Information Theory

Entropy and Mutual Information

Entropy

$$H(X) = \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = E_p \left[\log \frac{1}{p(x)} \right]$$

$$(X \text{ is constant}) \ 0 \leq H(X) \leq \log |\mathcal{X}| \ (X \text{ is uniform})$$

joint entropy

$$H(X, Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x, y)}$$

conditional entropy

$$H(X \mid Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{1}{p(x \mid y)}$$

Relative Entropy (K-L Distance)

$$D(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \geq 0$$

Mutual Information

$$I(X; Y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$
$$= D(p(x, y) \parallel p(x)p(y))$$

Conditional Mutual Information

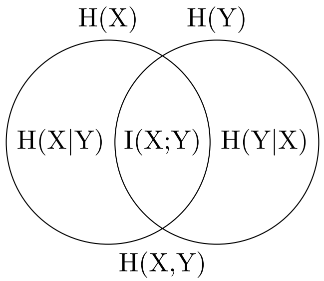
$$I(X; Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z)$$

Chain Rule

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

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

Venn Diagram



Inequalities

$f(x)$ is convex over (a, b) if for $x_1, x_2 \in (a, b)$ and $\lambda \in [0, 1]$,

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

Jensen's Inequality for convex f and random variable X,

$$E(f(X)) \geq f(E(X))$$

Information Can't Hurt

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

Log Sum Inequality for nonnegative $a_{1,\ldots,n}$ and $b_{1,\ldots,n}$

$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i}$$

with equality iff $\frac{a_i}{b_i}$ is constant

$D(p \parallel q)$ is convex in the pair (p, q)

$H(X)$ is concave of p

$I(X; Y)$ is concave of $p(x)$ for fixed $p(y \mid x)$

$I(X; Y)$ is convex of $p(y \mid x)$ for fixed $p(x)$

Markov Chain $X \to Y \to Z$ if $p(x, y, z) = p(x)p(y|x)p(z|y)$

Data-processing Inequality if $X \to Y \to Z$ forms a Markov Chain, then $I(X; Y) \geq I(X; Z)$

$T(X)$ is a sufficient statistic if $I(\theta; X) = I(\theta; T(X))$

Fano's Inequality

For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$

$$H(X \mid Y) \leq H(X \mid \hat{X}) \leq H(P_e) + P_e \log |\mathcal{X}|$$

Data Compression

Code

source code C for X is a mapping from $\mathcal{X}$ to the set of finite-length strings from a D-ary alphabet $\mathcal{D}^*$

expected length $L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$

instantaneous(prefix) $\Rightarrow$ uniquely decodable $\Rightarrow$ nonsingular

Kraft Inequality

For any instantaneous code over a D-ary alphabet,

$$\sum_{x \in \mathcal{X}} D^{-l(x)} \leq 1$$

(any uniquely decodable D-ary code also satisfies this)

Boundary on Optimal Code Length

$$H_D(X) \leq L^* < H_D(X) + 1$$

Wrong Code

for code assignment $l(x) = \left\lceil \log \frac{1}{q(x)} \right\rceil$ under real pmf $p(x)$,

$$H(p) + D(p \parallel q) \leq E_p(l(x)) < H(p) + D(p \parallel q) + 1$$

Asymptotic Equipartition Property

AEP

If $X_1, \ldots, X_n \sim p(x)$ are i.i.d,

$$-\frac{1}{n} \log p(X_1, \ldots, X_n) \xrightarrow{P} H(X)$$

Typical Set

$A_\epsilon^{(n)}$ is the set of sequences $(x_1, \ldots, x_n) \in \mathcal{X}^n$ where

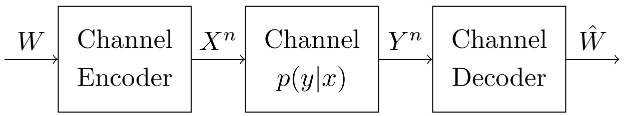
$$2^{-n(H(X)+\epsilon)} \leq p(x_1, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}$$

Consequences of AEP

- If $(x_1, \dots, x_n) \in A_\epsilon^{(n)}$, then
$$H(X) - \epsilon \le -\frac{1}{n} \log p(x_1, \dots, x_n) \le H(X) + \epsilon$$
- $\left| A_\epsilon^{(n)} \right| \le 2^{n(H(X)+\epsilon)}$
- For n sufficiently large, $Pr\left\{A_\epsilon^{(n)}\right\} > 1 - \epsilon$ and
$$\left| A_\epsilon^{(n)} \right| \ge (1-\epsilon) 2^{n(H(X)-\epsilon)}$$

Channel Capacity

Communication System



Discrete Memoryless Channel

A discrete channel is denoted by $(\mathcal{X}, p(y \mid x), \mathcal{Y})$
A discrete memoryless channel is a channel that satisfies
$$p(y_k \mid x^k, y^{k-1}) = p(y_k \mid x_k)$$
If a channel is used without feedback,
$$p(x_k \mid x^{k-1}, y^{k-1}) = p(x_k \mid x^{k-1})$$
Then for a DMC (without feedback by default),
$$p(y^n \mid x^n) = \prod_{i=1}^{n} p(y_i \mid x_i)$$

Channel Capacity

$$C = \max_{p(x)} I(X;Y)$$
For a weakly symmetric channel, i.e. the rows of the transition matrix $p(y \mid x)$ are permutations of each other,
$$C = \log |\mathcal{Y}| - H(\text{row of transition matrix})$$
achieved when X is uniform

Jointly Typical Sequences

$$A_\epsilon^{(n)} = \left\{ (x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \right.$$
$$\left. \left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \left| -\frac{1}{n} \log p(x^n, y^n) - H(X,Y) \right| < \epsilon \right\}$$
$$\text{where } p(y^n \mid x^n) = \prod_{i=1}^{n} p(y_i \mid x_i)$$

Joint AEP

If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$, then

- $Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)} \right) \le 2^{-n(I(X;Y)-3\epsilon)}$
- For n sufficiently large,
$$Pr\left((\tilde{X}^n, \tilde{Y}^n) \in A_\epsilon^{(n)} \right) \ge (1-\epsilon) 2^{-n(I(X;Y)+3\epsilon)}$$

Channel Coding Theorem

an (M, n) code: $X^n : \{1, \dots, M\} \xrightarrow{encode} \mathcal{X}^n$
$g : \mathcal{Y}^n \xrightarrow{decode} \{1, \dots, M\}$
probability of error $\lambda_i = Pr(g(Y^n) \ne i \mid X^n = x^n(i))$
maximal probability of error $\lambda^{(n)} = \max_{i \in \{1, \dots, M\}} \lambda_i$
average probability of error $P_e^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i$
rate $R = \frac{\log M}{n}$
rate R is achievable if $\exists$ a sequence of $(\lceil 2^{nR} \rceil, n)$ codes s.t.$\lambda^{(n)} \to 0$ as $n \to \infty$
rate R is achievable $\Leftrightarrow R \le C$

Capacity of Parallel Channels

$$C = \log\left(2^{C_1} + 2^{C_2}\right)$$

Differential Entropy

Differential Entropy

$$h(X) = -\int_S f(x) \log f(x) dx$$

joint entropy
$$h(X_1, \dots, X_n) = -\int f(x^n) \log f(x^n) dx^n$$
conditional entropy
$$h(X \mid Y) = -\int f(x,y) \log f(x \mid y) dx dy$$

AEP for Continuous Random Variables

If $X_1, \dots, X_n \sim f(x)$ are i.i.d,
$$-\frac{1}{n} \log f(X_1, \dots, X_n) \xrightarrow{P} h(X)$$

Typical Set for Cont. Random Variables

$A_\epsilon^{(n)} = \left\{ (x_1, \dots, x_n) : \left| -\frac{1}{n} \log f(x_1, \dots, x_n) - h(X) \right| \le \epsilon \right\}$

- $\text{Vol}\left(A_\epsilon^{(n)}\right) = \int_{A_\epsilon^{(n)}} dx_1 \cdots dx_n \le 2^{n(h(X)+\epsilon)}$
- For n sufficiently large, $Pr\left(A_\epsilon^{(n)}\right) > 1 - \epsilon$ and
$$\text{Vol}\left(A_\epsilon^{(n)}\right) \ge (1-\epsilon) 2^{n(h(X)-\epsilon)}$$

Entropy of Normal Distribution

$$h(\mathcal{N}(\mu, \sigma^2)) = \frac{1}{2} \log 2\pi e \sigma^2$$
$h(\mathcal{N}_n(\mu, K)) = \frac{1}{2} \log(2\pi e)^n |K|$ (K is the covariance matrix)

Relative Entropy and Mutual Information

relative entropy $D(f \parallel g) = \int f \log \frac{f}{g}$
mutual information $I(X;Y) = \int f(x,y) \log \frac{f(x,y))}{f(x)f(y)} dx dy$

Properties of Differential Entropy

$I(X;Y) = D(f(x,y) \parallel f(x)f(y)) \ge 0$
$h(X \mid Y) \le h(X)$ equality iff X, Y are independent
$h(X_1, \dots, X_n) \le \sum h(X_i)$ equality iff X_i are independent
$h(X + c) = h(X) \quad h(aX) = h(X) + \log |a|$

Gaussian Channel

Gaussian Channel with Power Constraint

$$Y_i = X_i + Z_i,\ Z_i \sim \mathcal{N}(0, N),\ \frac{1}{n}\sum_{i=1}^{n} x_i^2 \leq P$$

when $X \sim \mathcal{N}(0, P)$, maximum capacity is achieved,

$$C = \max_{E(X^2)\leq P} I(X;Y) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

Parallel Gaussian Channels

For k independent parallel Gaussian channels,

$$C = \sum_{j=1}^{k} \frac{1}{2}\log\left(1 + \frac{P_j}{N_j}\right)$$

power is allotted by water-filling, i.e. $P_i = (v - N_i)^+$, where v is chosen such that $\sum_{i=1}^{k} P_i = P$