# Lecture 4

## What can we learn privately?

## Recall:

### Probably Approximately Correct (PAC)

$X$ instance space — set of all instances
(input space / domain)

$h : X \to \{+1, -1\}$ concept — a function to label elements

$\mathcal{H}$ concept class — a collection of labeling functions

$h^*$ target concept — $h^* \in \mathcal{H}$ and label all instances correctly

$D$ target distribution — distribution over instances

sample / training data set $\begin{cases} (x_1, h^*(x_1)) \\ (x_2, h^*(x_2)) \\ \vdots \\ (x_m, h^*(x_m)) \end{cases}$

Agnostic case $h^*$ DOES NOT exist

training set $= \begin{cases} (x_1, y_1) \\ \vdots \\ (x_m, y_m) \end{cases}$

# Recall

true error : $\mathrm{err}(h) := \Pr_{(x,y)\sim D}\left[ h(x) \neq y \right]$

Data set

training set

$$= \overset{Z}{\boxed{\begin{array}{c} (x_1, y_1) \\ (x_2, y_2) \\ \vdots \\ (x_m, y_m) \end{array}}}$$

samples from D

$$\hat{\mathrm{err}}(h) = \frac{\left|\{ i \mid i \in [m] : h(x_i) \neq y_i \}\right|}{m}$$

ERM : empirical risk minimizer

$$\text{output} := \arg\min_{h} \hat{\mathrm{err}}(h)$$

↳ training error

Recall

## PAC learning (Probably Approximately Correct)

Suppose that we have a concept class $\mathcal{H}$ over $X$. We say that $\mathcal{H}$ is PAC learnable if there exists an algorithm $A$ s.t:

$\forall\ h \in \mathcal{H}$, $\forall\ D$ over $X$, $\forall\ \alpha, \beta \in (0, 0.5]$

$A$ receives $\alpha, \beta$, and samples $\langle x_1, y_1 \rangle, \ldots, \langle x_m, y_m \rangle$ where $x_i$'s are iid samples from $D$.

proper $[\in \mathcal{C}]$

Then, w. p. $\geq 1-\beta$, $A$ outputs $\hat{h}$ s.t.

$$err(\hat{h}) \leq \min_{h \in \mathcal{H}} err(h) + \alpha$$

The probability is taken over the randomness in the samples and any internal coin flips of $A$.

$\boxed{\text{* ERM works for a finite class } \mathcal{H} \text{ if we have enough samples.}}$

- Problem setup:

samples $(x_1, y_1), \ldots, (x_m, y_m) \sim D$

$h \in \mathcal{H}$ : $err(h) := \Pr_{(x,y) \sim D} [h(x) \neq y]$

- Goal

find $\hat{h} \in \mathcal{H}$ s.t. with probability $1-\beta$, $err(\hat{h}) \leq \min_{h \in \mathcal{H}} err(h) + \alpha$

proof via uniform convergence.

set $\alpha' = \frac{\alpha}{3}$, $\beta' = \beta/2$.

we show if $m \geq O\left(\frac{\log(|\mathcal{H}|) + \log(1/\beta')}{\alpha'^2}\right)$

then

$\Pr[\forall h \in \mathcal{H} : |\hat{err}(h) - err(h)| < \alpha'] \geq 1-\beta'$

For all $h \in \mathcal{H}$, we define:

$X_h :=$ # samples in $z$ where $h(x_i) \neq y_i$

$$\hat{err}(h) = \frac{X_h}{m}$$

Each sample is mislabeled w.p.

$$err(h) = \Pr_{(x,y)\sim D}[h(x) \neq y].$$

$\Rightarrow X_h$ is a binomial random variable.

$$X_h \sim Bin(m, err(h))$$

Via Hoeffding bound:

$$\Pr[|\hat{err}(h) - err(h)| > \alpha'] =$$

$$\Pr\left[\left|\frac{X_h}{m} - err(h)\right| > \alpha'\right] \leq 2e^{-\frac{m\alpha'^2}{3}}$$

$$\Pr\left[\exists h \in \mathcal{H} : |\hat{err}(h) - err(h)| > \alpha'\right]$$

$$\leq \sum_{h \in \mathcal{H}} \Pr\left[|\hat{err}(h) - err(h)| > \alpha'\right]$$

$$\leq |\mathcal{H}| \cdot 2 \cdot e^{-\frac{m\alpha'^2}{3}} \leq \beta'$$

This holds for all

$$m \geq \frac{3\left(\ln(2|\mathcal{H}|) + \ln(1/\beta')\right)}{\alpha'^2}$$

$$\Rightarrow \Pr\left[\forall h \in \mathcal{H} : |\hat{err}(h) - err(h)| \leq \alpha'\right]$$

$$\geq 1 - \beta'$$

The guarantee of the uniform convergence

Recall

$\boxed{\text{Uniform convergence (UC) implies agnostic PAC learnability via EMR.}}$

Outcome of ERM minimizes the $\hat{err}(h)$:

$$* \quad \hat{h}_{ERM} := \arg\min_{h \in \mathcal{H}} \hat{err}(h)$$

$$h^* := \arg\min_{h \in \mathcal{H}} err(h)$$

$$\left.\begin{aligned}
&* \quad \Rightarrow \quad \hat{err}(\hat{h}_{ERM}) \leq \hat{err}(h^*) \\
&UC \quad \underset{1-\beta'}{\overset{w.p.}{\Longrightarrow}} \quad \begin{cases} \hat{err}(\hat{h}_{ERM}) \leq err(h_{ERM}) + \alpha' \\ \hat{err}(h^*) \leq err(h^*) + \alpha' \end{cases}
\end{aligned}\right\}$$

$$\Rightarrow \quad \hat{err}(\hat{h}_{ERM}) \leq err(h^*) + 2\alpha'$$

$$\leq \min_{h \in \mathcal{H}} err(h) + 2\alpha'$$

as desierd in PAC learning.

# Exponential mechanism:

Help us to pick an element privately

→ an element that maximize a score or utility

Algorithm:

Input: data set $z$, set of options score function $u$, privacy parameter $\varepsilon$,

Output: pick $h \in \mathcal{H}$ with prob.

$$\propto e^{\frac{\varepsilon \cdot u(h, z)}{2\Delta}}$$

$\Delta$ is sensitivity of $h$:

$$\Delta u := \max_{h, z, z'} |u(h, z) - u(h, z')|$$

neighboring

## Proof of privacy:

for all $h$ :

$$\frac{e^{\frac{\varepsilon \cdot u(h,z)}{2\Delta}}}{e^{\frac{\varepsilon \cdot u(h,z')}{2\Delta}}} = e^{\varepsilon \frac{|h(z) - h(z')|}{\Delta}} \leq e^{\varepsilon/2}$$

$$Pr[\text{output} = h \mid z] = \frac{e^{\frac{\varepsilon\, u(h,z)}{2\Delta}}}{\sum_{h'} e^{\frac{\varepsilon\, u(h,z)}{2}}}$$

$$\leq \frac{e^{\frac{\varepsilon \cdot u(h,z')}{2\Delta}} \cdot e^{\varepsilon/2}}{\sum_{h'} e^{\frac{\varepsilon\, u(h,z')}{2\Delta}} \cdot e^{-\varepsilon/2}}$$

$$= e^{\varepsilon} \cdot Pr[\text{output} = h \mid z']$$

## Score performance of output $\hat{h}$

**Thm** $\Pr\left[ u(\hat{h}) \leq OPT - \frac{2\Delta}{\varepsilon}(\ln(H) + t)\right]$

$\leq e^{-t}$

$\hookrightarrow OPT = \max_h u(h, z)$

proof:

$c := OPT - \frac{2\Delta}{\varepsilon}(\ln(H) + t)$

$$\Pr[u(\hat{h}, z) \leq c] = \sum_{\substack{h: \\ u(h,z) \leq c}} \Pr[\text{pick } h \mid z]$$

$$\leq \frac{\sum_{h : u(h,z) \leq c} e^{\frac{\varepsilon u(h,z)}{2\Delta}}}{\sum_{h \in \mathcal{H}} e^{\frac{\varepsilon u(h,z)}{2\Delta}}}$$

$$\leq \frac{|\mathcal{H}| \, e^{\frac{\varepsilon c}{2\Delta}}}{e^{\frac{\varepsilon OPT}{2\Delta}}}$$

$$= \exp\left(\ln(H) + \frac{\varepsilon}{2\Delta}(c - OPT)\right) \leq e^{-t}$$

# Private PAC learning

Exponential mechanism chooses an item with high utility. We can use this mechanism to mimic ERM. We privately select a concept that labels a large number of samples correctly.

For all $h \in \mathcal{H}$, we define $u(h, z)$ to be the number of correctly labeled samples:

$$u(h, \underset{\text{training set}}{z}) = \left| \{ i \mid h(x_i) = y_i \} \right|$$

Note that $\hat{err}(h) = 1 - \frac{u(h,z)}{m}$

ERM minimizes $\hat{err}(h)$. $\approx$ Exponential mechanism Privately maximizes $u(h,z)$

Note $\hat{h}_{ERM}$ maximizes $u(h, z)$:

$$OPT = u(\hat{h}_{ERM}, z)$$

Let $\hat{h}_{\mu}$ be the outcome of the exponential mechanism.

- we have shown w.p. $1-\beta'$:

$\Delta$ of $u$ is 1

$$\underbrace{u(\hat{h}_{\mu}, z)}_{= m - m.\hat{err}(\hat{h}_{\mu})} > \underbrace{u(\hat{h}_{ERM}, z)}_{OPT} - \frac{2\Delta}{\varepsilon}\left(\ln(|H|) + \log(\tfrac{1}{\beta'})\right)$$

$$\Rightarrow \cancel{m} - m.\hat{err}(\hat{h}_{\mu}) > \cancel{m} - m.\hat{err}(h_{\mu}) - \frac{2}{\varepsilon}\left(\ln(|H|) + \ln(1/\beta')\right)$$

$$\left(\times \frac{-1}{m}\right) \Rightarrow \hat{err}(\hat{h}_{\mu}) < \hat{err}(\hat{h}_{ERM}) + \frac{2}{m\varepsilon}\left(\ln(|H|) + \ln(1/\beta')\right)$$

Thus, for all $m \geq \dfrac{2(\ln(|H|) + \ln(1/\beta'))}{\varepsilon\alpha'}$, we get:

$$\hat{err}(\hat{h}_{\mu}) \leq \hat{err}(\hat{h}_{ERM}) + \alpha' \quad (1)$$

- if $m \geq \theta\left(\frac{\log(|H|) + \log(1/\beta')}{\alpha'^2}\right)$, with probability of $1-\beta'$ uniform convergence holds:

$$\forall h \quad |\hat{err}(h) - err(h)| < \alpha' \quad (2)$$

Finally lets analyze the error of $\hat{h}_M$:

Using (1) and (2), with probability $1 - \underbrace{2\beta'}_{\beta :=}$:

$$\begin{aligned}
err(\hat{h}_M) &\leq \hat{err}(\hat{h}_M) + \alpha' && \text{via (2)} \\
&\leq \hat{err}(\hat{h}_{ERM}) + 2\alpha' && \text{via (1)} \\
&\leq \hat{err}(h^*) + 2\alpha' && \text{using ERM minimizes } \hat{err}. \\
&\leq err(h^*) + \underbrace{3\alpha'}_{\alpha :=} && \text{via (2)}
\end{aligned}$$

$\Rightarrow$ If $m \geq \theta\left(\left(\log(|H|) + \log(1/\beta)\right) \cdot \left(\frac{1}{\alpha^2} + \frac{1}{\alpha\varepsilon}\right)\right)$

then $\Pr\left[err(\hat{h}_M) \leq \min_{h \in H} err(h) + \alpha\right] \geq 1-\beta$

Note that for $\varepsilon > \alpha$ privacy is "free": That is the sample complexity increases by a constant factor. Otherwise, it's linear in terms of $\alpha$ and $\varepsilon$.