Independence & overfitting.

From last time: Hoeffding's inequality says: if $Z_i \in [0, \delta)$, then

$$\frac{1}{n} \sum Z_i \leq \frac{1}{n} \sum Z_i + (\delta - \alpha) \sqrt{\frac{\ln(\delta/\alpha)}{2n}}$$

Thus, ask for fixed $\delta$.

$$\Pr[\text{sgn}(f(x)) \neq Y] \leq \frac{1}{n} \sum \text{sgn}(f(x)) + \sqrt{\frac{\ln(\delta/\alpha)}{2n}}$$

Alternatively, suppose \(y \leq 1, 1\)

$$|f'(x)| \leq R \quad \text{and} \quad \ell \text{ is } \text{Lipschitz}.$$ Then

$$|\ell(-f'(x)) - \ell(0)| \leq R \cdot |f'(x)| \leq R \cdot R$$

$$\Rightarrow \ell(-f'(x)) \leq \ell(0) + R \cdot R,$$

$$\ell(\text{sgn}(f(x))) \geq R(0) - R \cdot R$$

$$\Rightarrow Z_+ := \ell(-f'(x)) \in [\ell(0) - R \cdot R, \ell(0) + R \cdot R],$$

w/ $\Pr \geq 1 - \delta$.

$$R_E(f) \leq \hat{R}_E(f) + 2 \cdot R \cdot \sqrt{\frac{\ln(1/\delta)}{2n}}.$$

Note: If we want both bounds to hold, we must union bound and $h(\mathcal{H})$ becomes $h(\mathcal{H}/\delta)$.
As mentioned before, this bound is for just one function, but we need it for now. Why does this issue exist?

First view of issue: independence.

We proved the preceding bounds by constructing
\[ Z_i := l(t(x_i), y_i), \]
and applying Hoeffding for some other Chernoff bound. Hoeffding requires \((Z_1, \ldots, Z_n)\) independent (when using MBA), but if we use \((x_i, y_i)_{i=1}^n\) to select \(t\), then all elements of \((Z_1, \ldots, Z_n)\) could be correlated.

Second view: a bad example.

Here’s a pathological example. Suppose \(X \in \{0, 1\} \text{ uniformly, and not only } Y \in \{-1, +1\}, \text{ but in fact } Y = X\).

Consider a rule learner \(t\) that memories \((x_i, y_i)_{i=1}^n\) and outputs

\[ \hat{f}(x) := \sum_{i=1}^n y_i \quad \text{if } x = x_i \text{ for some } i, \]
\[ \hat{f}(x) := 0 \quad \text{otherwise.} \]

Then \(\Pr[\hat{f}(x) \neq y] = 1\) but \(\frac{1}{n} \sum \mathbb{1}\{\hat{f}(x) \neq y\} = 0\).
This leads to the concept of **overfitting** meaning the training error is much lower than test error. Since we use training data to select predictors, we will have a non-independence issue but there is also a question of the degree to which train and test errors differ.

Remarks

* Can't we just use two stages?

1. Train $\mathbf{f}$ on $(\mathbf{X}, \mathbf{y})_{in}$, then output estimate $\hat{y}(x) = \mathbb{E}(y | \mathbf{f}(x))$.

Sure but we've sampled the data we train with.

- Suppose the estimate is bad; it's ok, rule to check other functions; indeed, this is "validating".
- But then we must union bound over things that are both where we started...

* Online learning, smoothing, coherence.

With online learning, FTRL was able to get "coherence" of past and future by forcing predictors to change little over time.

In present setting, this controls overfitting by not being too sensitive to individual examples.
Remarks

* SGD example and martingales.

Similarly, for SGD with $\text{Adam}$, we proved:

$$\frac{1}{n} \sum w_i^2 \leq \frac{D G}{n} \left( 1 + \sqrt{8 \ln \left( \frac{n}{\delta} \right)} \right).$$

(for instance, can take $L$ as a Lipschitz loss, $f = \sum x_i (\omega, x_i, y_i)$.)

The choice $\frac{1}{n} \sum w_i$ is a random variable that is correlated with $\sum_{i=1}^{n} L(x_i, y_i) w_i$.

To avoid independence issues, a few things were used:

* Algorithm uses each data point once and minimally changes its predictor for each (isn't trying to "memorise").

* Analysis is algorithm-dependent, moreover uses a Martingale bound together with precisely an "use-once" property.

It even one example is re-used, we break martingale!

So that proof had this sensitivity too.
Finite classes & primitive covers

The easiest way to handle deviations is to control the random variable

\[ \sup_{f \in \mathcal{F}} |R(f) - \hat{R}(f)|. \]

This is called uniform deviations, uniform law of large numbers, etc.

Since it is a bound that holds uniformly over a class of functions \( \mathcal{F} \).

+ This sidesteps the independence issue: we apply it for every function the algorithm could consider.

In some sense we apply bound to \( \mathcal{F} \) "before seeing data".

- Very easy to be loose & non-obtained to algorithm behavior. Overcoming this leads to lots of Soph's traction.

Even so, this will be the approach we study for rest of course.
There are a number of algorithm-dependent approaches. One is "Stability" by Bousquet & Elisseeff.

Uniform deviation bounds will add a term to our familiar "Hoeffding-style" $O\left(\frac{\ln(n/k)}{n}\right)$ that measures complexity of $f$.

We will only need a one-sided bound on $\sup_{f \in F} R(f) - \hat{R}(f)$, which gives the other direction by applying it to $-R - \hat{R}$ (in most cases...).

Some literature has uniform deviations $\sup_{f \in F} |R(f) - \hat{R}(f)|$.

(Measure theory note.) Measure theory only guarantees existence of random variables created via countable operations (roughly speaking).

Meanwhile, $\sup_{f \in F} R(f) - \hat{R}(f)$ can be problematic for uncountable $F$ (and there exist bad counterexamples, see classical paper "Learnability & the Vapnik-Chervonenkis Dimension" Blumer, Ehrenfeucht, Haussler, Warmuth, specifically appendix A mention of Shai Ben-David).

A "friendly" fix is to work with classes that have a countable dense subset.
Let's start with finite classes.

**Theorem.** Suppose \( l(f(x), y) \in [a, b] \) almost surely \( \forall f \in \mathcal{F} \).

With probability at least \( 1 - \delta \),

\[
\sup_{f \in \mathcal{F}} R(f) - \hat{R}(f) \leq (b-a) \sqrt{\frac{\ln (1/\delta) + \ln (1/\delta')}{2n}}.
\]

**Proof.** If \( |\gamma| = 0 \), bound immediate, thus suppose \( |\gamma| \geq 0 \).

Define \( \delta' = \delta / |\gamma| \). By Hoeffding, for any fixed \( f \in \mathcal{F} \)

\[
\Pr \left[ R(f) - \hat{R}(f) \geq (b-a) \sqrt{\frac{\ln (1/\delta')}{2n}} \right] \leq \exp \left( -\frac{2n \ln (1/\delta')}{(b-a)^2 \frac{\ln (1/\delta')}{2n}} \right) = \exp \left( \ln (\delta') \right) = \delta'.
\]

Thus

\[
\Pr \left[ \exists f \in \mathcal{F} \quad R(f) - \hat{R}(f) \geq \varepsilon \right] \leq \sum_{f \in \mathcal{F}} \Pr \left[ R(f) - \hat{R}(f) \geq \varepsilon \right] \leq 1/|\gamma| \cdot \delta' = \delta.
\]

(Statement of theorem is negation of this.)

**Remark.** When is this union bound tight?

Union bound is equality when events are disjoint.

Thus we should avoid controlling events which are too similar.
For many classes, the union bound approach is applied by first applying a discretization.

**Definition.** Given set $U$, scale $\varepsilon$, norm $\|\cdot\|$, a subset $V \subseteq U$ is a (proper) cover when

$$\sup_{u \in U} \inf_{v \in V} \|u - v\| \leq \varepsilon.$$ 

The (proper) covering number $\mathcal{N}(U, \varepsilon, \|\cdot\|)$ is the smallest cardinality of a cover.

We will use covers $\mathcal{C}$ as follows. Given a loss $l$, features $f$, and inputs $\bar{x}$, let $(\mathcal{C} \times \bar{x})$ denote all realizations of $\bar{x}$ across $\mathcal{C}$, e.g.

$$\langle \mathcal{C} \times \bar{x} \rangle_{\text{feas}} = \{ (l(f(x), y))_{(x, y) \in \bar{x}} : f \in \mathcal{C} \},$$

$$\langle \mathcal{C} \times \bar{x} \rangle_{\text{feas}} = \{ l(-f(x, y))_{(x, y) \in \bar{x}} : (x, y) \in \bar{x} \}.$$
Theorem (trivial/ uniform covering) [This is not the end of covering!]

Let input set $x$ be given, along with probability distribution on it. Suppose $\log \epsilon \leq \log \beta$ a.s.,

Then with $\rho \geq 1.5$,

$$\sup_{f \in \mathcal{G}} R(f) - \hat{R}(f) \leq \inf_{\epsilon > 0} \left( 2\epsilon + (b-c) \right) \sqrt{\frac{\ln (\mathcal{N}(\mathcal{H}(\epsilon), \epsilon, l. l. l.))}{2n}}$$

Proof. Fix $\epsilon > 0$. Suppose $\mathcal{N}(\mathcal{H}(\epsilon), \epsilon, l. l. l.) < \infty$ (else trivial),

let $\mathcal{G}$ be any minimum cardinality cover.

By finite class bound & boundedness of $\mathcal{H}$,

$$\sup_{f \in \mathcal{G}} R(f) - \hat{R}(f) \leq (b-c) \sqrt{\frac{\ln (\mathcal{N}(\mathcal{H}(\epsilon), \epsilon, l. l. l.))}{2n}}$$

For any $f \in \mathcal{G}$, let $\hat{f}$ denote closest element in $\mathcal{G}$; by choice of $\mathcal{H}$, qed definition cover,

$$\sup_{(x, y) \in \mathcal{Z}} |l(f(x), y) - l(f^*(x), y)| \leq \epsilon.$$

Thus

$$\sup_{f \in \mathcal{G}} R(f) - \hat{R}(f) = \sup_{f \in \mathcal{G}} \left( E[l(f(x), y) - l(f^*(x), y)] + R(f) - \hat{R}(f) \right)$$

$$\leq \sup_{f \in \mathcal{G}} \left( 2\epsilon + R(f) - \hat{R}(f) \right) = \sup_{f \in \mathcal{G}} \left( 2\epsilon + R(f) - \hat{R}(f) \right).$$
Remarks.

* Quick example (will be worked out later in detail).

Suppose \( l((x,y),y) = l(-<w,xy>) \) with \(|w| \leq l, \|w\| \leq l \) for all \( y \).

\( l \) is \( 1 \)-Lipschitz. Then have a cover of size \( O(\frac{1}{\varepsilon}) \) by covering unit ball in \( \mathbb{R}^n \):

\[
|l(<w,-xy>) - l(<w',-xy>)| \leq |<w,-xy> - <w',-xy>|
\leq \|xy\| \cdot \|w-w'\| \leq \varepsilon.
\]

So bound is:

\[
\sup_{\|w\|\leq l} Re(l(w) - R_e(w)) \leq \inf_{\varepsilon \geq 0} 2\varepsilon + 2\sqrt{\frac{O(d(h;\varepsilon)) + h(\varepsilon)}{2n}} = O\left(\sqrt{\frac{d \ln(n) + h(\varepsilon)}{2n}}\right).
\]

Choose \( \varepsilon = \frac{1}{2n} \).

We can still remove the "\( d \)"!

* Covering was introduced by Kolmogorov & Tikhomirov, but originally it's the oldest idea in analysis (e.g., let \( h \) compact).

* Covers can be "improper" by not requiring them to be subsets of target. But then need extra steps to deal with how they are bounded.

* Above example gives one benefit of \( l(h) \): poly \( (\frac{1}{\varepsilon}) \) scary.
This approach has a serious weakness (the extra "d" factor with SGD is not the worst!). The covering number is still infinite in many trivial cases, for instance univocally thresholds

\[ x \mapsto \| L x \|_\infty \] where \( x_0, y_0 \leq 0, 0.5 \).

This is due to uniform norm on infinite set.