LECTURE 1

- Big data:
  - (a) # observations (data points) $\longrightarrow$ Big
  - (b) # dimensions (parameters) $\longrightarrow$ Big

- Examples:
  1. Income of Kyiv population
     = 3,000,000 observations, dimension = 1.
  2. Images of people on FB
     each observation = 100×100 image
     Each pixel = dimension $\Rightarrow$ #dimensions = 10^4.
     HD.
  3. Other HD examples: text; sound; video; genome; medical history; chess games.

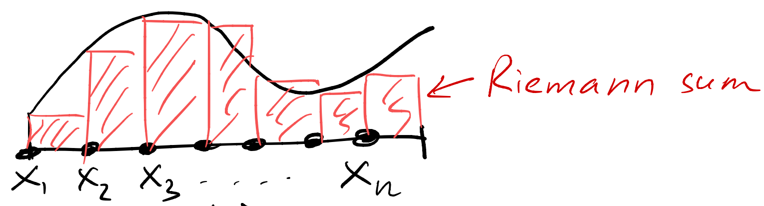
- Empirical Observation: it is exponentially harder to deal with large # of dimensions then with large # of observations.
  - $\hookrightarrow$ classical statistics, probability (via limit thms)
  - $\longrightarrow$ HDS, HDP (new).

WHY exponentially harder?   Example:

Problem (HD integration) Numerically compute the integral

$$\int_0^1 \cdots \int_0^1 f(\underbrace{x_1, \dots, x_d}_{\text{parameters}})\, dx_1 \cdots dx_d = \int_{[0,1]^d} f(x)\,dx$$

of a given function f (e.g. f = model of income, $\int f\,dx$ = population income)

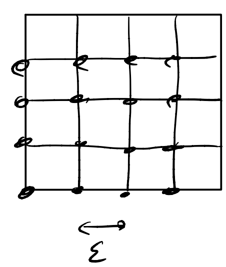


- If $d = 1$: use the grid

$$\int_{[0,1]^d} f(x)\,dx \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i).$$

resolution $= \varepsilon \Rightarrow n = 1/\varepsilon$ pts.

- If $d = 2$: same



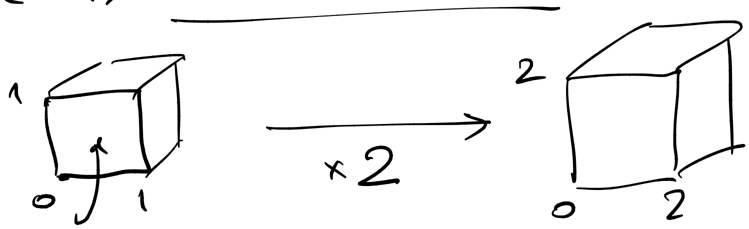
$\Rightarrow n = (1/\varepsilon)^2$ pts.

- For general d, $N = (1/\varepsilon)^d$ pts.

<u>Exponential</u> in d $\Rightarrow$ too large.

$\Rightarrow$ complexity of many alg's is <u>exponential in dimension</u>

- There is <u>too much room</u> in H.D.'s:



volume = 1   volume = 2^d !

"<u>THE CURSE OF DIMENSIONALITY</u>"

Probability for rescue: Monte-Carlo method

- Instead of choosing x_i on the grid, choose them at random (independently, uniformly in $[0,1]^d$)

$\Rightarrow f(x_i)$ are i.i.d. r.v.'s. $\quad \frac{1}{d}\sum_{i=1}^{d} f(x_i) \overset{?}{\approx} \int_{[0,1]^d} f(x)\,dx$

- Will use the following standard facts of probability theory:

1. Expectation of a r.v. X with density $p(x)$: (def)

$$\mathbb{E}X = \int_{-\infty}^{\infty} x \cdot p(x)\,dx$$

More generally,

$$\mathbb{E}f(X) = \int_{-\infty}^{\infty} f(x)\,p(x)\,dx$$

More generally, if X is a random vector in $\mathbb{R}^n$,

$$\mathbb{E}f(X) = \int_{\mathbb{R}^n} f(x)\,p(x)\,dx$$

2. Variance of a r.v. X:

$$\mathrm{Var}(X) = \mathbb{E}(X - \mathbb{E}X)^2$$

3. Linearity: (a) $\mathbb{E}(X_1 + \dots + X_n) = \mathbb{E}X_1 + \dots + \mathbb{E}X_n$

(b) If X_i are independent,

$$\mathrm{Var}(X_1 + \dots + X_n) = \mathrm{Var}(X_1) + \dots + \mathrm{Var}(X_n)$$

4. (Strong) Law of large numbers (SLLN):

If $X, X_1, X_2, \dots$ are independent and identically distributed (iid) random variables, then

$$\frac{1}{N}\sum_{i=1}^{N} X_i \longrightarrow \mathbb{E}X \quad \text{almost surely (a.s)}$$

i.e. with probability $=1$.

In our situation,

- $$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right] = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbb{E}f(X_i)}_{\substack{\parallel \\ \mathbb{E}f(X) \text{ by identical distribution}}} = \mathbb{E}f(X)$$

$$= \int_{\mathbb{R}^d} f(x)\,p(x)\,dx, \quad \text{where density } p(x) = \begin{cases} 1, & x \in [0,1]^d \\ 0, & \text{---} \end{cases}$$

(uniform distribution)

$$= \int_{[0,1]^d} f(x)\,dx.$$

$\Rightarrow$ we have an unbiased estimator of the integral

- SLLN $\Rightarrow$ $\boxed{\frac{1}{d}\sum_{i=1}^{d} f(X_i) \longrightarrow \int_{[0,1]^d} f(x)\,dx \quad \text{a.s.}}$

- Rate of convergence? L^2 error:

$$\mathbb{E}\Bigg(\underbrace{\frac{1}{n}\sum_{i=1}^{n} f(X_i)}_{\substack{\parallel \\ : \\ Z}} - \underbrace{\int_0^1 f(x)\,dx}_{\substack{\parallel \\ \mathbb{E}Z}}\Bigg)^2 = \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n} \underbrace{\mathrm{Var}(f(X_i))}_{\substack{\parallel \\ \mathrm{Var}(f(X)) \text{ by identical distribution}}} = \frac{\mathrm{Var}(f(X))}{n}$$

$$\leq \frac{1}{n} \quad \text{e.g. if } |f(x)| \leq 1 \quad \forall x.$$

- Taking square root $\Rightarrow$ expected error $= \boxed{O(1/\sqrt{n})}$

regardless of dimension !!