

Lecture Notes-1

Arnab Basu and Aleyna Kof

¹ Bilkent University

² `arnab@bilkent.edu.tr`

`aleyna.kof@bilkent.edu.tr`

1 Markov Inequality

$$X \geq 0, P[X \geq \epsilon] \leq \frac{E[X]}{\epsilon} \quad (1)$$

Why is this true?

$$\begin{cases} 1, & X \geq a \\ 0, & \text{otherwise} \end{cases}$$

What is $E[X]$? $E[I] \leq \frac{E[X]}{a} = P[X \geq a]$

2 Tchebysev Inequality

X is a random variable

$$P[\underbrace{|X - E[X]| \geq \epsilon}_{P[(X - E[X])^2 \geq \epsilon^2]}] \leq \frac{Var(X)}{\epsilon^2}$$

$$P[(X - E[X])^2 \geq \epsilon^2] \leq \frac{E[(X - E(X))^2]}{\epsilon^2} = \frac{Var(X)}{\epsilon^2} \quad (2)$$

3 Strong Law of Large Numbers (SLLN)

It talks about the distribution of the random variables (it does not have to be identical) $X_1, X_2, \dots, X_n$ independent random variables (identical) $\sim X$

What does independent random variable mean?

Consider $X, Y \Rightarrow P[X|Y] = P[X]$ and $P[Y|X] = \frac{P[XY]}{P[X]} = \frac{P[X]P[Y]}{P[X]} = P[Y]$

Any subset of independent random variable sequence is independent. Take two random variable from $X_1, \dots, X_n \Rightarrow X_1, X_2$ are independent. Then,

$$P\left[\left|\underbrace{\frac{X_1 + \dots + X_n}{n}}_{F_n} - E[X]\right| \geq \epsilon\right] \leq \frac{Var(X)}{n\epsilon^2} \quad (3)$$

Basically says, if there is huge dataset, sample average will not deviate from the

actual average more than ϵ (Consider $n \rightarrow \infty, \frac{Var(X)}{n\epsilon^2} \rightarrow 0$)

$$P[|F_n - E[X]| \geq \epsilon] = P[|F_n - E[X]|^2 \geq \epsilon^2] \leq \frac{E[|F_n - E[X]|^2]}{\epsilon^2} = \quad (4)$$

$$\begin{aligned} P[|F_n - E[X]| \geq \epsilon] &= P[|F_n - E[X]|^2 \geq \epsilon^2] \\ &= \frac{1}{n^2} \\ E[|X_1 + X_2 + \dots + X_n - nE[X]|^2] &= \frac{1}{n^2} E[(X_1 - E[X]) + (X_2 - E[X]) \dots (X_n - E[X])^2] \\ &= \frac{1}{n^2} \sum_{i=1}^n E[X_i - E[X]]^2 \\ &= \frac{1}{n} Var(X) \end{aligned} \quad (5)$$

(In case of no correlation: $E[(X_i - E[X])(X_j - E[X])] = 0$)

4 C^γ Diffeomorphism

continuously differentiable isomorphism

Why sample?

- To capture the essentials of the data source high dimensional spaces, volume index (which the mass of the data at) has to be calculated
- substitution (algebraic) as geometric operations (Substitution is mapping from real line (algebraically) and substitution is mapping from $(-2\pi, 2\pi)$ (geometrically))

4.1 Isomorphism

A **continuous bijection** with continuous inverse Linear isomorphism: the isomorphism is linear

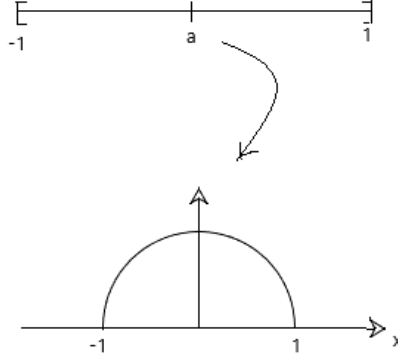
Diffeomorphism helps to take a set in certain region and maps it to another region preserving certain features

The aim of these substitution/calculation is to find $\int e^{x^2}$. Take $\phi(a) = (a, \sqrt{1-a^2})$,

$$\int_{-1}^{+1} (1-a^2) da = \frac{4}{3} \quad (6)$$

$$\phi(t) = (\cos t, \sin t) \longrightarrow [-\frac{\pi}{2}, \frac{\pi}{2}]$$

$$\int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} (\cos^2 t) dt = \frac{\pi}{2} \quad (7)$$



The way the substitution is made is significant!

$$\int_a^b f(x)dx = \lim_{n \rightarrow \infty} \sum f(x) \Delta x, \quad \Delta x = \frac{b-a}{n} \quad (8)$$

Integration geometrically means dividing into grids and then summing up the grids. However, all the grids (increment size) do not need to be the same (geometrically).

What to do? Scaling is a way to handle it.

Riemann Integral (the actual definition of the integral)

It allows uneven increments

$$\lim_{m \times m(-) \downarrow 0} \sum_i f(x)(x_{i+1} - x_i) \quad (9)$$

$$\sum_i F(a_i) \Delta a_i = \sum_i f \circ \phi(a_i) d(a_i) \quad (10)$$

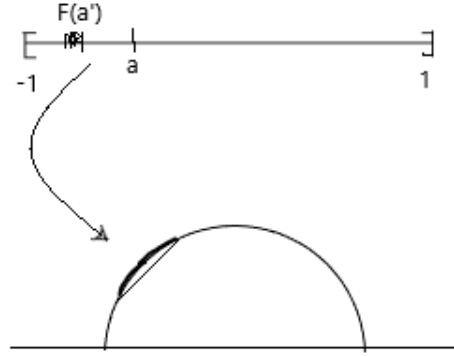
Since $a_i \Rightarrow 0$, arc length and quad length will be approximately same. Therefore, we can change (10) to

$$\sum_i F(a_i) \Delta a_i = \sum_i f \circ \phi(a_i) \times \underbrace{L(a_i)}_{\phi(a_{i+1}) - \phi(a_i)}, \quad F(a_i) = \frac{f \circ \phi(a_i) |\phi(a_{i+1}) - \phi(a_i)|}{a_{i+1} - a_i} \quad (11)$$

$$F(a) = f \circ \frac{d\phi}{da}$$

For two variables, ϕ_1, ϕ_2 ,

$$\begin{vmatrix} \frac{\delta \phi_1}{\delta x} & \frac{\delta \phi_1}{\delta y} \\ \frac{\delta \phi_2}{\delta x} & \frac{\delta \phi_2}{\delta y} \end{vmatrix}$$



Why diffeomorphism? It is required to calculate the derivative
 $A, B \subseteq \mathbb{R}^d, A \xrightarrow{\phi} B$

$$\int_A f \circ \phi |D\phi| = \int_B f \quad (12)$$

It is required to sample data points. Then, it has a distribution f and you can **pull back** from "bad" domain to "nice" one.

It is about pulling back maps of the data sets, not changing the essentials of the datasets. Therefore, the data which is sampled is same.

$$\int_{-\infty}^{\infty} e^{x^2} dx = I \quad (13)$$

In two dimension,

$$\int_{-\infty}^{\infty} e^{x^2+y^2} \underbrace{dxdy}_{\text{volume element}} = I \quad (14)$$

Take

$$(X = \gamma \cos \theta \text{ and } Y = \gamma \sin \theta) \Rightarrow \begin{vmatrix} \cos \theta & -\gamma \sin \theta \\ \sin \theta & \gamma \cos \theta \end{vmatrix} \quad (15)$$

$$I^2 = \int_0^{\infty} e^{-\gamma^2} \gamma d\gamma d\theta = \pi \quad (16)$$

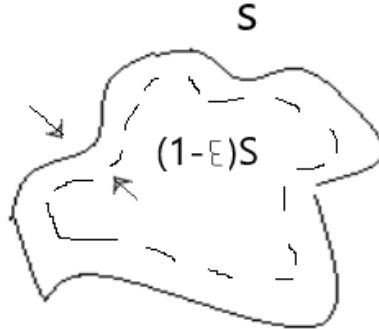


Fig. 1. The intuition behind high dimensional objects

5 High Dimensional Objects

Intuition: As in figure

$$d \gg 1$$

$$\frac{\text{Vol}((1-\epsilon)S)}{\text{Vol}(S)} = (1-\epsilon)^d \underbrace{\leq}_{(1-x) \leq e^{-x}} e^{-\epsilon d} \rightarrow 0 \quad (17)$$

In large datasets, the mass is most likely lie on frontier.