Uniform Convergence:

Let Z = input space with underlying distribution $\mathcal{D}$.

F = class of functions $Z \to \{0,1\}$ with $VC(F) < \infty$

Theorem: Suppose m samples $S = \{z_1, \ldots, z_m\}$ are drawn from $\mathcal{D}$. Then for any $0 < \delta < 1$, w.p. $\geq 1-\delta$,

$$\sup_{f \in F} \left| \frac{1}{m} \sum_i f(z_i) - \mathbb{E} f \right| \leq \sqrt{\frac{8 \ln 2S(F,m)}{m}} + \sqrt{\frac{1}{2m} \ln(1/\delta)}$$

where $S(F,m)$ = #ways m points can be labelled by $F \leq \left(\frac{me}{d}\right)^d$

[Proof will be done soon].

Application 1. Generalization bounds for learning.

$Z = X \times Y$

For each $h \in H$, define $f_h : Z \to \{0,1\}$ by $f_h(x,y) = 1(h(x) \neq y)$

Then,

$$err_D(h) = \mathbb{E} f_h$$

$$err(h,S) = \frac{1}{m} \sum_i f_h(z_i)$$

If $d = VC(H)$, w.p. $\geq 1-\delta$,

$$\sup_{h \in H} \left| err(h) - err(h,S) \right| \leq \sqrt{\frac{8d \ln em}{m}} + \sqrt{\frac{1}{2m} \ln(1/\delta)}$$

[Implication for ERM: add]

Application 2. density estimates in $\mathbb{R}^d$

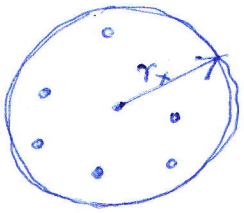
* Given m samples from a distribution, over $\mathbb{R}^d$, construct an estimate of its density function.

To estimate density at point $x \in \mathbb{R}^d$,

- r_x = smallest radius such that $B(x, r_x)$ has k samples
- estimate $= \frac{k/m}{\left(\text{volume of a ball of radius } r_x\right)} = \frac{k/m}{v_d \cdot r_x^d}$

where v_d = volume of unit ball in $\mathbb{R}^d$.

eg. $v_2 = \pi$, $v_3 = \frac{4}{3}\pi$, etc.



1. How large should k be?
2. How accurate are these estimates?

$\mathcal{F} = \{ \text{balls in } \mathbb{R}^d \} = \{ f_{\mu,r} : \mu \in \mathbb{R}^d, r > 0 \}$

$f_{\mu,r}(x) = 1(\|x-\mu\| \le r)$

$VC(\mathcal{F}) = d+1.$

By uniform convergence, w.p. ≥ 0.9, for all balls in $\mathbb{R}^d$ simultaneously,

$$\left| \frac{\#\text{ samples in ball}}{m} - \text{the probability of the ball} \right| \le O\left(\sqrt{\frac{d \log m}{m}}\right)$$

For any $x \in \mathbb{R}^d$

true probability of $B(x, r_x) = \frac{k}{m} \pm c\sqrt{\frac{d \log m}{m}}$

Reasonable choice: $k = 2c\sqrt{dm \log m}$

$\therefore$ Reasonable density estimates in $\mathbb{R}^d$ using $O(d \log d)$ samples!

[With $O(d \log d)$ points, every ball in space has roughly the right number of points]

Uniform Convergence Proof.

Theorem: Let $\mathcal{D}$ be a distribution over $\mathcal{Z}$ and $\mathcal{F}$ a class of functions $\mathcal{Z} \to \{0,1\}$. If $Z_1, \ldots, Z_m$ in D i.i.d then w.p. $\geq 1-\delta$,

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_i f(Z_i) - \mathbb{E} f \right| \leq \sqrt{\frac{8 \ln 2 S(\mathcal{F}, m)}{m}} + \sqrt{\frac{\ln(1/\delta)}{m}}$$

Intuition: Let $Z = (Z_1, \ldots, Z_m)$. $\mathcal{F}|_Z$ is small, of size $S(\mathcal{F}, m)$

Need to exploit this to bound:

$$\Pr\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_i f(Z_i) - \underbrace{\mathbb{E}[f]}_{\text{infinitely many possibilities}} \right| \geq \varepsilon \right)$$

To get rid of $\mathbb{E}f$,
we use a resampling trick: draw another m i.i.d samples $Z' = (Z_1', \ldots, Z_m')$ from the same distribution, and show that

$$\Pr\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_i f(Z_i) - \mathbb{E} f \right| \geq \varepsilon \right) \lesssim \Pr\left(\sup_{f \in \mathcal{F}} \left| \frac{1}{m} \sum_i f(Z_i) - \frac{1}{m} \sum f(Z_i') \right| \geq \varepsilon \right)$$

[**VC Intuition:** Another way to view the second expression:

1. Draw $2m$ samples T from D
2. Randomly partition into two groups Z, Z' with m samples each

After (1), $\mathcal{F}$ is down to $\mathcal{F}|_T$, which has size $\leq S(\mathcal{F}, 2m)$
For any of these $f \in \mathcal{F}|_T$, the chance that a random partition of T results in two very different averages is tiny $\leq e^{-\varepsilon^2 m}$

Overall failure probability $\lesssim S(\mathcal{F}, 2m) e^{-\varepsilon^2 m}$.

Our proof is based on this intuition, with some more sophesticated tweaks.

Step 1: $\sup_{f \in F} \left| \frac{1}{m} \sum_i f(Z_i) - \mathbb{E} f \right|$ is tightly concentrated around its expectation, viz. $\mathbb{E}\left[\sup_{f \in F} \left| \frac{1}{m} \sum_i f(Z_i) - \mathbb{E} f \right| \right]$.

Lemma: Let $g(Z) = g(Z_1, \ldots, Z_m) = \sup_{f \in F} \left| \frac{1}{m} \sum_i f(Z_i) - \mathbb{E} f \right|$. Then:

$$\Pr_Z \left(g(Z) \geq \mathbb{E}_Z g + \epsilon \right) \leq e^{-2\epsilon^2 m}$$

Proof: Generalization of Hoeffding's inequality. For all i,

$$\left| g(Z_1, \ldots, Z_i', \ldots, Z_m) - g(Z_1, \ldots, Z_i'', \ldots, Z_m) \right| \leq \frac{1}{m} \left| f(Z_i) - f(Z_i') \right|$$

* $f(Z_i)$ are i.i.d and 0/1 random variables.
* g is $\frac{1}{m}$-Lipschitz wrt the L_1-metric.

The lemma thus follows.

From now on, we will work with

$$\mathbb{E}[g] = \mathbb{E}\left[\sup_{f \in F} \left| \frac{1}{m} \sum_i f(Z_i) - \mathbb{E} f \right| \right]$$

Step 2: Symmetrization

Which is bigger?

$$\mathbb{E}|g| \quad \text{or} \quad |\mathbb{E} g|$$

$$\mathbb{E} \sup_{g \in G} g \quad \text{or} \quad \sup_{g \in G} \mathbb{E} g$$

$$\mathbb{E}_Z \sup_{f\in F} \left| \frac{1}{m}\sum f(z_i) - \mathbb{E}f \right|$$

$$= \mathbb{E}_Z \sup_{f\in F} \left| \mathbb{E}_{Z'} \left[\frac{1}{m}\sum_i f(z_i) - \frac{1}{m}\sum_i f(z_i') \right] \right|$$

$$= \frac{1}{m} \mathbb{E}_Z \sup_{f\in F} \left| \mathbb{E}_{Z'} \sum_i \left(f(z_i) - f(z_i') \right) \right|$$

$$\le \frac{1}{m} \mathbb{E}_Z \mathbb{E}_{Z'} \sup_{f\in F} \left| \sum_i f(z_i) - f(z_i') \right|$$

Recall that $f(z_i) - f(z_i') \overset{d}{=} f(z_i') - f(z_i)$ (symmetric)

We introduce extra r.v. $\sigma_1, \dots, \sigma_m \in \{+1, -1\}$; where $\sigma_i = 1$ w.p. $1/2$ iid and:

$$= \frac{1}{m} \mathbb{E}_Z \mathbb{E}_{Z'} \mathbb{E}_\sigma \sup_{f\in F} \left| \sum_i \sigma_i \left(f(z_i) - f(z_i') \right) \right|$$

$$\le \frac{1}{m} \mathbb{E}_{Z,Z'} \mathbb{E}_\sigma \sup_{f\in F} \left[\left| \sum_i \sigma_i f(z_i) \right| + \left| \sum_i \sigma_i f(z_i') \right| \right]$$

$$\le \frac{2}{m} \mathbb{E}_{Z} \mathbb{E}_\sigma \sup_{f\in F} \left| \sum_i \sigma_i f(z_i) \right|$$

Step 3: Conditioning.

Note for any Z, $F|_Z$ has size $\le S(F,m)$, and we will show:

$$\mathbb{E}_\sigma \sup_{f\in F} \left| \sum_i \sigma_i f(z_i) \right| = \mathbb{E}_\sigma \max_{a\in F|_Z} \left| \sum_i \sigma_i a_i \right| \le \sqrt{2m \ln 2S(F,m)}$$

This completes the theorem:

$$\sup_{f\in F} \left| \frac{1}{m}\sum_i f(z_i) - \mathbb{E}f \right| \le \mathbb{E} \sup_{f\in F} \left| \frac{1}{m}\sum f(z_i) - \mathbb{E}f \right| + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

(w.p. $\ge 1-\delta$)

$$\le \frac{2}{m}\sqrt{2m \ln 2S(F,m)} + \sqrt{\frac{\ln(1/\delta)}{2m}}$$

Last detail:

Lemma: (Massart) For finite $A \subset \mathbb{R}^m$, with $R = \max_i \|a_i\|$,

$$\mathbb{E}_\sigma \max_{a \in A} \sum_i \sigma_i a_i \leq R \cdot \sqrt{2 \ln |A|}$$

[redacted]

2. Using Hoeffding's inequality, we could show a slightly weaker bound:

$$P_\sigma \left(\sum_i \sigma_i a_i \geq r \right) \leq e^{-2r^2/\|a\|^2}$$

We'll use a slightly different approach.

$$e^{t\, \mathbb{E}_\sigma [\max_{a \in A} \sum_i \sigma_i a_i]} \leq \mathbb{E}_\sigma e^{t \max_a \sum_i \sigma_i a_i} \qquad \text{(Jensen)}$$

$$= \mathbb{E}_\sigma \max_a e^{t \sum_i \sigma_i a_i}$$

$$\leq \sum_a \mathbb{E}_\sigma e^{t \sum_i a_i \sigma_i}$$

$$= \sum_a \mathbb{E}_\sigma e^{t\sigma_1 a_1} \cdot \mathbb{E}_\sigma e^{t\sigma_2 a_2} \cdot \mathbb{E}_\sigma e^{t a_3 \sigma_3} \cdots$$

$$\leq \sum_i e^{t^2 a_1^2/2} \cdot e^{t^2 a_2^2/2} \cdots \qquad \text{(as in Hoeffding proof)}$$

$$\leq |A| e^{t^2 R^2/2}$$

$$\therefore \mathbb{E}_\sigma [\max_a \sum_i \sigma_i a_i] \leq \frac{\ln |A|}{t} + \frac{t R^2}{2}$$

Pick $t = \sqrt{\frac{2 \ln |A|}{R^2}}$

Uniform Convergence: Realizable Case.

- Let $S(H, m)$ = max #ways H can label a set of m points from X.

If $d = VC(H)$, then $S(H, m) \leq \left(\frac{me}{d}\right)^d$

We can shown that:

Theorem: In the realizable case, suppose we draw m examples iid from D. For any $\varepsilon > 0$,

$$\Pr\left[\exists h \in H \text{ with } err(h) > \varepsilon \text{ s.t. } h \text{ is consistent with the examples}\right] \leq 2S(H, 2m)\, e^{-\varepsilon m}$$

[Proof follows differently - by using Bernstein's inequality]

Restating this: If the learner outputs some $h^* \in H$ consistent with the examples, then: w.p. $\geq 1-\delta$,

$$err_D(h^*) \leq \frac{2}{m} \ln \frac{2S(H, 2m)}{\delta}$$

$$\leq \frac{2}{m}\left[\ln \frac{2}{\delta} + d \ln \frac{em}{d}\right] \quad \text{where } d = VC(H).$$

Lower Bound: If $VC(H) = d$, then any PAC algorithm will draw $\Omega(d/\varepsilon)$ examples in the worst case.

Proof: Suppose h shatters a set $x_1, \ldots, x_d$. Pick a labelling $y = (y_1, \ldots, y_d)$ u.a.r. Realizable.

Consider distribution D:

mass $1 - 4\varepsilon$ on (x_1, y_1)

mass $4\varepsilon/(d-1)$ on (x_i, y_i) $\quad i > 1$

Suppose a learner draws a sample of size $\leq \frac{d-1}{32\varepsilon}$ from $\mathcal{D}$, and H is consistent with these samples.

If S includes N of the pts $\{x_1, \ldots, x_d\}$. Then:

$$\mathbb{E}[N] \leq 4\varepsilon \cdot \frac{d-1}{32\varepsilon} \leq \frac{d-1}{8}$$

$$\Pr\left(N \geq \frac{d-1}{2}\right) \leq \frac{1}{4} \quad \text{(by Markov)}$$

w.p. $\geq 3/4$, the learner sees half the samples and has to guess on the other half, randomly.

$$P_y\left(\text{err}_D(h) \leq \varepsilon \,\middle|\, N < \frac{d-1}{2}\right)$$

$$\leq P_y\left(h \text{ coincides with } y \text{ on more than half of the unseen points} \,\middle|\, N < \frac{d-1}{2}\right)$$

$$\leq 1/2$$

w.p. $\geq 1/4$, over choice of y, S, $\text{err}_D(h) > \varepsilon$

$\therefore > \frac{d-1}{32\varepsilon}$ samples are needed.

$\exists$ a bad distribution $\mathcal{D}$ for which $\Omega(d/\varepsilon)$ examples are needed. [May not be needed for <u>all</u> distributions.]