12.1 Recap

A class of functions $\mathcal{F}$ is called a Glivenko-Cantelli (GC) class if
$$\lim_{n \to \infty} P \left( \sup_{f \in \{ \mathbb{E}_n f(X) - \mathbb{E} f(X) \mid \epsilon \} > \epsilon \right) = 0$$
Here the notation $\hat{\mathbb{E}}_n = \frac{1}{n} \sum_{i=1}^{n} f(X^{(i)})$ denotes the empirical expectation, and the RVs $X^{(i)}, i = 1, \ldots, n$ are i.i.d. samples $\sim \mathbb{P}$. We say that a uniform law of large numbers (ULN) holds for $\mathcal{F}$.

12.1.1 Example where ULN does not hold

Let $\mathcal{F} = \mathcal{F}_+ \cup \mathcal{F}_-$ where
$$\mathcal{F}_+ = \{ f : [0, 1] \to \{+1, -1\} : |\{ x : f(x) = +1 \}| < \infty \}, \quad \text{and}$$
$$\mathcal{F}_- = \{ f : [0, 1] \to \{+1, -1\} : |\{ x : f(x) = -1 \}| < \infty \}.$$ 

Suppose that $X \sim \text{unif}[0, 1]$ and $\mathbb{P}(Y = +1 | X = x) = .95 \forall x \in [0, 1]$. Then the Bayes classifier always predicts $+1$ (so it’s in $\mathcal{F}_-$) and $R^* = .95$. Consider a finite sample $\{(X^{(i)}, Y^{(i)})\}_{i=1}^{n}$. For a classifier $f$, $\hat{R}_n(f) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X^{(i)} \neq f(X^{(i)}))$ is the empirical risk. But for any finite sample size $n$, the empirical risk $\inf_{f \in \mathcal{F}} \hat{R}_n(f) = 0$; moreover, we can achieve this minimum value of zero risk by choosing some $\hat{f}_n \in \mathcal{F}_+$. But for any $f \in \mathcal{F}_+$, the population risk is $R(f_n) = 0.95$, since $f_n$ can be $+1$ for at most a finite number of points (and this has zero probability for the uniform distribution over $X$). So we have generated a sequence of functions $f_n$ such that $R_n(f_n) = 0$ but $R(f_n) = 0.95$ for all $n = 1, 2, \ldots$, showing that minimum empirical risk does not converge to population risk over this class.

12.2 The GC theorem

**Theorem 12.1.** (Classical Glivenko-Cantelli) Let $Z^{(i)}$ $i = 1, \ldots, n$ be i.i.d. with CDF $F(t) = \mathbb{P}(Z \leq t)$. Define empirical CDF $\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(Z^{(i)} \leq t)$. Then
$$\lim_{n \to \infty} P \left( \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| > \epsilon \right) \to 0$$
Or, in other words, $||\hat{F}_n - F||_\infty$ goes to zero in probability.

### 12.2.1 Example

$F = \{I(\cdot \leq t) : t \in \mathbb{R}\}$. $\hat{F}_n[\mid Z \leq t] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(Z^{(i)} \leq t) = \hat{F}_n(t)$. So the theorem says that the class $F$ of half-intervals is a Glivenko-Cantelli class.

### 12.2.2 The proof

**Proof:** (Proof follows Vapnik and Chervonenkis, 1971; see also Pollard, 1984).

(A) Symmetrization. Let $\{\tilde{Z}^{(i)}\}_{i=1}^{n}$ be a new i.i.d. sample independent of $\{Z^{(i)}\}_{i=1}^{n}$. Then for $n \epsilon^2 \geq 2$,

$$
P \left( \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| > \epsilon \right) \leq 2P \left( \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - \tilde{F}_n(t)| > \epsilon \right)
$$

where $\tilde{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\tilde{Z}^{(i)} \leq t)$. The proof of this claim is left as an exercise (use Chebyshev inequality).

(B) Introduce random signs. Let the RVs $\sigma^{(i)}$ $i = 1, ..., n$ be i.i.d. with distribution

$$
P(\sigma^{(i)} = +1) = P(\sigma^{(i)} = -1) = \frac{1}{2}.
$$

Then we have

$$
P \left( \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} (\mathbb{I}(Z^{(i)} \leq t) - \mathbb{I}(\tilde{Z}^{(i)} \leq t)) \right| > \frac{\epsilon}{2} \right)
$$

$$
= P \left( \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} (\mathbb{I}(Z^{(i)} \leq t) - \mathbb{I}(\tilde{Z}^{(i)} \leq t)) \right| > \frac{\epsilon}{2} \right)
$$

$$
\leq 2P \left( \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} \mathbb{I}(Z^{(i)} \leq t) \right| > \frac{\epsilon}{4} \right)
$$

where the equality is due to symmetry since $Z^{(i)}, \tilde{Z}^{(i)}$ are altogether i.i.d. and the inequality is by union bound.

(C) Conditioning. Define

$$
C = P \left( \sup_{t \in \mathbb{R}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} \mathbb{I}(Z^{(i)} \leq t) \right| > \frac{\epsilon}{4} |Z^{(1)}, ..., Z^{(n)}\right)
$$
WLOG, reorder data s.t. \( Z^{(1)} \leq \cdots \leq Z^{(n)} \). Then it is clear that changing \( t \) can only give \( n + 1 \) different values for \((\mathbb{I}(Z^{(1)} \leq t), \ldots, \mathbb{I}(Z^{(n)} \leq t))\). Hence, by union bound,

\[
C \leq (n + 1) \sup_{t \in \mathbb{R}} \mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} \mathbb{I}(Z^{(i)} \leq t) - \epsilon \right| > \frac{\epsilon}{4} \left| Z^{(1)}, \ldots, Z^{(n)} \right| \right)
\]

(D) Tail bounds. For any fixed \( t \in \mathbb{R} \), fixed \( Z^{(1)}, \ldots, Z^{(n)} \), since \( \sigma^{(i)} \mathbb{I}(Z^{(i)} \leq t) \) are bounded random variables,

\[
\mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^{n} \sigma^{(i)} \mathbb{I}(Z^{(i)} \leq t) - \epsilon \right| > \frac{\epsilon}{4} \left| Z^{(1)}, \ldots, Z^{(n)} \right| \right) \leq 2 \exp(-n\epsilon^2/32)
\]

by the Hoeffding bound. Note that this bound holds uniformly, for any choice of \( t \) and any choice of the data \( Z^{(i)}, i = 1, \ldots, n \). Hence, we can take the supremum over \( t \in \mathbb{R} \) and take expectations over the data \( Z^i \) to conclude that

\[
\mathbb{P}(\|\hat{F}_n - F\|_\infty > \epsilon) \leq 8(n + 1) \exp(-n\epsilon^2/32).
\]

Note that this converges to zero as \( n \to +\infty \) for any fixed \( \epsilon > 0 \), which completes the proof. \( \square \)

(For further discussion, see Convergence of Stochastic Processes, Pollard, 1984, chapter 2; posted on course webpage).

### 12.3 Concentration Inequalities

To complete the proof, it remains for us to prove the Hoeffding bound, which provides a sub-Gaussian tail bound for sums of bounded RVs.

#### 12.3.1 Recap

Let us recall some elementary tail bounds from probability:

**Theorem 12.2.** (Markov) For nonnegative \( Z \) with finite mean, \( \mathbb{P}(Z > t) \leq \mathbb{E}Z/t \).

**Theorem 12.3.** (Chebyshev) For \( X \) with finite variance, \( \mathbb{P}(|X - \mathbb{E}X| > \epsilon) \leq \text{Var}X/\epsilon^2 \).

**Theorem 12.4.** (Chernoff) For \( X \) with MGF defined at a positive \( \lambda \), \( \mathbb{P}(X > t) \leq \inf_{\lambda > 0} \mathbb{E}e^{\lambda X}/e^{\lambda t} \).
12.3.2 Hoeffding’s inequality

**Theorem 12.5.** (Hoeffding 1963) Let $X^{(1)}, \ldots, X^{(n)}$ be independent with $X^{(i)} \in [a_i, b_i]$. Then $\forall \epsilon > 0$,

$$
P\left(\left|\frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \mathbb{E}X^{(i)})\right| > \epsilon\right) \leq 2 \exp\left(-\frac{2n\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)
$$

**Proof:** By Chernoff,

$$
P\left(\frac{1}{n}\sum_{i=1}^{n}(X^{(i)} - \mathbb{E}X^{(i)}) > \epsilon\right) \leq e^{-\lambda \epsilon n} \mathbb{E}\exp\left(\lambda\sum_{i=1}^{n}(X^{(i)} - \mathbb{E}X^{(i)})\right) = e^{-\lambda \epsilon n} \prod_{i=1}^{n} \mathbb{E}e^{\lambda(X^{(i)} - \mathbb{E}X^{(i)})}
$$

We claim that $\mathbb{E}e^{\lambda(X^{(i)} - \mathbb{E}X^{(i)})} \leq \exp(\lambda^2(b_i - a_i)^2/8)$ (i.e., sub-Gaussian). Having this we’d be done after minimizing the Hoeffding bound

$$
\log \leq \exp\left(n \left(\inf_{\lambda > 0} \left(-\lambda \epsilon + \frac{\lambda^2}{8n} \sum_{i=1}^{n}(b_i - a_i)^2\right)\right)\right)
$$

So, we will prove our claim. Note that $e^{\lambda x}$ is convex in $x$. So, $\forall x \in [a, b]$,

$$
e^{\lambda x} = e^{\frac{x-a}{b-a}b + \frac{b-x}{b-a}a} \leq e^{\frac{x-a}{b-a}b} + e^{\frac{b-x}{b-a}a}
$$

The proof will be finished next lecture...  □