# Recap

Joint Entropy: $H(X,Y) = H(X) + \underbrace{H(Y \mid X)}_{\leq H(Y)}$ → Conditional Entropy

Subadditivity : $H(X_1, \ldots, X_m) = \sum_{i=1}^{m} H(X_i \mid X_1 \ldots X_{i-1})$

$$\leq H(X_1) + \cdots + H(X_m)$$

Shearer's Lemma: Collection $\mathcal{F}$ of subsets of $\{1, \ldots m\}$
Each $i \in [m]$ in at least $t$ subsets

$$t \cdot H(X_1, \ldots, X_m) \leq \sum_{S \in \mathcal{F}} H(X_S)$$

# Mutual Information

$H(Y|X) \leq H(Y)$, but by how much?

$$I(X;Y) = H(Y) - H(Y|X)$$

$$= H(Y) + H(X) - H(Y|X) - H(X)$$

$$= \underbrace{H(Y) + H(X) - H(X,Y)}_{\text{Symmetric in } X \text{ and } Y}$$

$$= H(X) - H(X|Y)$$

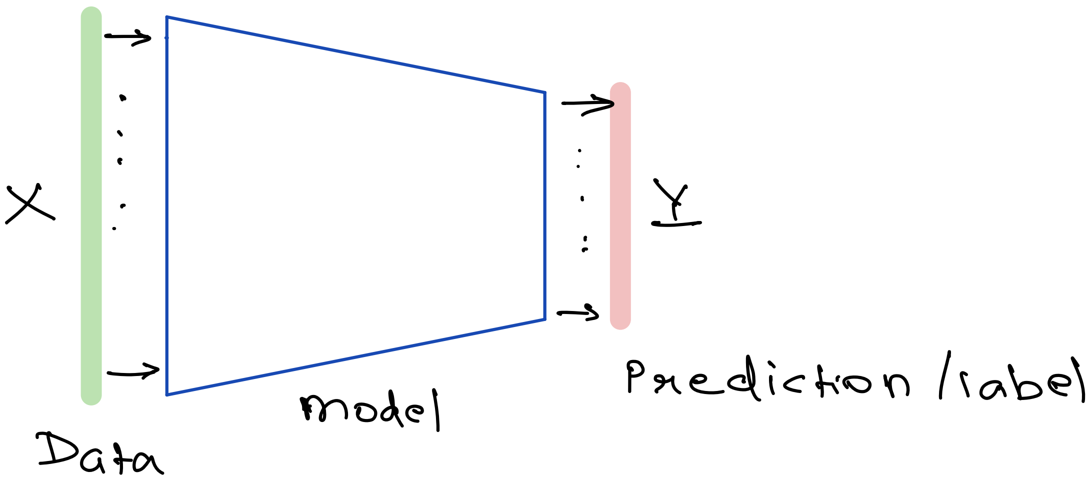
$$I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$$

- $\underbrace{0 \leq I(X;Y) \leq H(X)}_{\text{When?}}$ $\quad (I(X;Y) \leq \min\{H(X), H(Y)\})$

Interesting when choosing $Y$ (with other constraints)

- Error-correcting codes
- Statistical estimator
- Deep learning
- Communication Protocols
- Streaming algorithms

## Eg. Learning representations from data




$$Z = \text{enc}(X)$$

encoding / embedding

$$\text{maximize} \quad I(Z;\underline{Y})$$

(Mutual Information Coding)

$$\text{maximize} \quad I(Z;\underline{Y}) - \beta\, I(Z;X)$$

(Information Bottleneck)

E.g.

$$(X, Y, Z) = \begin{cases} 0\,0\,0 & \text{w.p. } 1/4 \\ 0\,1\,1 & \text{w.p. } 1/4 \\ 1\,0\,1 & \text{w.p. } 1/4 \\ 1\,1\,0 & \text{w.p. } 1/4 \end{cases}$$

$$X + Y + Z \equiv 0 \pmod 2$$

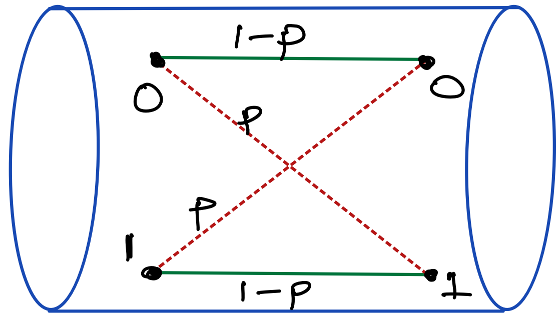
$$I(X;Y) = 0$$

$$I(X;Y|Z) = \underbrace{H(X|Z)}_{=1} - \underbrace{H(X|Y,Z)}_{=0}$$

Conditioning does not necessarily reduce $I(X;Y)$

E.g.

$$\left.\begin{array}{ll} 0 & \text{w.p. } 1/2 \\ 1 & \text{w.p. } 1/2 \end{array}\right\} X \longrightarrow$$




$$\longrightarrow \underline{Y}$$

Binary Symmetric Channel

$I(X ; \underline{Y}) = ?$ $\qquad 1 - H_2(p)$

Chain rule

$$I((X_1, \ldots, X_n); Y) = \sum_{i=1}^{n} I(X_i ; Y \mid X_1 \ldots X_{i-1})$$

$$H(X_1 \ldots X_n) - H(X_1 \ldots X_n \mid Y)$$

$$\sum \underbrace{H(X_i \mid X_1 \ldots X_{i-1}) - H(X_i \mid Y, X_1 \ldots X_{i-1})}$$

$$I(X_i ; Y \mid X_1 \ldots X_{i-1})$$

# Data Processing

- $I(X;Y) \geq I(X;g(Y))$ for any function $g$

**Proof:**

$$H(X) - \boxed{H(X|Y)}$$

$$= H(X) - \underbrace{\boxed{H(X|Y,g(Y))}}_{H(X|Y,g(Y)) \leq H(X|g(Y))}$$

$$\geq H(X) - H(X|g(Y))$$

$$= I(X;g(Y))$$

$$I(Y;Y) \geq I(Y;g(Y))$$

$$\|$$

$$H(Y) \geq H(g(Y)) - \cancel{H(g(Y)|Y)}$$

$$0$$

# Markov Chains

$$X \longleftrightarrow Y \longleftrightarrow Z$$

$X, Z$ independent given $Y$

▷ $I(X;Y) \geq I(X;Z)$

<u>Proof:</u>

$$\left.\begin{array}{l} H(X) - H(X|Y) \\ H(X) - \underbrace{H(X|Y,Z)}_{\leq H(X|Z)} \end{array}\right\} \begin{array}{l} I(X;Z|Y) = 0 \\ \text{by conditional independence} \end{array}$$

$$\geq H(X) - H(X|Z)$$

$$= I(X;Z)$$

# Sufficient Statistics

For random variables $X, Y$:

$g(Y)$ is a sufficient statistic of $Y$ for $X$ if

$$I(X;Y) = I(X;g(Y))$$

(No loss in data processing)

Ex: $X = \begin{cases} p_1 & \text{w. p. } 1/2 \rightarrow Y = (Y_1, \ldots, Y_n) \text{ iid } Y_i = \begin{cases} 1 & \text{wp } p_1 \\ 0 & \text{wp } 1-p_1 \end{cases} \\ p_2 & \text{w. p. } 1/2 \rightarrow Y = (Y_1, \ldots, Y_n) \text{ iid } Y_i = \begin{cases} 1 & \text{wp } p_2 \\ 0 & \text{wp } 1-p_2 \end{cases} \end{cases}$

$g(Y) = Y_1 + \ldots + Y_n = \#$ of 1's

Prove: $I(X;Y) = I(X;g(Y))$

# Fano's inequality

$$\underbrace{X}_{\text{Hidden}} \longrightarrow \underbrace{Y}_{\text{Observed}} \longrightarrow \underbrace{\hat{X}}_{\substack{\text{Predictor}\\ \text{for } X}}$$

$$\mathrm{Supp}(X) = \mathrm{Supp}(\hat{X}) = \mathcal{X}$$

$$p_e = \mathbb{P}(\hat{X} \neq X)$$

▶ $$H_2(p_e) + p_e \cdot \log(|\mathcal{X}| - 1) \geq \underbrace{H(X \mid \hat{X})}_{\to \text{Data Processing}} \geq H(X \mid Y)$$

<u>Proof:</u> $E = \mathbb{1}\{\hat{X} \neq X\}$ $\qquad \mathbb{P}(E=1) = p_e$

$$I(X, E \mid \hat{X}) = H(X \mid \hat{X}) - H(X \mid E, \hat{X})$$

$$I(X, E \mid \hat{X}) = H(E \mid \hat{X}) - \underline{\cancel{H(E \mid X, \hat{X})}} = 0$$

$$\underbrace{H(E \mid \hat{X})}_{\leq H(E)} + \underbrace{H(X \mid E, \hat{X})}_{p_e H(X \mid \hat{X}, E=1) + \cancel{(1-p_e) H(X \mid \hat{X}, E=0)}\ 0} = H(X \mid \hat{X})$$