# Information and Entropy

- Finite-alphabet random variable
- Information and entropy
- Joint and conditional entropy
- Mutual information
- Relative entropy
- Data processing inequality
- Asymptotic equipartition property
- Typical sets

# Entropy of a Random Variable

- Consider a discrete, finite-alphabet random variable *X*

$$\mathcal{A}_X = \{\alpha_0, \alpha_1, \ldots, \alpha_{K-1}\}$$

$$f_X(x) = P(X = x) \quad \forall x \in \mathcal{A}_X$$

- **"Information"** associated with the event *X*=*x*

$$h_X(x) = -\log_2 f_X(x)$$

- **"Entropy of *X*"** is the expected value of that information

$$H(X) = E\{h_X(X)\} = -\sum_{x \in \mathcal{A}_X} f_X(x) \log_2 f_X(x)$$

- Unit: bits

# Information and Entropy: Properties

- Information $h_X(x) \geq 0$
- Information $h_X(x)$ strictly increases with decreasing probability $f_X(x)$
- Boundedness of entropy

$$0 \leq H(X) \leq \log_2(|\mathcal{A}_X|)$$

equality if only one outcome can occur

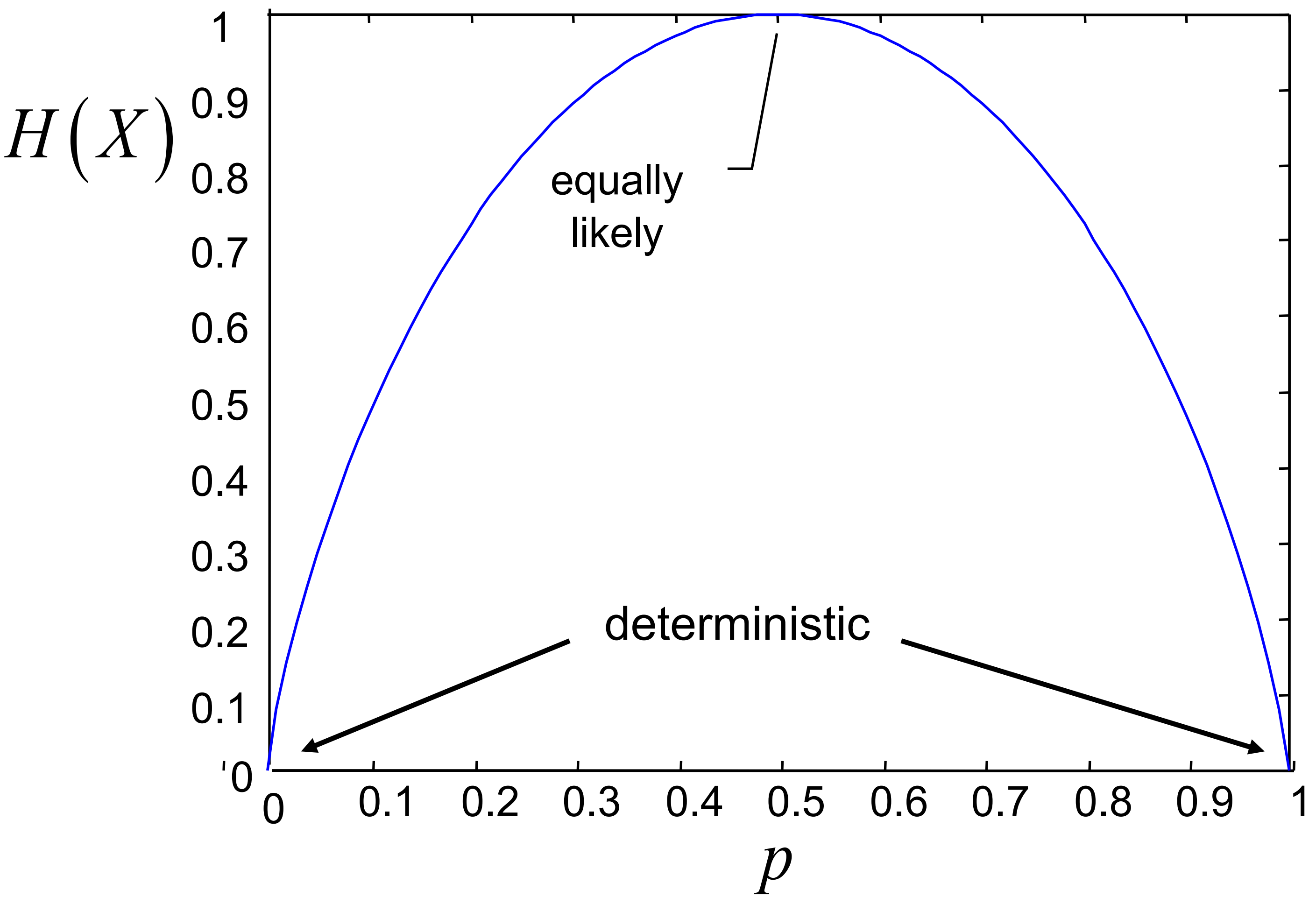
equality if all outcomes are equally likely

- Very likely and very unlikely events do not substantially change entropy

$$-p\log_2 p \to 0 \quad \text{for } p \to 0 \text{ or } p \to 1$$

# Example: Binary Random Variable

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



# Joint Entropy

- Consider random vectors (with discrete, finite-alphabet components)

$$\mathbf{X} = \left( X_0, X_1, \ldots, X_{m-1} \right)$$

- Entropy

$$\boxed{H(\mathbf{X}) = E\left[ -\log_2 f_{\mathbf{X}}\left(\mathbf{X}\right) \right] = E\left[ h_{\mathbf{X}}\left(\mathbf{X}\right) \right]}$$

- Long-hand

$$H(\mathbf{X}) = H(X_0, X_1, \ldots, X_{m-1})$$

$$= -\sum_{x_0} \sum_{x_1} \cdots \sum_{x_{m-1}} f_{\mathbf{X}}\left(x_0, x_1, \ldots, x_{m-1}\right) \log_2 f_{\mathbf{X}}\left(x_0, x_1, \ldots, x_{m-1}\right)$$

# Conditional Entropy

- Consider two discrete finite-alphabet r.v. *X* and *Y*

$$\boxed{\begin{aligned} H(X|Y) &= E\left[-\log_2 f_{X|Y}(x,y)\right] = -\sum_y \sum_x f_{X,Y}(x,y) \log_2 f_{X|Y}(x,y) \\ &= -\sum_y f_Y(y) \sum_x f_{X|Y}(x,y) \log_2 f_{X|Y}(x,y) \end{aligned}}$$

- Conditional entropy *H(X|Y)* is average additional information, if *Y* is already known
- Joint entropy: $H(X,Y) = E\left[-\log_2 f_{X,Y}(X,Y)\right]$

$$\begin{aligned} &= E\left[-\log_2 \left(f_Y(y) f_{X|Y}(X,Y)\right)\right] \\ &= E\left[-\log_2 f_Y(y)\right] + E\left[-\log_2 f_{X|Y}(X,Y)\right] \\ &= H(Y) + H(X|Y) \end{aligned}$$

(Chain rule)

# Mutual Information

- "Mutual information" is the average information that random variables *X* and *Y* convey about each other
  - Reduction in uncertainty about $x$, if $y$ is observed
  - Reduction in uncertainty about $y$, if $x$ is observed

$$\boxed{\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) = H(Y) - H(Y \mid X) \\ &= \sum_x \sum_y f_{X,Y}(x,y) \log_2 \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \end{aligned}}$$

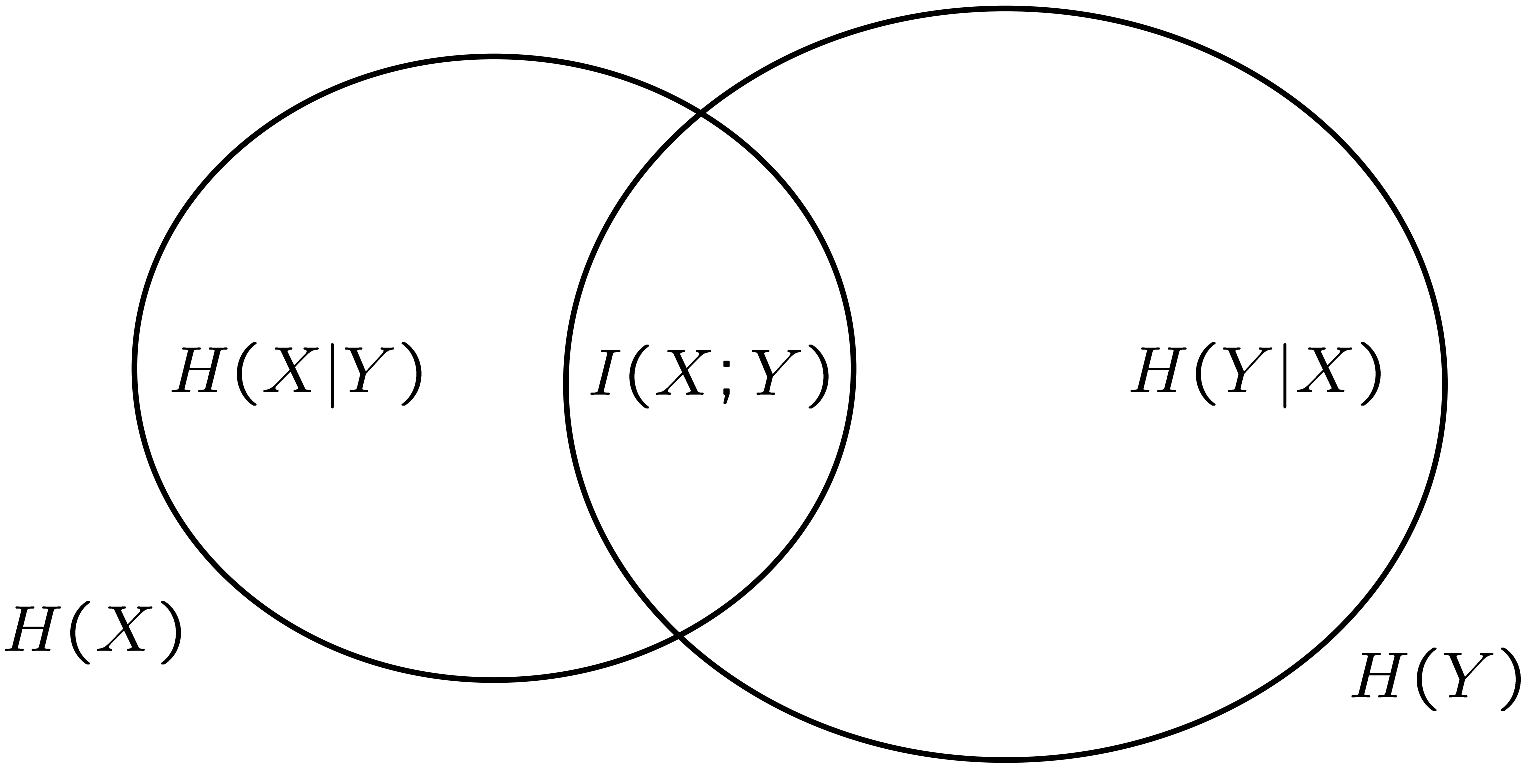
- Properties

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

$$I(X;Y) \le H(X)$$

$$I(X;Y) \le H(Y)$$

# Mutual Information




$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

# Conditioning Reduces Entropy

- Property of mutual information

$$0 \leq I(X;Y) = H(X) - H(X|Y)$$

- Knowing another random variable Y can only reduce the uncertainty in X

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

- Equality if and only if X and Y are independent

# Chain Rule for Entropy

$$H(X_1, X_2) = H(X_1) + H(X_2|X_1)$$

$$\begin{aligned} H(X_1, X_2, X_3) &= H(X_1) + H(X_2, X_3|X_1) \\ &= H(X_1) + H(X_2|X_1) + H(X_3|X_2, X_1) \\ &\vdots \end{aligned}$$

$$H(X_1, X_2, \ldots, X_n) = \sum_{i=1}^{n} H(X_i|X_{i-1}, \ldots, X_1)$$

# Joint Entropy and Statistical Dependency

- Theorem (Independence bound on entropy)

$$\boxed{H(X_0, X_1, X_2, \ldots, X_{m-1}) \leq H(X_0) + H(X_1) + \ldots + H(X_{m-1})}$$

Equality for statistical independence of $X_0, X_1, X_2, \ldots, X_{m-1}$

- Exploiting statistical dependencies can reduce bit-rate
- Statistically independent components can be compressed and decompressed separately without loss

# Conditional Mutual Information

- Mutual information between X and Y given Z

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

- Chain rule for information

$$I(X_1, X_2, \ldots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, \ldots, X_1)$$

# Relative Entropy

- Kullback Leibler distance between two PMFs p(x) and q(x)

$$H(p \parallel q) = \sum_x p(x) \log_2 \frac{p(x)}{q(x)}$$

- In general

$$H(p \parallel q) \neq H(q \parallel p)$$

- Information inequality

$$H(p \parallel q) \geq 0$$

- With equality if and only if p(x)=q(x) for all x.

# Relative Entropy and Mutual Information

- Mutual information is the relative entropy between the joint distribution and the product distribution

$$I(X;Y) = H(f_{X,Y} \parallel f_X f_Y)$$

- Non-negativity of mutual information

$$I(X;Y) \geq 0$$

- With equality if and only if X and Y are independent.

# Data Processing Inequality

- No clever manipulation of the data can improve the inferences that can be made from the data
- X, Y, Z are said to **form a Markov chain in that order**

$$X \to Y \to Z$$

- if the joint PMF can be written as

$$f_{X,Y,Z} = f_X f_{Y|X} f_{Z|Y}$$

- Note, X and Z are conditionally independent given Y

$$f_{X,Z|Y} = f_{X|Y} f_{Z|Y}$$

# Data Processing Inequality

- Chain rule for mutual information

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned}$$

- If $X \to Y \to Z$, X and Z are conditionally independent given Y

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

- Non-negativity of mutual information: $I(X;Y|Z) \geq 0$

- Data processing inequality: If $X \to Y \to Z$

$$I(X;Y) \geq I(X;Z)$$

# Remark: Weak Law of Large Numbers

- **Law of Large Numbers:** The average of the results obtained from a large number of trials is close to the expected value

$$\frac{1}{n}(U_1 + U_2 + \cdots + U_n) \xrightarrow{p} E[U] \quad \text{for} \quad n \to \infty$$

- Convergence in probability (weak law)

$$\lim_{n\to\infty} \Pr\left( \left| \frac{1}{n} \sum_{i=1}^{n} U_i - E[U] \right| < \epsilon \right) = 1 \quad \text{for} \quad \epsilon > 0$$

# Asymptotic Equipartition Property

- Let $X_1$, $X_2$, … be i.i.d. with $f_X$

$$
\begin{aligned}
-\frac{1}{n}\log_2 f_{X_1,X_2,\dots,X_n} &= -\frac{1}{n}\sum_{i=1}^{n}\log_2 f_{X_i} \\
&\xrightarrow{p} E[-\log_2 f_X] \\
&= H(X)
\end{aligned}
$$

- Asymptotic Equipartition Property (AEP): If $X_1$, $X_2$, … are i.i.d. with $f_X$, then

$$
-\frac{1}{n}\log_2 f_{X_1,X_2,\dots,X_n} \xrightarrow{p} H(X)
$$

# Typical Set

- The typical set $A_\varepsilon^{(n)}$ with respect to $f_X$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property

$$2^{-n[H(X)+\epsilon]} \leq f_{X_1,X_2,\ldots,X_n} \leq 2^{-n[H(X)-\epsilon]}$$

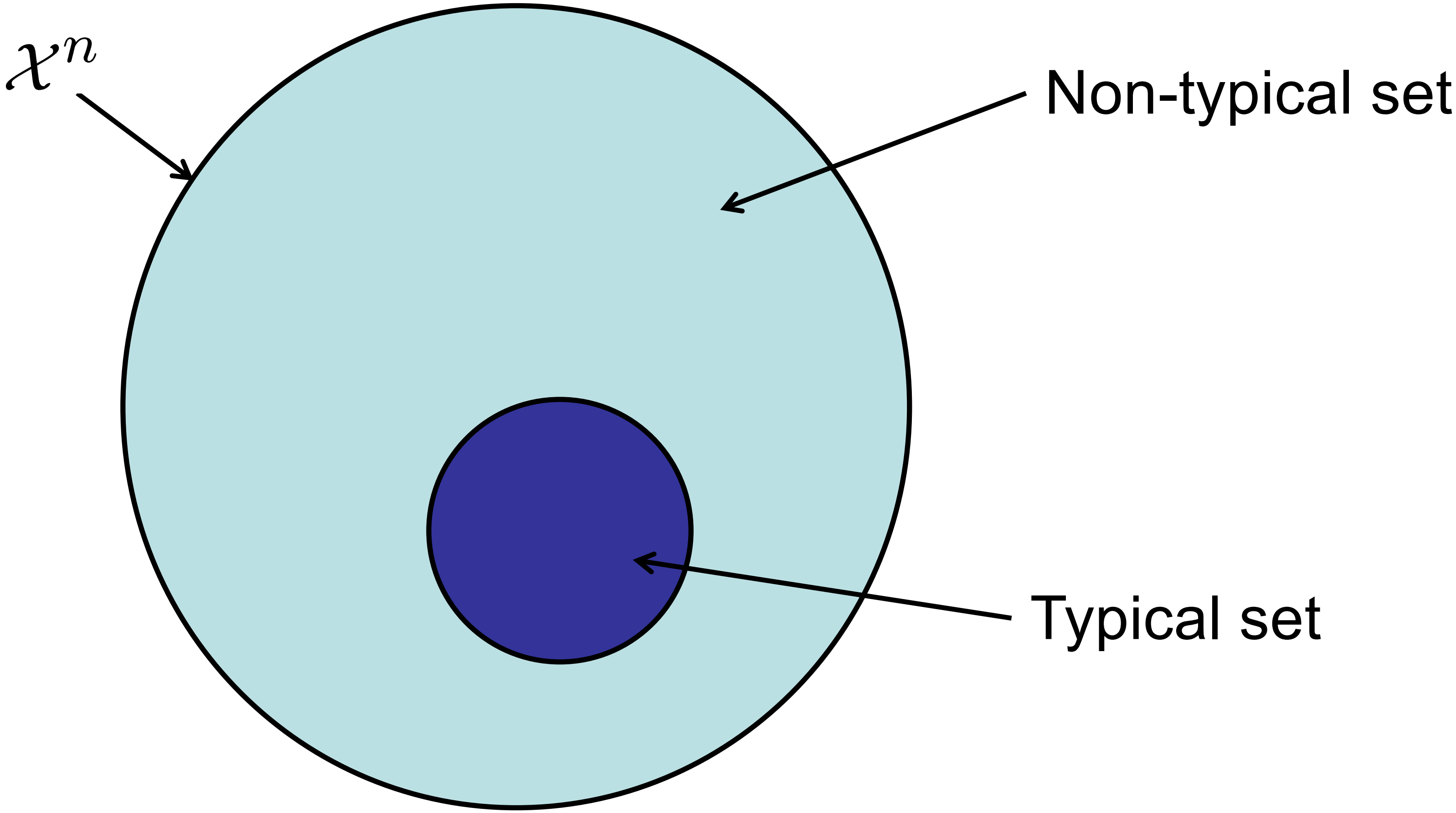
- The typical set has probability nearly 1

$$\Pr\left\{A_\epsilon^{(n)}\right\} > 1 - \epsilon$$

- All elements of the typical set are nearly equiprobable
- The number of elements in the typical set is nearly $2^{nH}$

$$\left|A_\epsilon^{(n)}\right| \leq 2^{n[H(X)+\epsilon]}$$

# Typical Sets and Source Coding




Represent sequences in $\mathcal{X}^n$ using nH(X) bits on average.