(1/\delta)}{2m}}. \tag{4}$$

The final step is to bound the expectation of $\varphi(S)$ in terms of the Rademacher complexity of $\mathcal{F}$. In order to do this, we introduce a "ghost sample" $\tilde{S} = \{\tilde{z}_1,\ldots,\tilde{z}_m\}$ independently drawn identically to S. Since $\mathsf{E}_{\tilde{S}}[\hat{\mathsf{E}}_{\tilde{S}}[h(z)] \,|\, S] = \mathsf{E}_D[h(z)]$ and $\mathsf{E}_{\tilde{S}}[\hat{\mathsf{E}}_S[h(z)] \,|\, S] = \hat{\mathsf{E}}_S[h(z)]$ we can rewrite the expectation

$$\begin{aligned}
\mathsf{E}_S\left[\sup_{h\in\mathcal{F}} \Big(\mathsf{E}_D[h(z)] - \hat{\mathsf{E}}_S[h(z)]\Big)\right] \;&=\; \mathsf{E}_S\left[\sup_{h\in\mathcal{F}} \mathsf{E}_{\tilde{S}}\left[\hat{\mathsf{E}}_{\tilde{S}}[h(z)] - \hat{\mathsf{E}}_S[h(z)] \;\middle|\; S\right]\right] \\
&=\; \mathsf{E}_S\left[\sup_{h\in\mathcal{F}} \mathsf{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^m h(\tilde{z}_i) - \frac{1}{m}\sum_{i=1}^m h(z_i) \;\middle|\; S\right]\right] \\
&=\; \mathsf{E}_S\left[\sup_{h\in\mathcal{F}} \mathsf{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^m \big(h(\tilde{z}_i) - h(z_i)\big) \;\middle|\; S\right]\right]
\end{aligned}$$

Since sup is a convex function, we can apply Jensen's Inequality to move the sup inside the expectation:

$$\mathsf{E}_S\left[\sup_{h\in\mathcal{F}}\mathsf{E}_{\tilde{S}}\left[\frac{1}{m}\sum_{i=1}^{m}\big(h(\tilde{z}_i)-h(z_i)\big)\;\middle|\;S\right]\right] \;\leq\; \mathsf{E}_{S,\tilde{S}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\big(h(\tilde{z}_i)-h(z_i)\big)\right]$$

Multiplying each term in the summation by a Rademacher variable σ_i will not change the expectation since $\mathsf{E}[\sigma_i]=0$. Furthermore, negating a Rademacher variable does not change its distribution. Combining these two facts,

$$\begin{aligned}
\mathsf{E}_{S,\tilde{S}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\big(h(\tilde{z}_i)-h(z_i)\big)\right] &= \mathsf{E}_{\sigma,S,\tilde{S}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i\big(h(\tilde{z}_i)-h(z_i)\big)\right]\\
&\leq \mathsf{E}_{\sigma,S,\tilde{S}}\left[\sup_{h\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(z_i)\right)+\sup_{h\in\mathcal{F}}\left(\frac{1}{m}\sum_{i=1}^{m}-\sigma_i h(\tilde{z}_i)\right)\right]\\
&= \mathsf{E}_{\sigma,S}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(z_i)\right] \;+\; \mathsf{E}_{\sigma,\tilde{S}}\left[\sup_{h\in\mathcal{F}}\frac{1}{m}\sum_{i=1}^{m}\sigma_i h(\tilde{z}_i)\right]\\
&= 2R_m(\mathcal{F})
\end{aligned}$$

Substituting this bound into (4) gives us exactly the desired result (1).

To obtain (2), we only need to notice that $\hat{R}_m(\mathcal{F})$ satisfies the precondition for McDiarmid's Inequality with constant $\frac{1}{m}$. A second application of McDiarmid's Inequality (now using confidence $\delta/2$ for each) bounds $\hat{R}_m(\mathcal{F})$ in terms of its expectation $R_m(\mathcal{F})$ and completes the proof. □

1.3.3 Connection to loss functions and error

Example 1. Let $X=\mathbb{R}^d$, $Y=\{-1,1\}$, and $Z=X\times Y$. For a concept class $\mathcal{H}\subseteq\{h:X\to Y\}$ we can let $L(\mathcal{H})=\{\ell_h|h\in\mathcal{H}\}$ where $\ell_h:Z\to\mathbb{R}$ is a loss funtion corresponding to the classifier h. This allows us to use Theorem 2 to obtain bounds on the misclassification rate of any hypothesis from $\mathcal{H}$. In particular, if we let $\mathcal{H}$ be the class of linear separators and $L(\mathcal{H})$ be the corresponding class of 0-1 loss functions, i.e. $\ell_h(z)=\ell_h(x,y)=\mathbb{1}_{h(x)\neq y}=\frac{1-yh(x)}{2}$ for each $h\in H$, then

$$\mathsf{E}_D[\ell_h(z)]=\mathsf{E}_D[\mathbb{1}_{h(x)\neq y}]=\mathsf{err}_D(h)$$

and

$$\hat{\mathsf{E}}_S[\ell_h(z)]=\frac{1}{m}\sum_{i=1}^{m}\mathbb{1}_{h(x_i)\neq y_i}=\widehat{\mathsf{err}}_S(h).$$

Taking $\mathcal{F}=L(\mathcal{H})$, Theorem 2 states that

$$\mathsf{err}_D(h)\;\leq\;\widehat{\mathsf{err}}_S(h)+2R_m(L(\mathcal{H}))+\sqrt{\frac{\ln(1/\delta)}{2m}}$$

which means we can bound the generalization error of a hypothesis in terms of its empirical error and the Rademacher complexity of the class of loss functions. In fact, Property 1 tells us that $R_m(L(\mathcal{H}))=\frac{1}{2}R_m(\mathcal{H})$, so we can ignore the loss function and write

$$\mathsf{err}_D(h)\;\leq\;\widehat{\mathsf{err}}_S(h)+R_m(\mathcal{H})+\sqrt{\frac{\ln(1/\delta)}{2m}}.$$

1.4 Recovering the VC Bound

Our next result will show how the general sample complexity result shown in the previous section can be used to recover the VC generalization bound shown earlier in the course. To do this, we will first need to prove two more facts: a basic property of Rademacher complexity regarding scalar multiplication and translation of function classes and another standard inequality known as Hoeffding's Inequality.

1.4.1 Preliminaries

Property 1. *Given any function class $\mathcal{F}$ and constants $a, b \in \mathbb{R}$, denote the function class $\{g | g(x) = af(x) + b\}$ by $a\mathcal{F} + b$. Then*

$$\hat{R}_m(a\mathcal{F} + b) = |a|\hat{R}_m(\mathcal{F})$$

and

$$R_m(a\mathcal{F} + b) = |a|R_m(\mathcal{F})$$

Proof. By definition, the empirical Rademacher complexity of $a\mathcal{F} + b$ is given by

$$\begin{aligned}
\hat{R}_m(a\mathcal{F} + b) &= \mathsf{E}_\sigma \left[\sup_{g \in a\mathcal{F}+b} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i g(z_i) \right) \right] \\
&= \mathsf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i (af(z_i) + b) \right) \right] \\
&= \mathsf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i af(z_i) + \frac{1}{m} \sum_{i=1}^{m} \sigma_i b \right) \right] \\
&= |a| \mathsf{E}_\sigma \left[\sup_{f \in \mathcal{F}} \left(\frac{1}{m} \sum_{i=1}^{m} \sigma_i f(z_i) \right) \right] \\
&= |a| \hat{R}_m(\mathcal{F})
\end{aligned}$$

and the analogous statement for $R_m(a\mathcal{F} + b)$ follows immediately by linearity of expectation. □

Theorem 3 (Hoeffding's Inequality)**.** *If X is a random variable with $\mathsf{E}[X] = 0$ and $a \leq X \leq b$, then for any real $s > 0$,*

$$\mathsf{E}[e^{sX}] \leq e^{s^2(b-a)^2/8}$$

Proof. Since e^{sx} is convex, we can write

$$e^{sx} \leq \frac{x-a}{b-a} e^{sb} + \frac{b-x}{b-a} e^{sa}.$$

Letting $p = -\frac{a}{b-a}$ and $\varphi(u) = -pu + \ln(1 - p + pe^u)$, we have

$$\begin{aligned}
\mathsf{E}[e^{sX}] &\leq \frac{b}{b-a} e^{sa} - \frac{a}{b-a} e^{sb} \\
&= (1-p) e^{-ps(b-a)} + p e^{s(b-a)} e^{-ps(b-a)} \\
&= \left((1-p) + p e^{s(b-a)} \right) e^{-ps(b-a)} \\
&= e^{\varphi(s(b-a))}
\end{aligned}$$

To complete the proof, we only need to obtain an upper bound for $\varphi(u)$, which we will do using a Taylor expansion. We compute

$$\varphi'(u) \;=\; -p + \frac{pe^u}{1-p+pe^u} \;=\; -p + \frac{p}{p+(1-p)e^{-u}}$$

and

$$\varphi''(u) \;=\; \frac{p(1-p)e^{-u}}{(p+(1-p)e^{-u})^2} \;\leq\; \frac{1}{4},$$

and then for some $t \in [0,u]$,

$$\begin{aligned}\varphi(u) \;&=\; \varphi(0) + u\varphi'(0) + \frac{u^2}{2}\varphi''(t)\\ &\leq\; 0 + 0 + \frac{u^2}{2}\left(\frac{1}{4}\right)\\ &\leq\; \frac{u^2}{8}\end{aligned}$$

This gives us $\varphi(s(b-a)) \leq \frac{s^2(b-a)^2}{8}$, which yields the desired result. □

1.4.2 Bounding the Rademacher complexity

Now we have the tools needed to prove the following theorem.

Theorem 4. *For any $A \subseteq \mathbb{R}^m$, let $R = \sup_{a\in A}\left(\sum_{i=1}^m a_i^2\right)^{1/2}$. Then*

$$\hat{R}_m(A) \;=\; \mathsf{E}_\sigma\left[\sup_{a\in A}\left(\frac{1}{m}\sum_{i=1}^m \sigma_i a_i\right)\right] \;\leq\; \frac{R\sqrt{2\ln|A|}}{m}$$

Proof. To set up the use of Hoeffding's Inequality, we start by taking the exponential of the empirical Rademacher complexity multiplied by some positive real constant s. By Jensen's Inequality,

$$\begin{aligned}\exp\left(s\mathsf{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^m \sigma_i a_i\right]\right) \;&\leq\; \mathsf{E}_\sigma\left[\exp\left(s\sup_{a\in A}\sum_{i=1}^m \sigma_i a_i\right)\right]\\ &=\; \mathsf{E}_\sigma\left[\sup_{a\in A}\left(\exp\left(s\sum_{i=1}^m \sigma_i a_i\right)\right)\right]\\ &\leq\; \sum_{a\in A}\mathsf{E}_\sigma\left[\exp\left(s\sum_{i=1}^m \sigma_i a_i\right)\right]\\ &=\; \sum_{a\in A}\mathsf{E}_\sigma\left[\prod_{i=1}^m \exp(s\sigma_i a_i)\right]\\ &=\; \sum_{a\in A}\prod_{i=1}^m \mathsf{E}_\sigma[\exp(s\sigma_i a_i)]\end{aligned}$$

where the last step uses the fact that the σ_i are independent. Now we can apply Hoeffding's Inequality since $\mathsf{E}_\sigma[\sigma_i a_i] = 0$ and $\sigma_i a_i \in [\alpha, \beta]$ where $\beta - \alpha = 2a_i$. This gives us

$$\begin{aligned}
\exp\left(s\mathsf{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^m \sigma_i a_i\right]\right) &\leq \sum_{a\in A}\prod_{i=1}^m \exp\left(\frac{s^2(2a_i)^2}{8}\right) \\
&= \sum_{a\in A}\exp\left(\frac{s^2}{2}\sum_{i=1}^m a_i^2\right) \\
&\leq |A|\exp\left(\frac{s^2R^2}{2}\right)
\end{aligned}$$

which means that

$$\mathsf{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^m \sigma_i a_i\right] \leq \frac{\ln|A|}{s} + \frac{sR^2}{2}.$$

Denote the right-hand side by $w(s)$. We would like to find the s that minimizes $w(s)$. Taking its derivative and setting it equal to zero, we find

$$w'(s) = -\frac{\ln|A|}{s^2} + \frac{R^2}{2} = 0 \quad \Rightarrow \quad s = \frac{\sqrt{2\ln|A|}}{R}.$$

Substituting this quantity back into the previous bound gives us

$$\begin{aligned}
\mathsf{E}_\sigma\left[\sup_{a\in A}\sum_{i=1}^m \sigma_i a_i\right] &\leq \frac{R\ln|A|}{\sqrt{2\ln|A|}} + \frac{R^2\sqrt{2\ln|A|}}{2R} \\
&= R\sqrt{2\ln|A|}
\end{aligned}$$

Dividing both sides by m yields the result stated in the theorem. □

1.4.3 Connection to VC theory

Example 2. For any finite concept class $\mathcal{H} \subseteq \{h : X \to \{-1, 1\}\}$ and example set $S = \{x_1, \ldots, x_m\}$, we can take $A = \{(h(x_1), \ldots, h(x_m)) | h \in \mathcal{H}\}$. Then $|A| = |\mathcal{H}|$ and $R = \sqrt{m}$. From Theorem 4, this means that

$$\hat{R}_m(\mathcal{H}) \leq \sqrt{\frac{2\ln|\mathcal{H}|}{m}}.$$

In general, (whether $\mathcal{H}$ is finite or not), we can take $A = \mathcal{H}[S]$, the set of distinct labelings of points in S using concepts from $\mathcal{H}$. Then $|A| = \mathcal{H}[m]$, the shatter coefficient of $\mathcal{H}$ on m points, and

$$\hat{R}_m(\mathcal{H}) \leq \sqrt{\frac{2\ln\mathcal{H}[m]}{m}}.$$

By Sauer's Lemma, $\mathcal{H}[m] \leq m^d$ where d is the VC dimension of $\mathcal{H}$, so we can further simplify this result to

$$\hat{R}_m(\mathcal{H}) \leq \sqrt{\frac{2d\ln m}{m}}.$$

A Additional Proofs

Proof of Lemma 1. Let $S = \{z_1, \ldots, z_m\}$ and $S' = \{z_1, \ldots, z'_j, \ldots, z_m\}$. Then by definition,

$$|\varphi(S) - \varphi(S')| \;=\; \left| \sup_{h \in \mathcal{F}} \Big(\mathsf{E}_D[h(z)] - \hat{\mathsf{E}}_S[h(z)] \Big) - \sup_{h \in \mathcal{F}} \Big(\mathsf{E}_D[h(z)] - \hat{\mathsf{E}}_{S'}[h(z)] \Big) \right|.$$

Letting $h^* \in \mathcal{F}$ be the maximizing function for the supremum in $\varphi(S)$, this becomes

$$|\varphi(S) - \varphi(S')| \;=\; \left| \mathsf{E}_D[h^*(z)] - \hat{\mathsf{E}}_S[h^*(z)] - \sup_{h \in \mathcal{F}} \Big(\mathsf{E}_D[h(z)] - \hat{\mathsf{E}}_{S'}[h(z)] \Big) \right|,$$

and by definition of supremum, h^* can at best maximize the $\varphi(S')$ term as well, so we have

$$\begin{aligned} |\varphi(S) - \varphi(S')| \;&\leq\; \left| \mathsf{E}_D[h^*(z)] - \hat{\mathsf{E}}_S[h^*(z)] - \mathsf{E}_D[h^*(z)] + \hat{\mathsf{E}}_{S'}[h^*(z)] \right| \\ &=\; \left| \hat{\mathsf{E}}_{S'}[h^*(z)] - \hat{\mathsf{E}}_S[h^*(z)] \right| \\ &=\; \left| \frac{1}{m} \sum_{z \in S} h^*(z) - \frac{1}{m} \sum_{z \in S'} h^*(z_i) \right| \end{aligned}$$

Since S and S' differ in only one element, this becomes

$$\begin{aligned} |\varphi(S) - \varphi(S')| \;&\leq\; \frac{1}{m} \left| \sum_{i \neq j} h^*(z_i) - \sum_{i \neq j} h^*(z_i) + h^*(z_j) - h^*(z'_j) \right| \\ &=\; \frac{1}{m} \left| h^*(z_j) - h^*(z'_j) \right| \\ &\leq\; \frac{1}{m} \end{aligned}$$

where the last step follows from fact that $h^* : Z \to [a, a+1]$ so $\sup_{z_j, z'_j \in Z} |h^*(z_j) - h^*(z'_j)| = 1$. □