

## Notes 10: Hypothesis testing

### 1. CHERNOFF BOUNDS

Due to Herman Rubin

$X_1, \dots, X_m$  independent  $\{0, 1\}$ -valued random variables

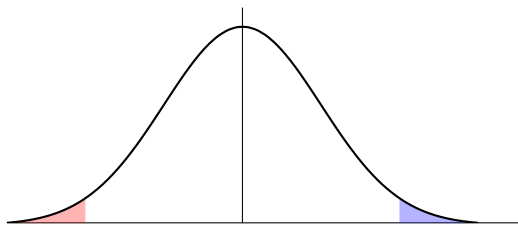
s.t.  $\mathbb{P}[X_i = 1] = \mathbb{E}[X_i] = p$  for  $1 \leq i \leq m$

$X \stackrel{\text{def}}{=} X_1 + \dots + X_m$  ( $\mathbb{E}[X] = mp$ )

**Theorem 1** (Multiplicative Chernoff). For all  $0 \leq \gamma \leq 1$ ,

$$\mathbb{P}[X \leq (1 - \gamma)mp] \leq e^{-\frac{1}{2}\gamma^2 mp}$$

$$\mathbb{P}[X \geq (1 + \gamma)mp] \leq e^{-\frac{1}{3}\gamma^2 mp}$$



Also true for  $X_1, \dots, X_m$  independent  $[0, 1]$ -valued (i.e. bounded) random variables

Many proofs; see e.g. Mulzer “Five Proofs of Chernoff’s Bound with Applications” if interested  
Exponential decay

### 2. HYPOTHESIS TESTING

Fix  $h \in \mathcal{H}$ , how can we test whether  $h$  is bad? (i.e.  $\text{err}_{\mathcal{D}}(h, c) = \mathbb{P}_{x \in \mathcal{D}}[h(x) \neq c(x)] \geq \varepsilon$ )

Solution: Draw  $m$  independent labelled samples  $(x^1, c(x^1)), \dots, (x^m, c(x^m))$ ,

Compute (empirical error)  $\widehat{\text{err}} \stackrel{\text{def}}{=} \frac{\#\text{samples s.t. } h(x^i) \neq c(x^i)}{m}$

By Chernoff bound,  $\widehat{\text{err}} \approx \text{err}_{\mathcal{D}}(h, c)$

e.g. If  $h$  is bad,  $p \stackrel{\text{def}}{=} \text{err}_{\mathcal{D}}(h, c) \geq \varepsilon$ ,

$$\mathbb{P}\left[\widehat{\text{err}} \leq \frac{\varepsilon}{2}\right] \leq e^{-\frac{1}{8}mp} \leq e^{-\frac{1}{8}\varepsilon m}$$

Further Improved Algorithm: Similar to Improved Algorithm

But only cover  $1 - \varepsilon/2$  fraction of positive samples using  $S_{i_1}, \dots, S_{i_k}$

Number of sets needed  $k \leq \text{OPT} \cdot \ln(2/\varepsilon)$  (why?)

Can show that  $O\left(\frac{1}{\varepsilon}(\ln \frac{1}{\delta} + s \ln \frac{1}{\varepsilon} \ln n)\right)$  samples suffices (exercise)

### 3. Hoeffding Inequality

**Theorem 2** (Hoeffding). Let  $X_1, \dots, X_n$  be independent random variables in  $[0, 1]$ . Let  $\bar{X} = \frac{1}{n} \sum_{1 \leq i \leq n} X_i$  be their empirical average. Then for any  $t \geq 0$ ,

$$\mathbb{P}[\bar{X} \geq \mathbb{E}[\bar{X}] + t] \leq \exp(-2nt^2).$$

See Wikipedia page on Hoeffding’s inequality for a proof if interested