§ 1.6* Entropy and Mutual Information for DRV

- A length N discrete random vector (DRV) $\underline{X} = (X_1, X_2, \dots, X_N)$ with realizations of $\underline{x} = (x_1, x_2, \dots, x_N)$

- **Entropy of $\underline{X}$:** Let $P(\underline{x}) = \Pr(X_1 = x_1, X_2 = x_2, \dots, X_N = x_N)$ denote the distribution of DRV $\underline{X}$,

$$H(\underline{X}) = \mathbb{E}\left[\log_2 P(\underline{x})^{-1}\right]$$
$$= \sum_{\underline{x}} P(\underline{x}) \log_2 P(\underline{x})^{-1} \text{ bits/vector}$$

§ 1.6* Entropy and Mutual Information for DRV

- Similarly, given $P\left(\underline{x},\underline{y}\right) = \Pr((x_1, x_2, \dots, x_N), (y_1, y_2, \dots, y_N))$ and $P\left(\underline{x}|\underline{y}\right) = \Pr((x_1, x_2, \dots, x_N)|(y_1, y_2, \dots, y_N))$

- **Joint Entropy**: $$H(\underline{X},\underline{Y}) = \mathbb{E}\left[\log_2 P\left(\underline{x},\underline{y}\right)^{-1}\right]$$
$$= \sum_{\underline{x}} \sum_{\underline{y}} P\left(\underline{x},\underline{y}\right) \log_2 P\left(\underline{x},\underline{y}\right)^{-1} \text{ bits/vector}$$

- **Conditional Entropy**: $$H(\underline{X}|\underline{Y}) = \mathbb{E}\left[\log_2 P\left(\underline{x}|\underline{y}\right)^{-1}\right]$$
$$= \sum_{\underline{x}} \sum_{\underline{y}} P\left(\underline{x},\underline{y}\right) \log_2 P\left(\underline{x}|\underline{y}\right)^{-1} \text{ bits/vector}$$

§ 1.6* Entropy and Mutual Information for DRV

- **Mutual Information**

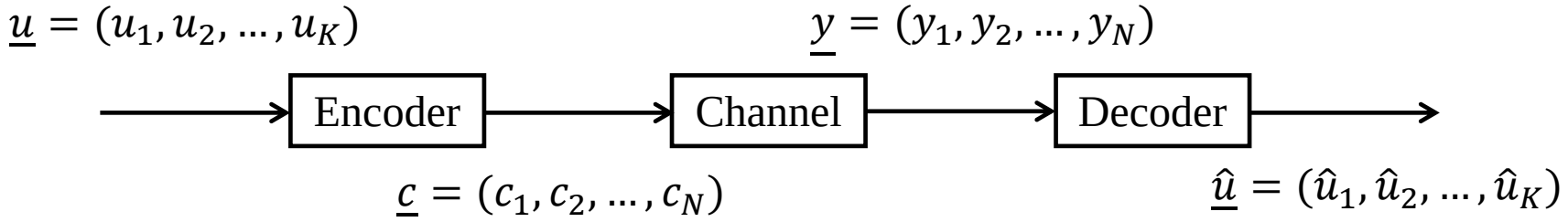
$$I(\underline{X},\underline{Y}) = H(\underline{X}) - H(\underline{X}|\underline{Y})$$

$$= \mathbb{E}\left[\log_2 \frac{P\left(\underline{x}\middle|\underline{y}\right)}{P(\underline{x})}\right]$$

$$= \sum_{\underline{x}}\sum_{\underline{y}} P\left(\underline{x},\underline{y}\right) \log_2 \frac{P\left(\underline{x}\middle|\underline{y}\right)}{P(\underline{x})}$$

§ 1.6* Entropy and Mutual Information for DRV

- **Data Processing Inequality** over a coded communication system



$$I\left(\underline{y}, \underline{u}\right) \geq I\left(\hat{\underline{u}}, \underline{u}\right)$$

Note: The decoded message $\hat{\underline{u}}$ provides less information about the original message $\underline{u}$ than the received vector $\underline{y}$. However, a digital system would need the decoder to estimate $\underline{u}$ through interpreting $\underline{y}$.

§ 1.6* Entropy and Mutual Information for DRV

- **Theorem 1.** If some symbols $X_1, X_2, \ldots, X_N$ of $\underline{X}$ are independent, i.e.,
$$P(\underline{X}) = P(X_1)P(X_2)\cdots P(X_N).$$

$$\boxed{I(\underline{X},\underline{Y}) \geq \sum_{i=1}^{N} I(X_i, Y_i)}$$

Proof:

$$I(\underline{X},\underline{Y}) = \mathbb{E}\left[\log_2 \frac{P\left(\underline{x}\middle|\underline{y}\right)}{P(\underline{x})}\right] = \mathbb{E}\left[\log_2 \frac{P\left(\underline{x}\middle|\underline{y}\right)}{P(x_1)P(x_2)\cdots P(x_N)}\right]$$

$$\sum_{i=1}^{N} I(X_i, Y_i) = \sum_{i=1}^{N} \mathbb{E}\left[\log_2 \frac{P(x_i|y_i)}{P(x_i)}\right] = \mathbb{E}\left[\log_2 \frac{P(x_1|y_1)P(x_2|y_2)\cdots P(x_N|y_N)}{P(x_1)P(x_2)\cdots P(x_N)}\right]$$

§ 1.6* Entropy and Mutual Information for DRV

$$\sum_{i=1}^{N} I(X_i, Y_i) - I(\underline{X}, \underline{Y})$$

$$= \mathbb{E}\left[\log_2 \frac{P(x_1|y_1)P(x_2|y_2)\dots P(x_N|y_N)}{P\left(\underline{x}\middle|\underline{y}\right)}\right]$$

(Applying Jensen's inequality to the above eq.)

$$\leq \log_2\left(\mathbb{E}\left[\frac{P(x_1|y_1)P(x_2|y_2)\dots P(x_N|y_N)}{P\left(\underline{x}\middle|\underline{y}\right)}\right]\right)$$

$$= \log_2\left(\sum_{\underline{x}}\sum_{\underline{y}} P(x_1|y_1)P(x_2|y_2)\dots P(x_N|y_N)P(\underline{y})\right)$$

$$= \log_2\left(\sum_{\underline{x}} P(x_1)P(x_2)\dots P(x_N)\right)$$

$$= 0$$

□

§ 1.6* Entropy and Mutual Information for DRV

- In a communication system,

Theorem 1 tells if $X_1, X_2, \ldots, X_N$ are independent, $\underline{Y}$ tells more information about $\underline{X}$ than the sum of each Y_i about X_i.

§ 1.6* Entropy and Mutual Information for DRV

- **Theorem 2.** If the channel is described as memoryless, i.e., $P\left(\underline{y}\middle|\underline{x}\right) = \prod_{i=1}^{N} P(y_i|x_i)$, we have

$$\boxed{I(\underline{X},\underline{Y}) \leq \sum_{i=1}^{N} I(X_i, Y_i)}$$

Proof:

$$I(\underline{X},\underline{Y}) = \mathbb{E}\left[\log_2 \frac{P\left(\underline{y}\middle|\underline{x}\right)}{P\left(\underline{y}\right)}\right] = \mathbb{E}\left[\log_2 \frac{P(y_1|x_1)P(y_2|x_2)\dots P(y_N|x_N)}{P\left(\underline{y}\right)}\right]$$

$$\sum_{i=1}^{N} I(X_i, Y_i) = \sum_{i=1}^{N} \mathbb{E}\left[\log_2 \frac{P(y_i|x_i)}{P(y_i)}\right] = \mathbb{E}\left[\log_2 \frac{P(y_1|x_1)P(y_2|x_2)\dots P(y_N|x_N)}{P(y_1)P(y_2)\dots P(y_N)}\right]$$

§ 1.6* Entropy and Mutual Information for DRV

$$I(\underline{X},\underline{Y}) - \sum_{i=1}^{N} I(X_i, Y_i) = \mathbb{E}\left[\log_2 \frac{P(y_1)P(y_2)\dots P(y_N)}{P\left(\underline{y}\right)}\right]$$

(Applying Jensen's inequality to the above eq.)

$$\leq \log_2\left(\mathbb{E}\left[\frac{P(y_1)P(y_2)\dots P(y_N)}{P\left(\underline{y}\right)}\right]\right)$$

$$= \log_2\left(\sum_{\underline{x}}\sum_{\underline{y}} P\left(\underline{x}\middle|\underline{y}\right) P(y_1)P(y_2)\dots P(y_N)\right)$$

$$= \log_2\left(\sum_{\underline{x}} P(\underline{x})\right)$$

$$= 0$$

□

§ 1.6* Entropy and Mutual Information for DRV

- In a communication system,

Theorem 2 tells if channel is discrete memoryless channel, $\underline{Y}$ tells less information about $\underline{X}$ than the sum of each Y_i about X_i.

§ 1.6* Entropy and Mutual Information for DRV

- **Corollary 5:**

If channel is discrete & memoryless and source is independent, we have

$$I(\underline{X}, \underline{Y}) = \sum_{i=1}^{N} I(X_i, Y_i)$$

This property will be used to prove the Shannon's Channel Coding Theorem.

References:

[1] Elements of Information Theory, by T. Cover and J. Thomas.

[2] Scriptum for the lectures, Applied Information Theory, by M. Bossert.