CSC 665: Rademacher complexity

Chicheng Zhang

September 17, 2019

1 Uniform convergence of empirical error to generalization error, revisited

In previous lectures, we have already established PAC learning guarantees for empirical risk minimization (ERM), when the hypothesis class $\mathcal{H}$ is finite. The strategy is the show that for all classifiers in $\mathcal{H}$, its empirical error concentrates around its generalization error with high probability, in other words, with probability $1-\delta$ over the choice of S, i.e. n iid samples from D,

$$\sup_{h\in\mathcal{H}} |\operatorname{err}(h,S) - \operatorname{err}(h,D)| = O\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{n}}\right). \tag{1}$$

This type of concentration is often called "uniform convergence" in learning theory literature, as long as the right hand side converges to 0 as m goes to infinity. Under uniform convergence, we can easily argue that with probability $1-\delta$, the ERM, $\hat{h}$, satisfies that

$$\operatorname{err}(\hat{h},D) - \min_{h'\in\mathcal{H}} \operatorname{err}(h',D) = O\left(\sqrt{\frac{\ln|\mathcal{H}| + \ln\frac{1}{\delta}}{n}}\right). \tag{2}$$

which is sufficient to ensure $\mathcal{H}$'s PAC learnability.

Now, turning back to the case when $\mathcal{H}$ is infinite. Under what conditions on $\mathcal{H}$ can we establish an analog of Equation (1) (hence establishing an analog of Equation (2))? In this note, we show that $\mathcal{H}$ having a finite VC dimension is sufficient to ensure uniform convergence.

Theorem 1. *There exists a numerical constant $c_1 > 0$ such that the following holds. Suppose hypothesis class $\mathcal{H}$ has VC dimension d. Then, given a set of n iid examples $(X_1,Y_1),\ldots,(X_n,Y_n)$ from distribution D, with probabilty $1-\delta$,*

$$\sup_{h\in\mathcal{H}} |\operatorname{err}(h,S) - \operatorname{err}(h,D)| \le c_1\sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{2}{\delta}}{n}}. \tag{3}$$

Consequently, ERM on $\mathcal{H}$ has an agnostic PAC sample complexity of

$$m(\epsilon,\delta) = O\left(\frac{1}{\epsilon^2}\left(d\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta}\right)\right).$$

To show Theorem 1, let us set up some useful notation, summarized in the following table.

Original notation	Shorthand notation	Explanation
$\mathcal{X} \times \mathcal{Y}$	$\mathcal{Z}$	instance space
$\{(X_i, Y_i)\}_{i=1}^n$	$\{Z_i\}_{i=1}^n$	training examples
$\mathbf{1}(h(x) \neq y)$	ℓ_h	0-1 loss associated with h
$\{\mathbf{1}(h(x) \neq y) : h \in \mathcal{H}\}$	$\mathcal{F} = \{\ell_h : h \in \mathcal{H}\}$	loss class associated with $\mathcal{H}$
$\mathrm{err}(h, D)$	$\mathbb{E}_{Z\sim D} f(Z)$ (abbrev. $\mathbb{E}_D f(Z)$)	generalization error of h
$\mathrm{err}(h, S)$	$\frac{1}{n}\sum_{i=1}^n f(Z_i) = \mathbb{E}_{Z\sim S} f(Z)$ (abbrev. $\mathbb{E}_S f(Z)$)	training error of h

Note that $\mathcal{F}$ is a mapping from $\mathcal{X} \times \mathcal{Y}$ to $\{0, 1\}$. Similar to the growth function of the original hypothesis class $\mathcal{H}$, we can also define the growth function of $\mathcal{F}$ to be

$$\mathcal{S}(\mathcal{F}, n) = \max_{z_1,\ldots,z_n} |\left\{(f(z_1), \ldots, f(x_n)) : f \in \mathcal{F}\right\}|.$$

We first have the following simple lemma that links the growth function of $\mathcal{F}$ and that of $\mathcal{H}$.

Lemma 1. $\mathcal{S}(\mathcal{F}, n) = \mathcal{S}(\mathcal{H}, n)$.

Proof. We first observe that for any set of labeled examples $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$,

$$\begin{aligned} & |\left\{(f(x_1, y_1), \ldots, f(x_n, y_n)) : f \in \mathcal{F}\right\}| \\ = \; & |\left\{(\mathbf{1}(h(x_1) \neq y_1), \ldots, \mathbf{1}(h(x_n) \neq y_n)) : f \in \mathcal{F}\right\}| \\ = \; & |\left\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\right\}| \end{aligned}$$

where the first equality is by the definition of $\mathcal{F}$, and the second equality is by observing that every labeling of h on $(x_1, \ldots, x_n)$ induces an unique pattern of misclassification on these n examples.

This implies that

$$\mathcal{S}(\mathcal{F}, n) = \max_{(x_1,y_1),\ldots,(x_n,y_n)} |\left\{(f(x_1, y_1), \ldots, f(x_n, y_n)) : f \in \mathcal{F}\right\}| = \max_{x_1,\ldots,x_n} |\left\{(h(x_1), \ldots, h(x_n)) : h \in \mathcal{H}\right\}| = \mathcal{S}(\mathcal{H}, n).$$

□

The following theorem is central to the proof of Theorem 1, which establishes uniform convergence of empirical loss to generalization loss for loss classes of small growth function. Its proof requires several important insights; we will defer it to the next section.

Theorem 2. *Suppose $Z_1, \ldots, Z_n$ is a set of iid examples, and $\mathcal{F} \subseteq (Z \to \{0, 1\})$ is the loss function class. Then, with probability $1 - \delta$,*

$$\sup_{f\in\mathcal{F}} |\mathbb{E}_S f(Z) - \mathbb{E}_D f(Z)| \leq \sqrt{\frac{32\left(\ln\frac{2}{\delta} + \ln\big(2\,\mathcal{S}(\mathcal{F}, n)\big)\right)}{n}}.$$

Proof of Theorem 1. Applying the notation of Table 1, and using Lemma 1, we immediately get that with probability $1 - \delta$,

$$\sup_{h\in\mathcal{H}} |\,\mathrm{err}(h, S) - \mathrm{err}(h, D)| \leq \sqrt{\frac{32\left(\ln\frac{2}{\delta} + \ln\big(2\,\mathcal{S}(\mathcal{H}, n)\big)\right)}{n}}.$$

By Sauer's Lemma, $\mathcal{S}(\mathcal{H}, n) \leq (\frac{en}{d})^d$, this implies that the right hand side is upper bounded by $c_1\sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{2}{\delta}}{n}}$ for some large constant c_1.

The sample complexity bound follows from the following observation: when Equation 3 holds, the ERM has excess error $2c_1\sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{2}{\delta}}{n}}$. Define $m(\epsilon, \delta) = \min\left\{m : 2c_1\sqrt{\frac{d\ln\frac{n}{d} + \ln\frac{2}{\delta}}{n}} \leq \epsilon\right\}$. It can be checked that $m(\epsilon, \delta) = O\left(\frac{1}{\epsilon^2}\left(d\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta}\right)\right)$.

This implies that when $n \geq m(\epsilon, \delta)$, with probability $1 - \delta$, ERM has excess error at most ϵ. □

2 Proof of Theorem 2

Before the actual proof, let us collect a few elementary but useful facts about supremum in the following lemma.

Lemma 2. *The following inequalities hold:*

1. $$\sup_{f\in\mathcal{F}}\big(A(f)+B(f)\big)\le\sup_{f\in\mathcal{F}}A(f)+\sup_{f\in\mathcal{F}}B(f).$$
   Equvalently,
   $$\sup_{f\in\mathcal{F}}C(f)\le\sup_{f\in\mathcal{F}}D(f)+\sup_{f\in\mathcal{F}}\big(C(f)-D(f)\big).$$

2. $$\sup_{f\in\mathcal{F}}\mathbb{E}\left[A(f)\right]\le\mathbb{E}\left[\sup_{f\in\mathcal{F}}A(f)\right].$$

If the sup's were max's, you should be able to prove these using basic algebra.[1] For completeness, we include a proof of this lemma in Appendix A.

Step 1: concentration. Let us first view $\sup_{f\in\mathcal{F}}\big|\mathbb{E}_S f(Z)-\mathbb{E}_D f(Z)\big|$ as a function of $Z_1,\dots,Z_n$, specifically, $g(Z_1,\dots,Z_n)$ where

$$g(z_1,\dots,z_n)=\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(z_i)-\mathbb{E}_D f(Z)\right|.$$

We claim g is $\frac{1}{n}$-sensitive. Indeed, for every coordinate i, consider an alternative input $z^{(i)}=(z_1,\dots,z_{i-1},z_i,z_{i+1},z_n)$, Denote by $M(f,z)\triangleq\big|\frac{1}{n}\sum_{i=1}^n f(z_i)-\mathbb{E}_D f(Z)\big|$.

$$M(f,z)-M(f,z^{(i)})\le\left|\frac{1}{n}(f(z_i)-f(z_i'))\right|\le\frac{1}{n}.$$

Hence,

$$\sup_{f\in\mathcal{F}}M(f,z)\le\sup_{f\in\mathcal{F}}M(f,z^{(i)})+\sup_{f\in\mathcal{F}}\Big(M(f,z)-M(f,z^{(i)})\Big)\le\sup_{f\in\mathcal{F}}M(f,z^{(i)})+\frac{1}{n},$$

where the first inequality is from item 1 of Lemma 2.

Applying McDiarmid's Inequality, with probability $1-\delta$,

$$\sup_{f\in\mathcal{F}}\big|\mathbb{E}_S f(Z)-\mathbb{E}_D f(Z)\big|\le\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Z_i)-\mathbb{E}_D f(Z)\right|+\sqrt{\frac{\ln\frac{2}{\delta}}{2n}}.$$

Therefore, the proof reduces to upper bounding the first term (expected maximum deviation), i.e.

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Z_i)-\mathbb{E}_D f(Z)\right|. \tag{4}$$

The expectation is over a supremum of an infinite collection of random variables (each one is associated with a function f in $\mathcal{F}$), which is a bit difficult to deal with. Our high-level strategy for the remaining steps, is to reduce the problem to bounding the expectation over a supremum of a finite collection of random variables.

[1] In fact, throughout this course, it is OK to think about the sup and inf's as max and min's under all circumstances.

Step 2: double sampling trick (transduction). For the moment, let us fix a set of training examples $S = z_1, \ldots, z_n$. Suppose that we obtained a fresh set of iid samples $S' = \left\{Z'_1, \ldots, Z'_n\right\}$ independent of S (we can think of S' as a set of validation example - the goal is to ensure for all classifiers, its training loss is close to its validation loss). Observe that $\mathbb{E}_D f(Z) = \mathbb{E}\frac{1}{n}\sum_{i=1}^n f(Z'_i)$.

Now, the term within the expectation of Equation (4) can be upper bounded as:

$$\begin{aligned}
& \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(z_i) - \mathbb{E}_D f(Z)\right| \\
= & \sup_{f\in\mathcal{F}}\left|\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n f(z_i) - \frac{1}{n}\sum_{i=1}^n f(Z'_i)\right]\right| \\
\leq & \sup_{f\in\mathcal{F}}\mathbb{E}\left|\frac{1}{n}\sum_{i=1}^n f_0(z_i) - \frac{1}{n}\sum_{i=1}^n f_0(Z'_i)\right| \\
\leq & \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(z_i) - \frac{1}{n}\sum_{i=1}^n f(Z'_i)\right|
\end{aligned}$$

where in the last three lines, the expectation is over the random draw of S'. The first inequality uses Jensen's Inequality and the convexity of $|x|$, and the second inequality uses item 2 of Lemma 2.

Now, we consider the randomness in training sample S. The above implies that,

$$\begin{aligned}
& \mathbb{E}_{S\sim D^n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}_D f(Z)\right| \\
\leq & \mathbb{E}_{S\sim D^n, S'\sim D^n}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Z_i) - \frac{1}{n}\sum_{i=1}^n f(Z'_i)\right| \\
= & \frac{1}{n}\mathbb{E}_{S\sim D^n, S'\sim D^n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(Z_i) - f(Z'_i))\right|, \qquad (5)
\end{aligned}$$

Here, note that for different realizations of S and S', the f that achieves the supremum can still be drastically different - therefore, we are still dealing with an infinite collection of random variables.

Step 3: symmetrization. Now here comes the crucial observation: define function

$$h(z_1, z'_1, \ldots, z_n, z'_n) = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}|\sum_{i=1}^n \sigma_i(f(z_i) - f(z'_i))|.$$

As the $2n$ random variables $(Z_1, Z'_1, Z_2, Z'_2, \ldots, Z_n, Z'_n)$ has exactly the same distribution law as, say, $(Z'_1, Z_1, Z_2, Z'_2, \ldots, Z_n, Z'_n)$ (switching the order of the first two samples), this implies that

$$\mathbb{E}h(Z_1, Z'_1, Z_2, Z'_2, \ldots, Z_n, Z'_n) = \mathbb{E}h(Z'_1, Z_1, Z_2, Z'_2, \ldots, Z_n, Z'_n).$$

Hence,

$$\frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n (f(Z_i) - f(Z'_i))\right| = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|(f(Z'_1) - f(Z_1)) + \sum_{i=2}^n (f(Z_i) - f(Z'_i))\right|.$$

More generally, for any fixed $(\sigma_1, \ldots, \sigma_n)$ in $\{-1, +1\}^n$,

$$\frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(f(Z_i) - f(Z_i'))\right| = \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=2}^{n}\sigma_i(f(Z_i) - f(Z_i'))\right|. \tag{6}$$

Suppose σ_i's are random variables drawn iid from R, i.e. $\mathrm{U}(\{-1, +1\})$ [2] (which is called the Rademacher distribution), by taking expectations over σ_i's on both sides of Equation (6), we have that

$$\begin{aligned}
& \frac{1}{n}\mathbb{E}_{S,S'\sim D^n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}(f(Z_i) - f(Z_i'))\right| \\
= \; & \frac{1}{n}\mathbb{E}_{S,S'\sim D^n,\sigma\sim\mathrm{R}^n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i(f(Z_i) - f(Z_i'))\right| \\
\leq \; & \frac{1}{n}\mathbb{E}_{S\sim D^n,\sigma\sim\mathrm{R}^n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(Z_i)\right| + \frac{1}{n}\mathbb{E}_{S'\sim D^n,\sigma\sim\mathrm{R}^n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(Z_i')\right| \\
= \; & \frac{2}{n}\mathbb{E}_{S\sim D^n,\sigma\sim\mathrm{R}^n}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(Z_i)\right|
\end{aligned} \tag{7}$$

where the first inequality uses the fact that $|A + B| \leq |A| + |B|$, and item 1 of Lemma 2, and the second equality uses the fact that S and S' come from the same distribution.

Definition 1. *The empirical Rademacher complexity of $\mathcal{F}$ with respect to sample S of size n, $\mathrm{Rad}_S(\mathcal{F})$, is defined as*

$$\mathrm{Rad}_S(\mathcal{F}) \triangleq \frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^{n}\sigma_i f(Z_i)\right|,$$

where the expectation is over $\sigma \sim \mathrm{R}^n$. The Rademacher complexity of $\mathcal{F}$ with respect to distribution D with sample size n, denoted as $\mathrm{Rad}_n(\mathcal{F})$, is defined as

$$\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\,\mathrm{Rad}_S(\mathcal{F}),$$

where the expectation is over $S \sim D^n$.

Using the above notation, the right hand side of Equation (7) can be written as $2\,\mathrm{Rad}_n(\mathcal{F})$.

Step 4: controlling empirical Rademacher complexities. To upper bound $\mathrm{Rad}_n(\mathcal{F})$, it suffices to give an uniform upper bound of $\mathrm{Rad}_S(\mathcal{F})$ for every fixed sample S of size n. Note that when S is fixed, there are at most $\mathcal{S}(\mathcal{F}, n)$ realizations of $(f(Z_1), \ldots, f(Z_n))$, where are elements of $\Pi_{\mathcal{F}}(S)$. Therefore,

$$\mathrm{Rad}_S(\mathcal{F}) = \frac{1}{n}\mathbb{E}\max_{(a_1,\ldots,a_n)\in\Pi_{\mathcal{F}}(S)}\left|\sum_{i=1}^{n}\sigma_i a_i\right|$$

Observe that we have successfully "tamed" an infinite collection of random variables to only a finite collection! It turns out that there is a classical lemma that can bound the expectation of the maximum of a finite collection of random variables, stated as follows:

[2] Here $\mathrm{U}(A)$ stands for the uniform distribution over set A.

Lemma 3 (Massart's Lemma)**.** *Suppose A is a finite subset of $\mathbb{R}^n$, and for all a in A, $\|a\|_2 \le R$. Then,*

$$\mathbb{E}\left[\max_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right] \le 2R\sqrt{\ln|A|}.$$

We now apply Massart's Lemma to our setting. Consider

$$A = \left\{(a_1,\dots,a_n) : a \in \Pi_{\mathcal{F}}(S)\right\} \cup \left\{(-a_1,\dots,-a_n) : a \in \Pi_{\mathcal{F}}(S)\right\},$$

we know that $|A| \le 2\,\mathcal{S}(\mathcal{F},n)$. In addition, for all $a \in A$, $\|a\|_2 = \sqrt{\sum_{i=1}^n a_i^2} \le \sqrt{n}$.

Therefore,

$$\mathbb{E}\max_{(a_1,\dots,a_n)\in\Pi_{\mathcal{F}}(S)}\left|\sum_{i=1}^{n}\sigma_i a_i\right| \le \mathbb{E}\max_{a\in A}\left|\sum_{i=1}^{n}\sigma_i a_i\right| \le 2\sqrt{n\ln\bigl(2\,\mathcal{S}(\mathcal{F},n)\bigr)}$$

This implies that $\mathrm{Rad}_S(\mathcal{F}) \le 2\sqrt{\frac{\ln(2\,\mathcal{S}(\mathcal{F},n))}{n}}$, and consequently, $\mathrm{Rad}_n(\mathcal{F}) = \mathbb{E}\,\mathrm{Rad}_S(\mathcal{F}) \le 2\sqrt{\frac{\ln(2\,\mathcal{S}(\mathcal{F},n))}{n}}$.

In summary, we have shown that with probability $1-\delta$,

$$\begin{aligned}
&\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i) - \mathbb{E}_D f(Z)\right| \\
&\le \mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}f(Z_i) - \mathbb{E}_D f(Z)\right| + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \\
&\le 2\,\mathrm{Rad}_n(\mathcal{F}) + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \\
&\le 4\sqrt{\frac{\ln\bigl(2\,\mathcal{S}(\mathcal{F},n)\bigr)}{n}} + \sqrt{\frac{\ln\frac{2}{\delta}}{2n}} \\
&\le \sqrt{\frac{32(\ln\bigl(2\,\mathcal{S}(\mathcal{F},n) + \ln\frac{2}{\delta}\bigr)}{n}},
\end{aligned}$$

where the last inequality uses the elementary fact that $\sqrt{A}+\sqrt{B} \le \sqrt{2(A+B)}$. □

Proof of Lemma 3. We use the machinery of moment generating functions developed in the lectures on concentration inequalities. First, observe that for any $t > 0$,

$$\begin{aligned}
\max_{a\in A}\sum_{i=1}^{n}\sigma_i a_i &= \frac{\max_{a\in A}\sum_{i=1}^{n} t\sigma_i a_i}{t} \\
&= \frac{\ln\left(\exp\{\max_{a\in A}\sum_{i=1}^{n} t\sigma_i a_i\}\right)}{t} \\
&\le \frac{\ln\left(\sum_{a\in A}\exp\{\sum_{i=1}^{n} t\sigma_i a_i\}\right)}{t}
\end{aligned}$$

Now, taking expectation on both sides, and note that $\ln(x)$ is a concave function, applying Jensen's inequality, we get that for any $t > 0$,

$$\mathbb{E}\left[\max_{a\in A}\sum_{i=1}^{n}\sigma_i a_i\right] \le \frac{\ln\left(\sum_{a\in A}\mathbb{E}\exp\{\sum_{i=1}^{n}t\sigma_i a_i\}\right)}{t} \tag{8}$$

For each a in A, let's look at the term $\mathbb{E}\exp\{\sum_{i=1}^{n}t\sigma_i a_i\}$. Observe that the exponent is a sequence of independent random variables, therefore, we can decompose them to $\prod_{i=1}^{n}\mathbb{E}\exp\{t\sigma_i a_i\}$. But we know how to bound each factor: by the key lemma in proving Hoeffding's Inequality, we know that for a zero-mean random variable X with range c, its moment generating function $\mathbb{E}e^{tX}$ is at most $\exp\left\{\frac{c^2t^2}{8}\right\}$. This implies that for all a in A, as random variable $\sigma_i a_i$ has mean zero and range $2a_i$,

$$\mathbb{E}\exp\left\{\sum_{i=1}^{n}t\sigma_i a_i\right\} = \prod_{i=1}^{n}\mathbb{E}\exp\{t\sigma_i a_i\} \le \exp\left\{\frac{t^2}{2}\cdot\sum_{i=1}^{n}a_i^2\right\} \le \exp\left\{\frac{t^2R^2}{2}\right\}.$$

Coming back to Equation (8), we have that the right hand side is at most

$$\frac{\ln|A| + t^2R^2/2}{t} = \frac{\ln|A|}{t} + \frac{tR^2}{2}.$$

As Equation (8) holds for any $t > 0$, we choose $t = \sqrt{\frac{2\ln|A|}{R}}$ to minimze the right hand side, which is $2R\sqrt{\ln|A|}$. □

A Proof of Lemma 2

We show the two items respectively.

1. For every $\epsilon > 0$, there exists an f_0 in $\mathcal{F}$ such that

   $$A(f_0) + B(f_0) \ge \sup_{f\in\mathcal{F}}\big(A(f) + B(f)\big) - \epsilon.$$

   As f_0 is in $\mathcal{F}$, it can be easily seen that,

   $$A(f_0) + B(f_0) \le \sup_{f\in\mathcal{F}}A(f) + \sup_{f\in\mathcal{F}}B(f).$$

   Combining the above two inequalities, this implies that for any $\epsilon > 0$,

   $$\sup_{f\in\mathcal{F}}\big(A(f) + B(f)\big) \le \sup_{f\in\mathcal{F}}A(f) + \sup_{f\in\mathcal{F}}B(f) + \epsilon.$$

   Taking $\epsilon \to 0$ on both sides of the above inequality, we get the first inequality. The second inequality follows by setting $A(f) = C(f)$ and $B(f) = C(f) - D(f)$.

2. For every $\epsilon > 0$, there exists an f_0 in $\mathcal{F}$ such that

   $$\mathbb{E}\left[A(f_0)\right] \ge \sup_{f\in\mathcal{F}}\mathbb{E}\left[A(f)\right] - \epsilon.$$

   As f_0 is in $\mathcal{F}$, it can be easily seen that,

   $$\mathbb{E}\left[A(f_0)\right] \le \mathbb{E}\left[\sup_{f\in\mathcal{F}}A(f)\right].$$

Combining the above two inequalities, this implies that for any $\epsilon > 0$,

$$\sup_{f \in \mathcal{F}} \mathbb{E}\left[A(f)\right] \leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} A(f)\right] + \epsilon.$$

Taking $\epsilon \to 0$ on both sides of the above inequality, we get the item.