

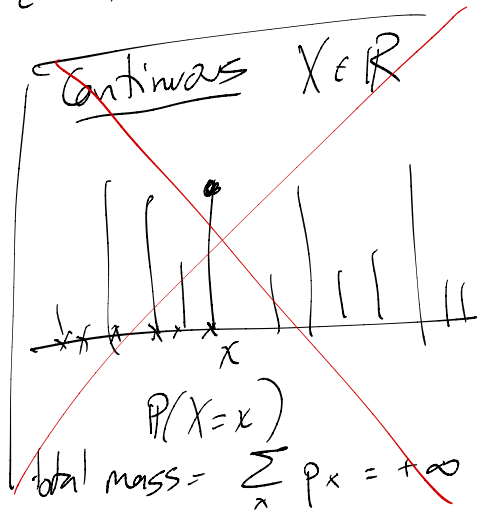
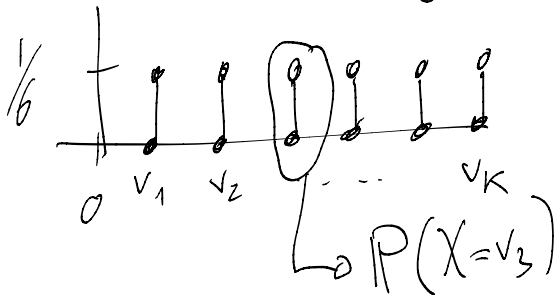
## 4. Generalization bounds

**X** = random variable  $\in \{v_1, v_2, \dots, v_K\}$ .

Ex: dice:  $K=6$ .  $v_1=1, v_2=2, \dots$

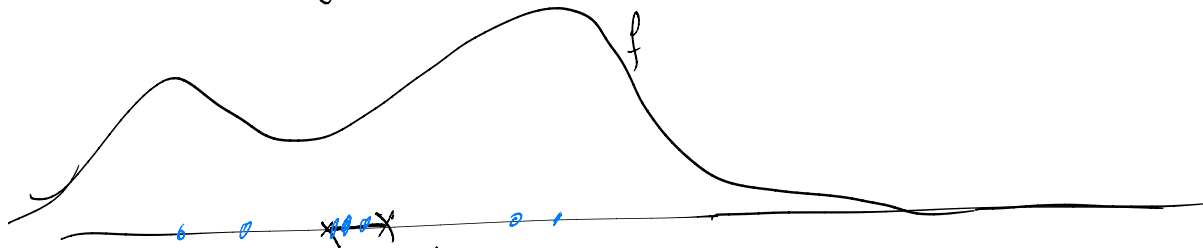
each of the  $v_i$  has probab  $p_i$

Ex: dice.  $p_i = \frac{1}{6}$



~~mass at a point +~~

mass of a very small interval containing the point



$$P(X \in [x, x+dx]) = \underbrace{f(x)dx}_{\text{density of } X}$$

back to discrete:

$$E[X] = \sum_i p_i v_i$$

"expectation" of  $x$

$$\begin{aligned} \underline{E}x: \quad & \frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \dots + \frac{1}{6} \cdot 6 = \frac{1}{6} (1+2+\dots+6) \\ & = \frac{1}{6} \cdot \frac{6(6+1)}{2} = 3.5 \end{aligned}$$



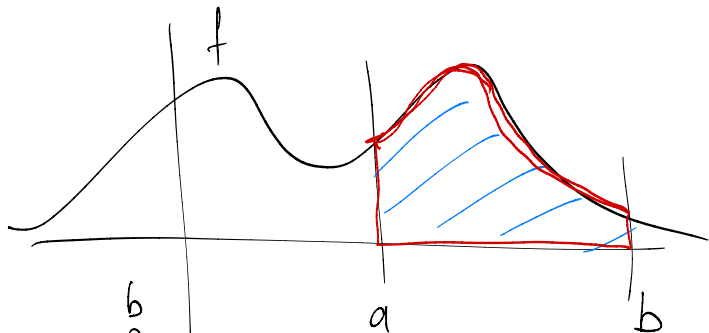
Continuous: (if we have a density)



$$\sum p_i x_i \longleftrightarrow \text{"} \sum f(x_i) dx_i \text{"}$$

$$\int f(x) dx$$

$\hookrightarrow s$



for "nice" functions,  $\int_a^b f(x) dx = \text{area under curve}$   $\mathcal{L}$

## Properties of $\mathbb{E}$ :

- $\mathbb{E}[1] = 1$

$$\left( X=1, \quad v_1=1, \quad p_1=1 \right. \\ \left. \sum p_i v_i = 1 \cdot 1 = 1 \right)$$

- $\mathbb{E}[X+Y] = \mathbb{E}[X] + \mathbb{E}[Y]$

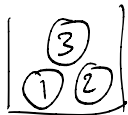
- $\forall \alpha \in \mathbb{R}, \mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$

Independence:  $X \sqcup Y \iff$   
"is independent"

$$P(X=x \text{ and } Y=y) = P(X=x) \cdot P(Y=y)$$

Ex: two dices

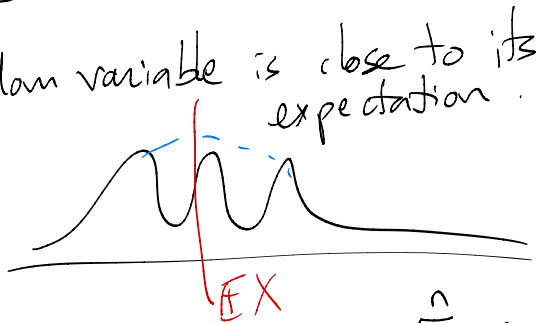
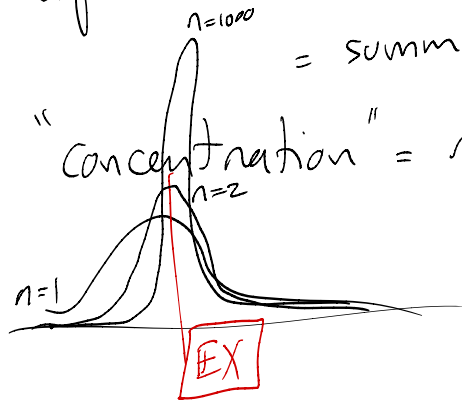
Counter-example: bucket.





random variable = model for reality.

expectation = theoretical average  
= summary information.



$X_1, \dots, X_n$   $n$  ind. random variables.  
same variance  $\sigma^2 > 0$ .

$$\bar{X}_n = \frac{1}{n} \sum_{i=1}^n X_i$$

Q:  $\text{Var}(\bar{X}_n)$ ?

Reminder:  $\text{Var}(X) = E[X^2] - E[X]^2$

Claim:  $\text{Var}(X) = E[(X - EX)^2]$  (König-Huygens)

$$\begin{aligned} &= E[X^2 - 2X(EX) + (EX)^2] \\ &= E[X^2] - 2(EX)^2 + (EX)^2 \quad \square \end{aligned}$$

summary  
info of order 2

$$\bullet \text{Var}(\lambda X) = \lambda^2 \text{Var}(X) \quad \left( \begin{aligned} &E[(\lambda X - E(\lambda X))^2] \\ &= E[(\lambda(X - EX))^2] \\ &= E[\lambda^2 (X - EX)^2] \end{aligned} \right)$$

$$\bullet \text{Var}(X+Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)$$

$$\text{where } \text{Cov}(X, Y) = E[(X - EX)(Y - EY)] \quad (X \perp Y \Rightarrow \text{Cov}(X, Y) = 0)$$

$$\text{Var}(\bar{X}_n) = \text{Var}\left(\frac{1}{n} \sum_{i=1}^n X_i\right)$$

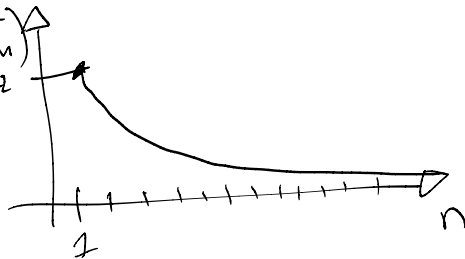
$$= \frac{1}{n^2} \text{Var}\left(\sum_{i=1}^n X_i\right)$$

$$= \frac{1}{n^2} \left( \text{Var}(X_1) + \text{Var}(X_2) + \dots + \text{Var}(X_n) + \right. \\ \left. 2 \text{Cov}(X_1, X_2) + 2 \text{Cov}(X_1, X_3) + \dots \right)$$

$= 0$

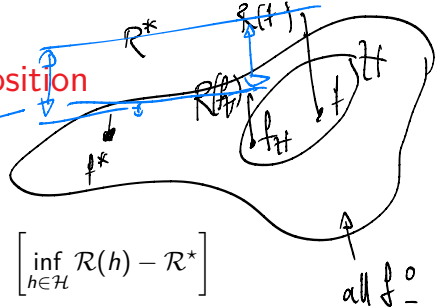
$$= \frac{1}{n^2} \cdot n \sigma^2$$

$\text{Var}(X_n)$



$$\text{Var}(\bar{X}_n) = \frac{\sigma^2}{n}$$

## Reminder: risk decomposition



### ► Reminder:

on  $f \in \mathcal{H}$

$$\underbrace{R(f) - R^*}_{\text{excess risk}} = \underbrace{\left[ R(f) - \inf_{h \in \mathcal{H}} R(h) \right]}_{\text{estimation error}} + \underbrace{\left[ \inf_{h \in \mathcal{H}} R(h) - R^* \right]}_{\text{approximation error}}$$

"how rich  $\mathcal{H}$  is"

### ► Estimation error:

- always non-negative
- random if there is randomness in the creation of  $f$
- characterizes how much we loose by picking the wrong predictor in a given class

### ► Approximation error: $= F(\mathcal{H})$

- deterministic, does not depend on  $f$ , **only on the class of functions  $\mathcal{H}$**
- characterizes how much we loose by restricting ourselves to a given class

$$\hat{f} \leftarrow \underset{g \in \mathcal{H}}{\operatorname{argmin}} \hat{\mathcal{R}}(g) \text{ (ERM)}$$

## Decomposition of the estimation error

► **Notation (i):**  $f_{\mathcal{H}} \in \arg \min_{f \in \mathcal{H}} \mathcal{R}(f)$ , best predictor in our function class

► **Notation (ii):**  $\hat{f}$  empirical risk minimizer

► **Useful decomposition:**

$\hat{\mathcal{R}}(\hat{f})$  minimal on  $\mathcal{H}$   
(for any  $g \in \mathcal{H}$ ,  $\hat{\mathcal{R}}(g) \geq \hat{\mathcal{R}}(\hat{f})$ )  
(def. of  $f_{\mathcal{H}}$ )

$$\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) = \mathcal{R}(\hat{f}) - \mathcal{R}(f_{\mathcal{H}})$$

$$= \mathcal{R}(\hat{f}) - \hat{\mathcal{R}}(\hat{f}) + \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{\mathcal{H}}) + \hat{\mathcal{R}}(f_{\mathcal{H}}) - \mathcal{R}(f_{\mathcal{H}})$$

$$\leq \sup_{f \in \mathcal{H}} \{ \mathcal{R}(f) - \hat{\mathcal{R}}(f) \} + \hat{\mathcal{R}}(\hat{f}) - \hat{\mathcal{R}}(f_{\mathcal{H}}) + \sup_{f \in \mathcal{H}} \{ \hat{\mathcal{R}}(f) - \mathcal{R}(f) \}$$

► middle term is  $\leq 0$  by definition, and we get

$$\mathcal{R}(\hat{f}) - \inf_{f \in \mathcal{H}} \mathcal{R}(f) \leq 2 \sup_{f \in \mathcal{H}} | \hat{\mathcal{R}}(f) - \mathcal{R}(f) | = \max_{f \in \mathcal{H}} F(f)$$

↑  
hard to get



## Decomposition of the estimation error, ctd.

- ▶ **Remark (i):** no more dependency in  $\hat{f}$ , we only need to control functions (but we do need **uniform control**)
- ▶ **Remark (ii):** if  $\hat{f}$  not global minimizer, say

$$\hat{\mathcal{R}}(\hat{f}) \leq \inf_{f \in \mathcal{H}} \hat{\mathcal{R}}(f) + \varepsilon,$$

we need to add  $\varepsilon$  to our bound

- ▶ **Remark (iii):** bound usually grows with size of  $\mathcal{H}$  and decreases with  $n$

## 4.1. Uniform bounds via concentration

## Concentration inequalities

- ▶ informally speaking: random variable is “close” to its expectation with high probability
- ▶ **Example:** Markov, Chebyshev
- ▶ more involved:

$\approx 1/60$

**Proposition (Hoeffding's inequality):** let  $Z_1, \dots, Z_n$  be independent random variables such that  $Z_i \in [0, 1]$  almost surely, then, for any  $t \geq 0$ ,

$$\mathbb{P} \left( \left| \frac{1}{n} \sum_{i=1}^n (Z_i - \mathbb{E}[Z_i]) \right| \geq t \right) \leq 2 \exp(-2nt^2) .$$

Boucheron, Lugosi, Massart, Concentration inequalities: A non-asymptotic theory ... Oxford Univ Press. 2013

$$\frac{1}{n} \sum_{i=1}^n (Z_i - \mathbb{E}Z_i) = \frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{n} \sum_{i=1}^n \mathbb{E}Z_i$$

$$Z_1, \dots, Z_n \text{ iid} \longrightarrow \underbrace{\frac{1}{n} \sum_{i=1}^n Z_i}_{\text{empirical average}} - \underbrace{\mathbb{E}Z_1}_{\text{true average / common expectation}}$$

Probabilistic upper bounds: assume  $\mathcal{H} = \{f\}$ .

$$P(\hat{R}(f) - R(f) \leq t) \geq 1 - \delta = 99\% \quad (\text{upper bound point of view})$$
$$\Leftrightarrow P(\hat{R}(f) - R(f) > t) \leq \delta = 1\% \quad (\text{ex: } \delta = 0.01)$$

$f$  = fixed function,  $t$  = fixed threshold.

$$P(\text{event}) = 1 - P(\text{converse event})$$

complement

$\hat{R}(f) - R(f) < 1000$ , with probability 0.8



Chebyshev:  $P(|Z - \mathbb{E}Z| > t) \leq \frac{\sigma^2}{t^2}$

$Z := \hat{R}(f)$ ,  $\mathbb{E}Z = R(f)$ .

Q:  $\text{Var}(Z)$ ? (recall that  $(X_i, Y_i)$  are independent <sup>identically distributed</sup>)

$$\text{Var}\left(\frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i))\right) = \frac{1}{n^2} \sum_i \text{Var}(\ell(Y_i, f(X_i))) + 0$$

$$= \frac{1}{n^2} \cdot n \cdot \text{Var}(\ell(Y, f(X)))$$

$$= \frac{\text{Var}(\ell(Y, f(X)))}{n}$$

$$\begin{aligned} \text{Var}(X+Y) &= \text{Var}(X) \\ &\quad + \text{Var}(Y) \\ &\quad + 2\text{Cov}(X, Y) \end{aligned}$$

Non-independent case:  $Z_1 = Z_2 = \dots = Z_n$ .

$$\begin{aligned}\text{Var}\left(\frac{1}{n} \sum_{i=1}^n Z_i\right) &= \text{Var}\left(\frac{1}{n}(Z_1 + Z_1 + \dots + Z_1)\right) \\ &= \text{Var}(Z_1)\end{aligned}$$

let us say  $l$  is bounded ( $l \leq l_\infty$ ).

EX:  $l = 0-1\text{-loss} \Rightarrow l_\infty = 1$ .  
( $l(y, \hat{y}) = \mathbb{1}_{y \neq \hat{y}}$ )

Q:  $\text{Var}(l(Y, f(X))) \leq ?$

$$\mathbb{E}\left[\left(\frac{l(Y, f(X)) - \mathbb{E}l}{\mathbb{E}l(Y, f(X)) \leq l_\infty}\right)^2\right] \leq \mathbb{E}[4l_\infty^2] \leq 4l_\infty^2$$

$|l - \mathbb{E}l| \leq 2l_\infty$   
 $\leq |l| + |\mathbb{E}l|$

better: Popoviciu's inequality:  $\frac{l_\infty^2}{2}$ .



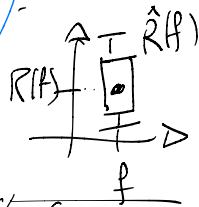
so far: •  $P(|\hat{R}(f) - R(f)| > t) \leq \frac{\text{Var}(\ell(Y, f(X)))}{n t^2}$

•  $\text{Var}(\ell(Y, f(X))) \leq 4 l_\infty^2$

$\Rightarrow P(|\hat{R}(f) - R(f)| > t) \leq$

$\frac{4 l_\infty^2}{n t^2} - \delta$

"hair cut":  $\frac{4 l_\infty^2}{n t^2} = \delta \Leftrightarrow t^2 = \frac{4 l_\infty^2}{n \delta}$



$\Leftrightarrow t = \frac{2 l_\infty}{\sqrt{n \delta}}$

with probab  $\geq 1 - \delta = 99\%$   $\delta = 0.01$

$\hat{R}(f) - R(f) \leq \frac{2 l_\infty}{\sqrt{n \delta}} \left( \frac{2.1}{\sqrt{n \cdot 0.1}} \right)$

with probability at least  $1-\delta$ ,  $\frac{1}{n}(0+1+0+0+\dots+n)$

$$\hat{R}(f) - R(f) \leq \frac{2l_\infty}{\sqrt{n}\sqrt{\delta}}$$

$$\hat{R}(f), R(f) \in [0, 1]$$

Example: 0-1 loss:  $l_\infty = 1$ , 99% proba:  $\delta = 0.01$   
 $\Rightarrow \sqrt{\delta} = 0.1$ . then: with proba  $\geq .99$ ,

$$\hat{R}(f) - R(f) \leq \frac{20}{\sqrt{n}}$$

$$n = 100: \sqrt{n} = 10, \quad \hat{R}(f) - R(f) \leq 2$$

$$n = 10^6: \sqrt{n} = 10^3$$

$$\frac{20}{1000} = 0.02$$



rate  $\rightarrow \frac{1}{\sqrt{n}} e^n$

really bad.

- $\rightarrow$  1) need  $n$  high
- 2) need smarter concentration inequalities.

## Single function

- ▶ assume that  $\mathcal{H} = \{f_0\}$  and  $\ell$  a bounded loss function
- ▶ then we can control

$$\sup_{f \in \mathcal{H}} \left| \hat{\mathcal{R}}(f) - \mathcal{R}(f) \right| = \hat{\mathcal{R}}(f_0) - \mathcal{R}(f_0) = \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f_0(X_i)) - \mathbb{E}[\ell(Y, f_0(X))] .$$

- ▶ indeed, since the observations are i.i.d., we can use Hoeffding on the  $Z_i := \ell(Y_i, f_0(X_i))$
- ▶ common expectation =  $\mathcal{R}(f_0)$
- ▶ for any  $\delta \in (0, 1/2)$ ,

$$\mathbb{P} \left( \left| \hat{\mathcal{R}}(f_0) - \mathcal{R}(f_0) \right| \geq \frac{1}{\sqrt{2n}} \sqrt{\log \frac{1}{\delta}} \right) \leq 2 \exp \left( -2n \frac{1}{2n} \log 1/\delta \right) = 2\delta .$$

Recall: (Hoeffding)  $P\left(\left|\frac{1}{n}\sum_{i=1}^n Z_i - \mathbb{E}Z\right| \geq t\right) \leq 2\exp(-2nt^2)$   
(1%)

say that  $Z_i = \ell(Y_i, H(X_i))$ ,  $|\ell| \leq 1$

$$\Rightarrow P(|\hat{R}(f) - R(f)| \geq t) \leq 2\exp(-2nt^2) = \delta$$

Q: what choice of threshold  $t$  to guarantee probab.  $< 1\%$ ?

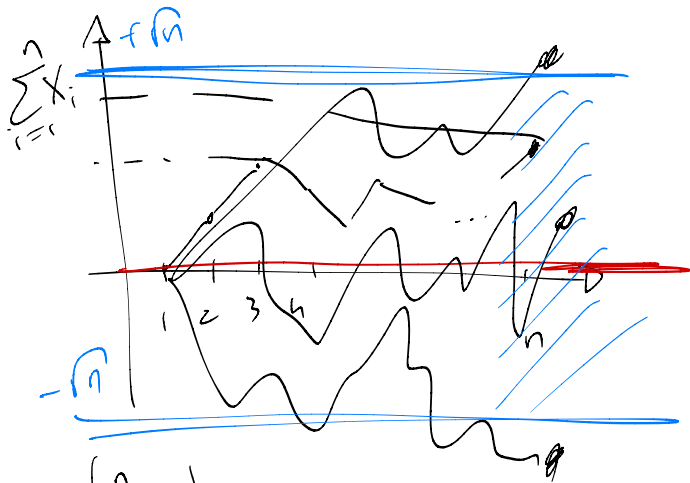
$$\exp(-2nt^2) = \delta/2 \quad \Rightarrow -2nt^2 = \log \frac{\delta}{2}$$

$$\Rightarrow t = \sqrt{\frac{1}{2n} \cdot -\log \frac{\delta}{2}}$$

$$P\left(\frac{1}{n}\sum - \mathbb{E} > \frac{1}{\sqrt{n}}\right) \leq \delta$$

$$\Leftrightarrow P\left(\sum_{i=1}^n X_i - n\mathbb{E}X_1 > \sqrt{n}\right) \leq \delta$$

"sums of indep. r.v.  
have typical deviations  
of order  $\sqrt{n}$ "



$$P\left(\left|\sum_{i=1}^n X_i\right| > r\sqrt{n}\right) \leq 0.01$$

$$X_i \in \{-1, +1\}$$

$$P(X_i = -1) = P(X_i = +1)$$

$$E\left[\sum_{i=1}^n X_i\right] = \sum_{i=1}^n E X_i = 0$$

$$E X_i = \sum \text{prob of outcome} \times \text{value of } \text{---}$$

$$= \frac{1}{2} \cdot (-1) + \frac{1}{2} \cdot (+1)$$

$$= 0.$$

Non-independent:  $X_1 = X_2 = \dots = X_n$ .

$$\sum X_i = n X_1 \quad P(X_1 = -1) = P(X_1 = +1) = \frac{1}{2}$$

$$\begin{aligned} P(|\sum X_i| > \sqrt{n}) &= P(|n X_1| > \sqrt{n}) \\ &= P(|X_1| > \frac{1}{\sqrt{n}}) \\ &= P(1 > \frac{1}{\sqrt{n}}) \\ &= 1 \quad (\text{whenever } n \geq 2). \end{aligned}$$