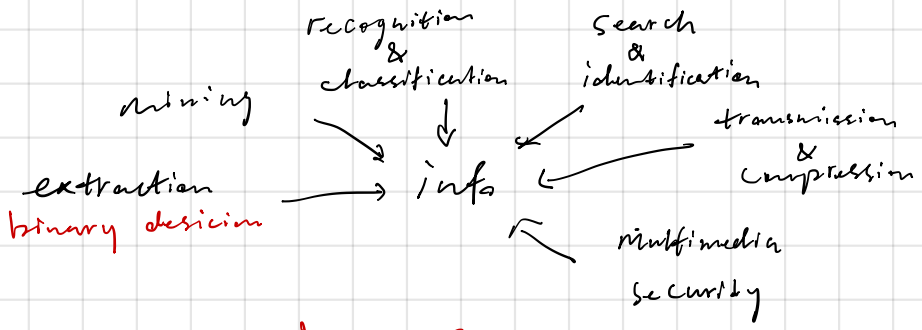
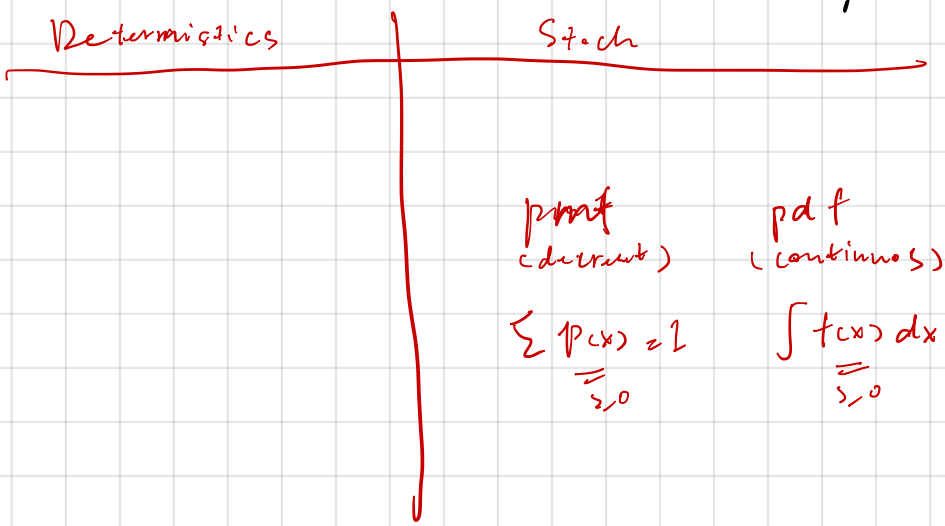
main probs of interest





prob (discrete) $\sum p(x) = 1$, ≥ 0

pdf (continuous) $\int f(x)\,dx = 1$, ≥ 0

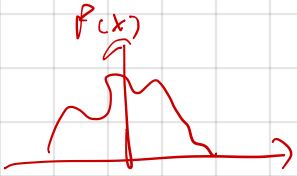
information theory & machine learning

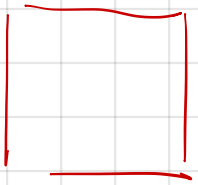
$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$X \sim f(x)$

$\mu = E(X) = \sum x p(x)$

$\sigma^2 = \sum (x-\mu)^2 p(x)$

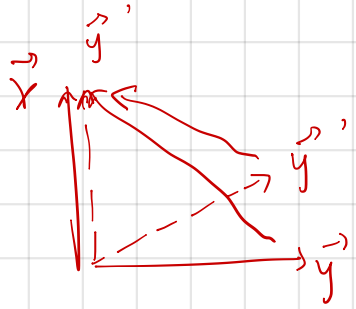




8 bit info

$2^8 = 256$ for one pixel

(μ, σ^2). most unpredicted distribution



$\langle \vec{x}, \vec{y} \rangle \to \frac{\pi}{2}$, $d_{xy} \to \max$ $\quad \langle \vec{x}, \vec{y} \rangle \to 0$ $d_{xy} \to \min$

$\cos \langle \vec{x}, \vec{y} \rangle \to 0$ $\quad \cos \langle \vec{x}, \vec{y} \rangle \to 1$

$x \sim p(x)$ $\quad y \sim q(x)$

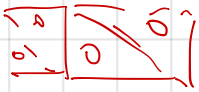
$\mathbb{E}_{p(x)} \mathbb{E}_{q(y)} \|x - y\|^2 = ?$ we don't know.

SVD

singular value decomposition

$A = \mathbb{R}^{N \times N} \Rightarrow A = U \Sigma V^T$

$m \times m$ $\quad m \times n$ $\quad n \times n$

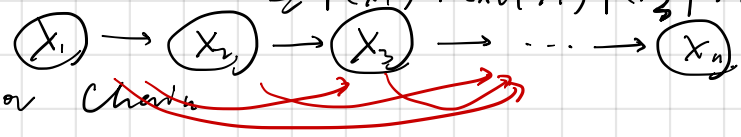
$\Sigma =$ 

Information Theory

Chain Rule $P(x, y) = P(y|x)P(x) = P(x|y)P(y)$

Bayes Rule $P(y|x) = \frac{P(x|y)P(y)}{P(x)}$

$$P(x_1, x_2, x_3 \cdots, x_n) = P(x_1 | x_2, x_3 \cdots x_n) P(x_2, x_3 \cdots x_n)$$
$$= P(x_1|x_2, x_3 \cdots x_n) P(x_2|x_3 \cdots x_n) P(x_3 \cdots x_n)$$
$$\vdots$$
$$= P(x_1|x_2, x_3 \cdots x_n) P(x_2|x_3 \cdots x_n) \cdots P(x_{n-1}|x_n) P(x_n)$$
$$= P(x_1) P(x_n|x_1) P(x_3|x_1 x_2) \cdots P(x_n|x_1, x_2 \cdots x_{n-1})$$



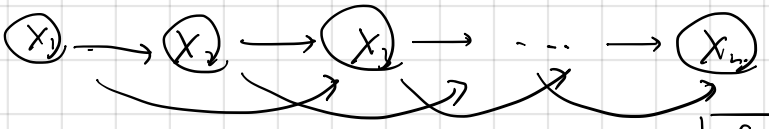
Markov Chain

first order:

$$P(x_1, x_2, x_3, \cdots, x_n) = P(x_1) P(x_2|x_1) P(x_3|x_2) \cdots P(x_n|x_{n-1})$$

second order:

$$P(x_1, x_2, x_3, \cdots, x_n) = P(x_1) P(x_2|x_1) P(x_3|x_1 x_2) \cdots P(x_n|x_{n-1}, x_{n-2})$$



forward and backward.

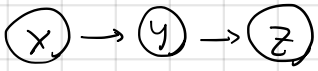
$$= \prod_{i=1}^{n} P(x_n | x_{n-1}, x_{n-2})$$

$P(x_2 | x_0, x_1) = P(x_2|x_1)$ base case.

$P(x_1 | x_{-1}, x_0) = P(x_1)$

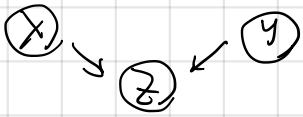
$P(x, y, z) = P(x) P(y|x) P(z|x, y)$

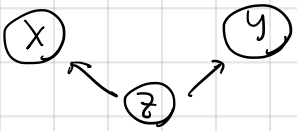
$P(x, y, z) = P(x) P(y) P(z)$



$P(x, y, z) = P(x) P(y|x) P(z|y)$

$P(x, y, z) = P(z) P(y|z) P(x|y)$





$P(x, y, z) = P(x)$

$X \sim P(X)$ $\quad X = [x_1, x_2, \cdots x_n]$, n samples

$X \in \mathcal{X}$

$$E_P(X) = \sum_{i=1}^{n} P_r(X = x_i)\, x_i$$

$$H(X) = -\sum_{i=1}^{n} P_r(X = x_i) \log_2 P_r(X = x_i) = -\sum_{x \in \mathcal{X}} P(x) \log_2 P(x)$$

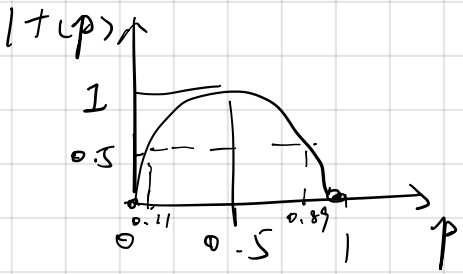
$$= E_P[-\log_2 P(x)]$$

$X \in \{0, 1\}$ $\quad X \in \mathcal{X}$

$P_r(X = 0) = p$

$P_r(X = 1) = 1 - p$

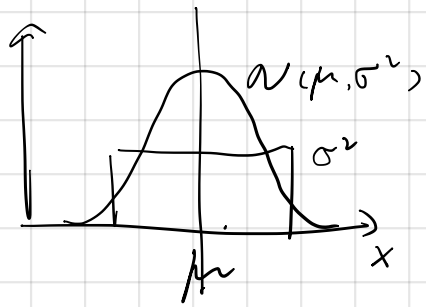
$$H(X) = -p \log_2 p - (1-p) \log_2 (1-p) = H(p)$$

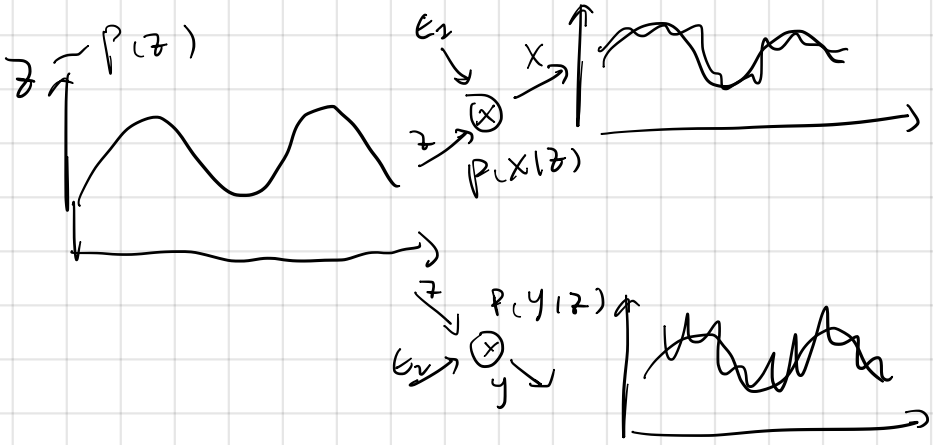


$-\log_2 p$

* uniform distribution is the most unexpected given the fixed domain $[a, b]$ so the entropy of it is the highest

* Gaussian distribution has the highest entropy given the fixed σ^2

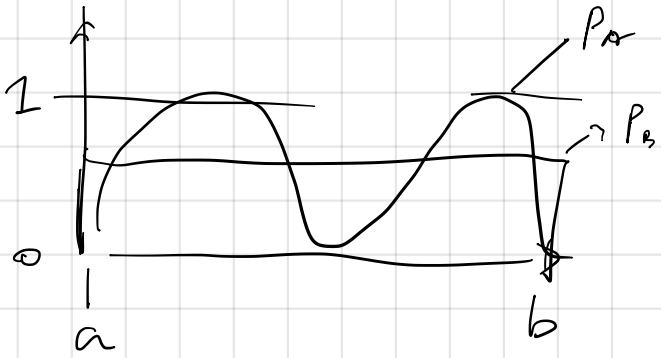




$X - z = \epsilon_1$

$\epsilon_1 \perp\!\!\!\perp \epsilon_2$ independent

$y - z = \epsilon_2$



$H(P_A) < H(P_B)$

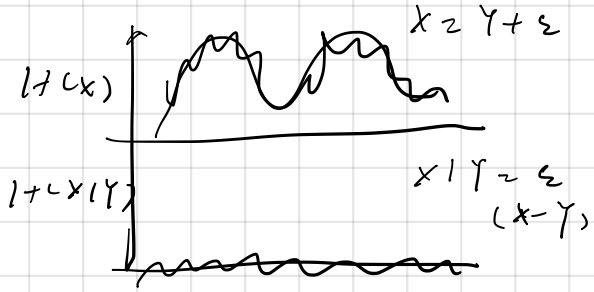
A ?

⚠ n independent $X_i \perp\!\!\!\perp X_j$

$H(X_1, X_2, \cdots X_n) = \sum_{i=1}^{n} H(X_i)$

Conditioning kills entropy

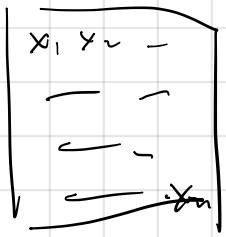
$H(X|Y) \le H(X)$



$H(X_1, X_2, \cdots X_n) \le \sum_{i=1}^{n} H(X_i)$

using 1st Markov chain to prove this

$H(X_1, X_2, \cdots X_n)$ Bits $\le 8n$ bits



$H(X_i) = \log_2 256 = 8$ bits

n

given $P(X, Y, Z)$. to know $P(X)$

marganilisation) $P(X) = \sum_{y} \sum_{z} P(X, Y, Z)$

$P(X) P(Y|X) P(Z|Y, X)$

$\sum 1$) ② $P(X)$

~~independence~~ & correllation

for all the moments. just second moment of the distribution