Record

LECTURE 1

- Big data:
  - (a) # observations (data points) → Big
  - (b) # dimensions (parameters describing each data pt.) → Big

- Examples:
  1. Income of Kyiv population
     = 3,000,000 observations, dimension = 1. low dim.
  2. Avatars of people on FB
     each avatar = 100×100 image
     Each pixel = dimension ⇒ #dimensions = 10^4. high dim.
  3. Other HD examples: text; sound; video; genome; medical history; chess games.

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

WHY exponentially harder? Example:

Problem (HD integration) Numerically compute the integral

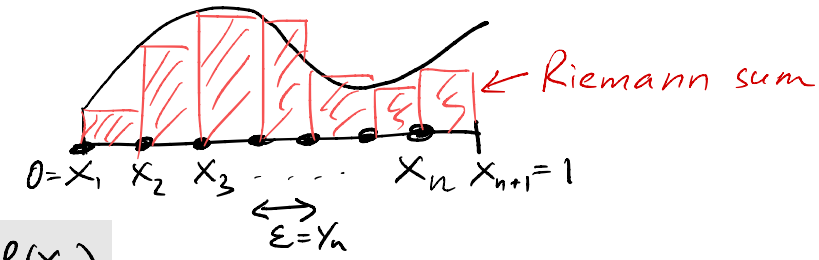
$$\int_0^1 \cdots \int_0^1 f(x_1, \ldots, x_d)\, dx_1 \cdots dx_d = \int_{[0,1]^d} f(x)\, dx$$

($x_1, \ldots, x_d$: parameters)

of a given function f

(e.g. f = model of income, $\int f\, dx$ = total income)

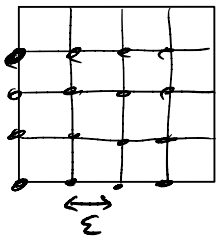
- If $d = 1$: use the grid



$$\int_0^1 f(x)\, dx \approx \sum_{i=1}^{n} (x_{i+1} - x_i) f(x_i) = \frac{1}{n} \sum_{i=1}^{n} f(x_i)$$

"Resolution" $\varepsilon = 1/n$

- If $d = 2$, do similarly but use the grid



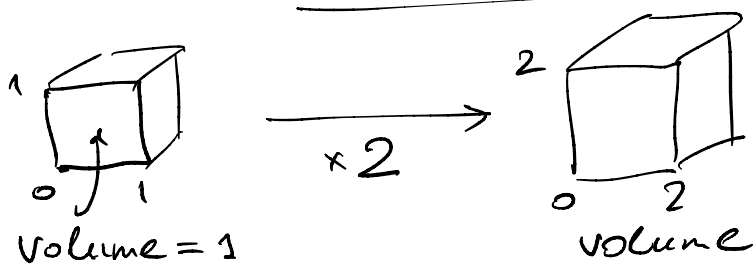
$\Rightarrow$ $n = (1/\varepsilon)^2$ pts.

- For general dimension d, we need $n = (1/\varepsilon)^d$ pts.

Exponential in d. Too large.

Complexity of many alg's is exponential in dimension

- Why? There is too much room in H.D.'s:



"THE CURSE OF DIMENSIONALITY"

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.

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

- Will use the following standard facts of probability theory:

① **Def** A r.v. X has **density** (= pdf = probability density function) $p(x)$ if $\mathbb{P}\{X \in A\} = \int_A p(x)\,dx$ $\forall$ Borel $A \subset \mathbb{R}$

In this case we say that X has a **continuous distribution**.

- $p(x) \geq 0$ $\forall x$; $\int_{-\infty}^{\infty} p(x)\,dx = \mathbb{P}\{X \in (-\infty,\infty)\} = 1$.
- Similarly for a **random vector** taking values in $\mathbb{R}^n$. Density $p: \mathbb{R}^n \to \mathbb{R}$

Example : uniform distribution. $X \sim \mathrm{Unif}([0,1])$ if $p(x) = \begin{cases} 1, & x \in [0,1] \\ 0, & \end{cases}$

$X \sim \mathrm{Unif}([0,1]^d)$ if $p(x) = \begin{cases} 1, & x \in [0,1]^d \\ 0, & \end{cases}$

② **Def** The **expected value** (expectation) of a r.v. X with density $p(x)$ is

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

More generally, $\forall$ function $f: \mathbb{R} \to \mathbb{R}$,

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

Similarly for a random vector X : $\mathbb{E}f(x) = \int_{\mathbb{R}^n} f(x)\,dx$.

② Variance of a r.v. X:

$$\mathrm{Var}(X) = \mathbb{E}\left(X - \mathbb{E}X\right)^2 = \mathbb{E}[X^2] - (\mathbb{E}X)^2$$

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

(b) If X_i are independent,

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

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

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

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

i.e. with probability $= 1$.

Back to our situation: $X_i \sim \text{Unif}([0,1]^n)$ independent and identically distributed (iid)

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

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

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

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

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

- Rate of convergence? L^2 error ("MSE"):

$$\mathbb{E}\left(\underbrace{\frac{1}{n}\sum_{i=1}^{n} f(X_i)}_{\| \; Z} - \underbrace{\int_{[0,1]^d} f(x)\, dx}_{\| \; \mathbb{E}Z}\right)^2 = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} f(X_i)\right)$$

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

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

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

regardless of dimension !! ☺