The goal of this lecture is to establish risk bounds that depend on the learning algorithm $A$ instead of the hypothesis class $H$. In particular, we would prove results that look like

With probability at least $1 - \delta$ over the sample set $S \sim D^m$,

$$|\ell(A(S); D) - \ell(A(S); S)| \leq \epsilon$$

i.e. we compare the performance of the same function, $A(S)$, on two different distributions, the uniform distribution on the sample set $S$ and the real distribution $D$. Contrast this to the approach we took in the Rademacher and VC theory, where we compared the performance of two functions, the output of the algorithm $A(S)$ and the best hypothesis $h^*$, on the same distribution $D$.

Notice two simple facts. First, $A(S)$ is a random variable, with the randomness comes from $S$. Second, $A(S)$ depends on $S$, so we cannot use Hoeffding’s inequality to compare $\ell(A(S); D)$ and $\ell(A(S); S)$.

## 1 Uniform stability

Before getting to the formal definitions, we introduce some notations

- $S^i = S \setminus \{x_i, y_i\}$, i.e. the set of $m - 1$ samples where the $i$th sample is removed.
- $S^i = S^i \cup \{x_i', y_i'\}$ for some worst $(x_i', y_i').$

**Definition 1 (Uniform stability).** Algorithm $A$ has uniform stability $\beta$ with respect to a loss function $\ell$ if for all $S$ and all $i \in [m],$

$$\max_{x,y} |\ell(A(S); (x, y)) - \ell(A(S^i); (x, y))| \leq \beta$$

i.e. the algorithm is "stable" with respect to removing a single sample at all points.

Note that $\beta$ depends on $m$ and we would want $\beta_m$ to be around $\frac{1}{m}$.

**Remark 1.** There are weaker notions of stability such as

- **Error stability:** For all $S$, for all $i \in [m]: |\ell(A(S); D) - \ell(A(S^i); D)| \leq \beta$, i.e. the average difference is small. This is a very week notion of stability.

- **Hypothesis stability:** For all $i \in [m]$, $E_{S,(x,y)} \left[ |\ell(A(S); (x, y)) - \ell(A(S^i); (x, y))| \right] \leq \beta$

**Remark 2.** By triangle inequality, uniform stability $\beta$ implies that for $S$, for all $i \in [m]$

$$\max_{x,y} |\ell(A(S); (x, y)) - \ell(A(S^i); (x, y))| \leq 2\beta$$

Now we show that uniform stability implies a bound in the form stated in the first paragraph.

**Theorem 2.** Suppose that $\ell \in [0, C]$ and $A$ has uniform stability $\beta$. Furthermore, suppose that $A$ is anonymous, i.e. $A(S) = A(S')$ if $S$ and $S'$ contains the same elements (but in different orders). Let $Z$ be the random variable defined by $Z = \ell(A(S); D) - \ell(A(S); S)$. Then w.p. at least $1 - \delta$,

$$Z \leq 2\beta + m(4\beta + C/m)\sqrt{\frac{\log(1/\delta)}{2m}}$$
Proof. We will show that $Z$ is concentrated, and then show that $E[Z]$ is small. For the first part, note that by triangle inequality:

$$|\ell(A(S); D) - \ell(A(S'); D)| \leq |\ell(A(S); D) - \ell(A(S'); D)| + |\ell(A(S); D) - \ell(A(S'); D)| \leq 2\beta$$

Also:

$$|\ell(A(S); S) - \ell(A(S'); S')| \leq \frac{1}{m} \sum_{j \neq i} |\ell(A(S); (x_j, y_j)) - \ell((A)(S'); (x_j, y_j))|$$

$$+ \frac{1}{m} |\ell(A(S); (x_i, y_i)) - \ell(A(S'); (x'_i, y'_i))|$$

$$\leq \frac{m - 1}{m} 2\beta + \frac{C}{m} \leq 2\beta + \frac{C}{m}$$

Hence, $|Z - Z'| \leq 4\beta + \frac{C}{m}$ where $Z'$ denotes the random variable where $S$ is replaced by $S'$. This implies that $\max_S Z - \min_S Z \leq m(4\beta + C/m)$; therefore $Z \leq E[Z] + m(4\beta + C/m)$.

To bound $E[Z]$, we will need some identities

- $E_S[\ell(A(S); (x_j, y_j))] = E_{S,(x', y')}[\ell(A(S'); (x', y'))]$. This identity holds because for $(x', y')$ drawn from $D$, $\Pr[S] = \Pr[S']$.

- $E_{S,(x', y')}[\ell(A(S); (x', y'))] = E_{S,(x', y')}[\ell(A(S'); (x', y'))]$. To see that this identity holds, let $S'$ be the set that contains the same elements as $S$ but with $(x_i, y_i)$ and $(x_j, y_j)$ exchange their order. Then since $A$ is anonymous, $A(S) = A(S')$. The identity then follow from the fact that $\Pr[S] = \Pr[S']$ and the previous identity.

With these identities, we have:

$$E[Z] = E_S[\ell(A(S); D) - \ell(A(S); S)]$$

$$= E_{S,(x', y')}[\ell(A(S); (x', y'))] - E_S[\ell(A(S); S)]$$

$$= E_{S,(x', y')}[\ell(A(S); (x', y'))] - \frac{1}{m} \sum_{j=1}^{m} E_S[\ell(A(S); (x_j, y_j))]$$

$$= E_{S,(x', y')}[\ell(A(S); (x', y'))] - \frac{1}{m} \sum_{j=1}^{m} E_S[\ell(A(S); (x_j, y_j))]$$

$$= E_{S,(x', y')}[\ell(A(S); (x', y'))] - \frac{1}{m} \sum_{j=1}^{m} E_{S,(x', y')}[\ell(A(S); (x', y'))]$$

$$E_{S,(x', y')}[\ell(A(S); (x', y'))] - \frac{1}{m} m E_{S,(x', y')}[\ell(A(S); (x', y'))]$$

$$E_{S,(x', y')}[\ell(A(S); (x', y'))] - \ell(A(S'; (x', y')))]$$

$$\leq 2\beta$$

Now, by McDiarmid’s inequality, we have that with probability at least $1 - \delta$,

$$Z \leq E[Z] + m(4\beta + c/m) \sqrt{\frac{\log(1/\delta)}{m}} \leq 2\beta + m(4\beta + c/m) \sqrt{\frac{\log(1/\delta)}{m}}.$$

The implication of this theorem is that if $A$ is uniformly stable and has good empirical risk, we have good bound on $A$’s risk even if the hypothesis class $\mathcal{H}$ is bad.
2 Regularized ERM

We give an example algorithm that has uniform stability. Consider the case where $y \in \{-1, 1\}$, is convex and differentiable. For simplicity, assume that $\ell(h; (x, y)) = \ell(yh(x))$, so $\ell$ is $\lambda$-lipschitz. (log loss is a nice loss here; hinge loss is not differentiable, but there’s a fix for that.) The algorithm we will use is Regularized ERM, which outputs the hypothesis that minimize the following quantity:

$$R(h) = \ell(h; S) + c\psi(h)$$

where $\psi(h)$ is some convex and differentiable function from $H$ to $\mathbb{R}$ and $c$ is some constant.

Consider the following quantities

- $\bar{h} = \arg\min_h R(h)$
- $R^{\lambda}(h) = \ell(h; S^{\lambda}) + c\psi(h)$
- $\bar{h}^{\lambda} = \arg\min_h R^\lambda(h)$

Note that $\bar{h}^{\lambda}$ is the output of the algorithm on $S^{\lambda}$. Therefore $\beta = \max_{x,y} \| \ell(h; (x, y)) - \ell(\bar{h}^{\lambda}; (x, y)) \|$ is exactly the quantity we want to bound. We have:

$$\beta = \max_{x,y} \| \ell(h; (x, y)) - \ell(\bar{h}^{\lambda}; (x, y)) \| \leq \max_{x,y} \| y\bar{h}(x) - y\bar{h}^{\lambda}(x) \| = \max_{x} \lambda \| \bar{h}(x) - \bar{h}^{\lambda}(x) \|. \tag{1}$$

Now we consider a concrete hypothesis class $H = \{ w \in \mathbb{R}^n \}$ and domain $\mathcal{X} = \{ x \in R^n : \|x\| \leq X \}$. Let $\psi(w) = \| w \|^2$.

To prove a concrete bound on $\beta$, we will use the Bregman divergence defined as follows.

**Definition 3.** The Bregman divergence of a function $f$ is defined by $B_f(v', v) = f(v') - f(v) - \langle v' - v, \nabla f(v) \rangle$.

**Lemma 4.** If $f$ is convex then $B_f \geq 0$.

**Lemma 5.** $cB_\psi(\bar{h}^{\lambda}||\bar{h}) + cB_\psi(\bar{h}||\bar{h}^{\lambda}) \leq \frac{2\lambda}{m} \max_x \| \bar{h}^{\lambda}(x) - \bar{h}(x) \|.$

**Proof.** We have:

$$cB_\psi(\bar{h}^{\lambda}||\bar{h}) + cB_\psi(\bar{h}||\bar{h}^{\lambda}) \leq cB_R(\bar{h}^{\lambda}||\bar{h}) + cB_R(\bar{h}||\bar{h}^{\lambda})$$

$$= R(\bar{h}^{\lambda}) - R(\bar{h}) + R(\bar{h}) - R(\bar{h}^{\lambda})$$

$$= \frac{1}{m} \left( \ell(\bar{h}^{\lambda}; (x_i, y_i)) - \ell(\bar{h}; (x_i, y_i)) \right) + \frac{1}{m(m-1)} \sum_{j \neq i} \left( \ell(\bar{h}; (x_j, y_j)) - \ell(\bar{h}^{\lambda}; (x_j, y_j)) \right)$$

$$\leq \frac{\lambda}{m} \| \bar{h}^{\lambda}(x_i) - \bar{h}(x_i) \| + \frac{\lambda}{m(m-1)} \sum_{j \neq i} \| \bar{h}(x_j) - \bar{h}^{\lambda}(x_j) \|$$

$$\leq \frac{\lambda}{m} \max_x \| \bar{h}^{\lambda}(x) - \bar{h}(x) \| + \frac{\lambda}{m(m-1)} (m-1) \max_x \| \bar{h}(x) - \bar{h}^{\lambda}(x) \|$$

$$\leq \frac{2\lambda}{m} \max_x \| \bar{h}(x) - \bar{h}^{\lambda}(x) \|$$

where the first inequality follows from the facts that $B_{f+g} = B_f + B_g$, that $\ell$ is convex and Lemma 4. \hfill \square

Now note that for our definition of $\psi$, $B_\psi(w||w') = \| w - w' \|^2$ is symmetric. Thus, the lemma implies

$$2c\| \bar{w}^{\lambda} - \bar{w} \|^2 \leq \frac{2\lambda}{m} \| \bar{h}^{\lambda}(x_i) - \bar{h}(x_i) \| \leq \frac{2\lambda}{m} X \| \bar{w}^{\lambda} - \bar{w} \|$$
Hence, we have

\[ \| \bar{w}^{\text{i}} - \bar{w} \| \leq \frac{\lambda X}{mc} \]

By (1), we have

\[ \beta \leq \max_x \lambda \left| \hat{h}(x) - \bar{h}^{\text{i}}(x) \right| \leq \lambda X \| \bar{w} - \bar{w}^{\text{i}} \| \leq \lambda X \cdot \frac{\lambda X}{mc} = \frac{\lambda^2 X^2}{mc} = O \left( \frac{1}{m} \right) \] (2)

This shows that the regularized ERM algorithm has uniform convergence for this setting. Choosing an appropriate \( c \) so as to make sure that the algorithm still has good empirical loss yields a good bound on its risk.