- **Self-information:** a symbol x_i from a random variable X

$$I(x_i) = -log(x_i)$$

- **Entropy:** a discrete random variable (D.R.V) X

$$H(X) = -\sum_{i=0}^{M} log(x_i)$$

  - If X is a fixed random variable, then $H(X) = 0$.
  - $H(p)$ is a concave function of X.

- **Joint entropy**: a pair of D.R.V. (X, Y) with joint distribution $p(x, y)$

$$H(X,Y) = -\sum_x \sum_y p(x,y) log\, p(x,y)$$

- **Conditional entropy:** If $(X, Y) \sim p(x, y)$

$$H(Y|X) = -\sum_x \sum_y p(x,y) log\, p(y|x)$$

  - $H(Y|X) \neq H(X|Y)$

- **Chain rule of entropy:**

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

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

- **Relative entropy:** between two probability mass function (pmf) $p(x)$ and $q(x)$

$$D(p \,||\, q) = \sum_x p(x) log \frac{p(x)}{q(x)}$$

- **Chain rule of the relative entropy:**

$$D(p(x,y) \,||\, q(x,y)) = D(p(x) || q(x)) + D(p(x|y) || q(x|y))$$

  - $D(q||q) = 0$

- **Mutual information:** let $p(x, y)$ be the the joint pmf of X and Y

$$I(X;Y) = \sum_x \sum_y p(x,y) log \frac{p(x,y)}{p(x)p(y)}$$

  - $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(X|Y) = H(X) + H(Y) - H(X, Y)$
  - Symmetric: $I(X; Y) = I(Y; X)$
  - Identity: $I(X; X) = H(X)$

- **Conditional mutual information:** (X and Y) given Z

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

- **Convex function:** a function called convex over an interval (a, b) if there exists x_1, $x_2 \in (a, b)$ and $1 \leq \alpha \leq 0$, we have

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2)$$

- **Jensen's inequality**: if f is a convex function and X is a random variable,

$$Ef(X) \geq f(EX)$$

- **Information inequality:** let $p(x)$, $q(x)$ be two pmf's of $x \in \Omega x$.

$$D(p||q) \geq 0, with\ equlaity\ if\ and\ only\ if\ p(x) = q(x)\ for\ all\ x$$

- **Uniform pmf maximizes entropy:**

$$H(X) \leq log(the\ number\ of\ element\ in\ the\ range\ of\ X)$$

  - With equality if and only if X has a uniform distribution over elements in X.

- **The more information, the better:**

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

  - Observation of another random variable Y can reduce uncertainty in X.

- **Efficiency of joint coding of sources:** let the distribution of X, X_2, …, X_n be $p(x_1, x_2, \ldots, x_n)$, then

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

- **Markov chain of random variables:** random variables X, Y, Z are said to form a Markov chain, $X \to Y \to Z$, if the conditional distribution of Z depends only on Y, and is conditionally independent of X.
  - $X \to Y \to Z$ if and only if $p(x, y, z) = p(x) * p(x|y) * p(z|y)$
  - $X \to Y \to Z$ if and only if $p(x, \underline{z}|y) = p(x|y) * p(z|y)$

- **Data processing inequality:** if $X \to Y \to Z$, then

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

- **Fano's inequality:** for any estimator X' s.t. $X \rightarrow Y \rightarrow X'$, with $P_e = Pr\{X \neq X'\}$,

$$H(P_e) + P_e\, log(elements\ in\ X) \geq H(X|X') \geq H(X|Y)$$

   - Weak Fano's inequality: $1 + P_e * \log(\Omega x) \geq H(X|Y)$.
   - Strong Fano's inequality: $1 + P_e * \log(\Omega x - 1) \geq H(X|Y)$.

- **Law of large numbers:**
   - **Weak law:** if $X_1, X_2, ...$ are i.i.d. $\sim p(x)$ with mean μ, then there exists $\varepsilon > 0$

$$\lim_{n\to\infty} p\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| < \varepsilon\right) = 1$$

   - **Strong law:** if $X, X_2, ...$ are i.i.d. $\sim p(x)$ with mean μ,

$$p\left(\lim_{n\to\infty}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right) = \mu\right) = 1$$

- **Asymptotic equipartition property:** if $X_1, X_2, ...$ re i.i.d. $\sim p(x)$, then

$$-\frac{1}{n}\log p(X_1, X_2, \dots, X_n) \rightarrow H(X)$$

- **Typical set:** the typical set with $A_\varepsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \dots, x_n) \in \Omega_x^n$ with the property

$$2^{-n(H(x)+\varepsilon)} \leq p(x1, x2, \dots, xn) \leq 2^{n(H(x)-\varepsilon)}$$

   - The typical set has probability nearly 1.
   - All elements in the set are nearly euqiprobable.
   - The number of elements in the typical set is nearly 2^{nH}.
   - $|A_\varepsilon^{(n)}| \leq 2^{n(H(x)+\varepsilon)}$
   - $|A_\varepsilon^{(n)}| \geq (1-\varepsilon) * 2^{n(H(x)-\varepsilon)}$

- **High probability set:** let $X_1, X_2, \dots, X_n$ be i.i.d. $\sim p(x)$, $B_\delta^{(n)}$ be the smallest set with $Pr\{B_\delta^{(n)}\} \geq 1 - \delta$. For $\delta < 1/2$ and any $\delta' > 0$ if $Pr\{B_\delta^{(n)}\} \geq 1 - \delta$, then

$$\frac{1}{n}\log\left(B_\delta^{(n)}\right) > H - \delta'$$

- **Stationary process:** for every n and shift t, and for all $x_1, x_2, \dots, x_n \in \Omega x$

$$Pr\{X_1 = x_1, X_2 = x_2, \dots, X_n = x_n\} =$$
$$Pr\{X_{1+t} = x_{1+t}, X_{2+t} = x_{2+t}, \dots, X_{n+t} = x_{n+t}\}$$

- **Markov chain:** a discrete stochastic process $X_1, X_2, .., X_n$ is said to be a Markov chain or a Markov process if for $n = 1, 2, \dots$,n and for all $x_1, x_2, \dots, x_n, x_{n+1} \in \Omega x$

$$Pr\{X_{n+1} = x_{n+1} \mid X_1 = x_1, \dots, X_n = x_n\} = Pr\{X_{n+1} = x_{n+1} \mid X_n = x_n\}$$

- **Time invariance:** the Markov chain is said to be time invariant if the conditional probability $p(x_{n+1} \mid x_n)$ does not depend on n. For n = 1, 2, … and for all a, $b \in \Omega x$

$$Pr\{X_{n+1} = b \mid X_n = a\} = Pr\{X_2 = b \mid x_n = a\}$$

- **State probability:** if the pmf of state at time n is $p(x)$, the pmf at time n+1 is

$$p(x_{n+1}) = \sum_{x_n} p(x_n)\, p(X_{n+1} = x_{n+1} \mid X_n = x_n)$$

- **Entropy rate:** a stochastic process $\{X_i\}$ is defined by

$$\mathrm{H(X)} = \lim_{n\to\infty} \frac{1}{n} H(\mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_n)$$

  - $\mathrm{H'(X)} = \lim_{n\to\infty} H(\mathrm{X}_n \mid \mathrm{X}_1, \mathrm{X}_2, \dots, \mathrm{X}_{n-1})$
  - For a stationary stochastic process, $H(X) = H'(X)$
  - For a stationary Markov chain, $H(X) = H(X_2 \mid X_1) = -\sum_{ij} \mu_i P_{ij} \log(P_{ij})$