1 Information Measures

1.1 Entropy Basics

Axiomatic view of Information: **Non-negative**, **Zero for definite events**, **Monotonicity** (i.e. if $p \le p'$ then $\psi(p) \ge \psi(p')$, the less likely the event is, more information is learnt), **Continuity**, **Additive** under independence (i.e. $\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$).

Axiomatic view of Entropy: $\psi(\mathbf{p})$ is continuous, increasing in uniform case (i.e. uniform over larger set means more uncertain), and

$$\psi(p_1, \ldots, p_n) = \psi(p_1 + p_2, \ldots, p_n) + (p_1 + p_2)\psi\left(\frac{p_1}{p_1 + p_2} + \frac{p_2}{p_1 + p_2}\right)$$

Joint Entropy: Entropy of random vector (X, Y).

Conditional Entropy:

$$H(Y|X) = E_{P_{XY}}\left[\frac{1}{\log P_{Y|X}(Y|X)}\right] = \sum_x P_X(x) H(Y|X = x)$$

Entropy Properties:

- **Bounds**: $0 \le H(X) \le \log|\mathcal{X}|$
- **Differentiation**: $H'(p) = \log\frac{1-p}{p}$
- **Chain Rule**:

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

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

- **Conditioning reduces entropy**: $H(X|Y) \le H(X)$, with equality iff X and Y are independent.
- **Sub-additivity**: $H(X_1, \ldots, X_n) \le \sum H(X_i)$, with equality iff X_i are independent.
- **Applying Function**: For deterministic f, $H(f(X)) \le H(X)$, with equality iff f is one-one.

1.2 Relative Entropy

Relative Entropy or Kullback-Leibler Divergence: Measure of inefficiency of assuming distribution q, when truly it is p.

$$D(p \parallel q) = \sum p(x) \log\frac{p(x)}{q(x)} = E_p\left[\log\frac{p(X)}{q(X)}\right]$$

$$D(p(y|x) \parallel q(y|x)) = E_p\left[\log\frac{p(Y|X)}{q(Y|X)}\right]$$

- **Gibb's Inequality**: $D(p \parallel q) \ge 0$, with equality iff $p = q$.
- **Chain Rule**: $D(p(x,y) \parallel q(x,y)) = D(p(x) \parallel q(x)) + D(p(y|x) \parallel q(y|x))$.

1.3 Mutual Information

Mutual Information: Relative entropy between join distribution and product distribution:

$$I(X;Y) = \sum_{(x,y)} p(x,y) \log\frac{p(x,y)}{p(x)p(y)} = E_{P_{XY}}\left[\log\frac{p(X,Y)}{p(X)p(Y)}\right]$$

- **Expansion**:

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

- $I(X;X) = H(X)$
- $I(X;Y) = 0$ iff X and Y are independent.
- **Chain Rule**:

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

- **Data Processing Inequality**: If Z depends on (X,Y) only via Y, then $X \to Y \to Z$, and $I(X;Z) \le I(X;Y)$, with equality iff $I(X;Y|Z) = 0$. Processing cannot increase information available.
  - X and Z conditionally independent given Y.
  - $I(X;Y|Z) < I(X;Y)$ if $X \to Y \to Z$.
  - $P_{X|YZ} = P_{X|Y}$ and $P_{Z|XY} = P_{Z|Y}$.
- **Partial Sub-additivity**: If $Y_1, \ldots, Y_n$ are conditionally independent given $(X_1, \ldots, X_n)$, in addition Y_i only depends on X_i, then
$$I(X_1, \ldots, X_n; Y_1, \ldots Y_n) \le \sum I(X_i; Y_i)$$
If $X_1, \ldots, X_n$ are mutually independent, then it is $\ge$ instead.
- **Information Preserving Transform**: If $X \to f(X) \to Y$, then $H(Y|X) = H(Y|f(X))$, and $I(Y;X) = I(Y; f(X))$.
- $I(X;Y|Z)$ can be both smaller ($X = Y = Z$) or larger ($Z = X + Y$) than $I(X;Y)$.

1.4 Inequalities

- **Convex Function**: Function f is said to be convex over interval (a,b) if for every $x_1, x_2 \in (a,b)$ and $\lambda \in [0,1]$, $f(\lambda x_1 + (1-\lambda)x_2) \le \lambda f(x_1) + (1-\lambda) f(x_2)$.
- **Jensen's Inequality**: If f is convex function, then $E[f(X)] \ge f(E[X])$
- **Markov's Inequality**: $P(X \ge a) \le \frac{E[X]}{a}$
- **Chebyshev's Inequality**: $P(|X - E[X]| \ge a) \le \frac{V[X]}{a^2}$
- **Log-sum Inequality**: For non-negative numbers,

$$\sum a_i \log\frac{a_i}{b_i} \ge \left(\sum a_i\right) \log\frac{\sum a_i}{\sum b_i}$$

- $1 - \frac{1}{x} \le \ln x \le x - 1$
- $\alpha \le \lceil \alpha \rceil < \alpha + 1$

2 Symbol-Wise (Variable Length) Source Coding

2.1 Definitions and Basic Results

Source Code: A code C for a random variable X is a mapping from range of X to finite length strings of symbols from a D-ary alphabet.

Expected Length: $L(C) = \sum_{x \in \mathcal{X}} p(x) l(x)$ where $l(x) = |C(x)|$.

Non-singular Code: Code that satisfies $C(x) \neq C(x') \iff x \neq x'$. This suffices for uniquely decoding single letters.

Extension of Code: Extension C^* of a code C is mapping from all finite length strings of $\mathcal{X}$ to strings of D: $C(x_1 \ldots x_n) = C(x_1) \ldots C(x_n)$ (concatenation).

Uniquely Decodable Code: Code whose extension is non-singular. That is $C(x_1 \ldots x_n) \neq C(y_1 \ldots y_m) \iff x_1 \ldots x_n \neq y_1 \ldots y_m$.

Prefix-Free Code or Instantaneous Code: If no codeword is a prefix of another codeword.

Kraft's Inequality: For any prefix-free code must satisfy $\sum_{x \in \mathcal{X}} D^{-l(x)} \le 1$.

- **Proof:** Consider D-ary tree and mark the nodes corresponding to codewords. Consider a random walk from the root. Probability of hitting any marked node ≤ 1.
- **Or,** write the number of nodes in l_{max} depth in two ways.

Entropy Bound: For $X \sim P_X$ and any prefix free code C, $L(C) \ge H(X)$ with equality iff $P_X(x) = 2^{-l(x)}$.

- **Proof**: $L(C) - H(X) = \sum_x p(x) l(x) - \sum_x p(x) \log\frac{1}{p(x)} = \sum_x p(x) \log\frac{p(x)}{2^{-l(x)}} = D(p \parallel q) + \log\frac{1}{c} \ge 0$ where $c = \sum_x 2^{-l(x)}$ and $q(x) = \frac{2^{-l(x)}}{c}$.

2.2 Shannon-Fano Code

Shannon-Fano Code: Set $l(x) = \left\lceil \log\frac{1}{p(x)} \right\rceil$. Satisfies $H(X) \le L(C) < H(X) + 1$.

Wrong Shannon-Fano Code: If $l(x)$ is set according to PMF Q, but in reality it is P, then $H(X) + D(P \parallel Q) \le L(C) \le H(X) + D(P \parallel Q) + 1$.

Shannon-Fano-Elias Code:

- Suppose $F(x)$ is cumulative distribution function. Let $\bar{F}(x) = \sum_{a<x} p(a) + \frac{1}{2} p(x) \in (F(x-1), F(x))$.
- Taking binary of $\bar{F}(x)$ gives the code, but it can be of infinite length.
- So, truncate $\bar{F}(x)$ to $l(x) = \lceil \log\frac{1}{p(x)} \rceil + 1$ bits.
- $\bar{F}(x) -$ truncated $\bar{F}(x) < 2^{-l(x)} < \frac{p(x)}{2} = \bar{F}(x) - F(x-1)$. So truncated version still works.
- $H(X) + 1 \le L(C) < H(X) + 2$.

Competitive Optimality of Shannon-Fano Code: Shannon-Fano code $l(x)$ vs any other code $l'(x)$ satisfies $Pr[l(X) \ge l'(X) + c] \le \frac{1}{2^{c-1}}$. Probability that Shannon-Fano gives codeword of a randomly drawn symbol c bits bigger is exponentially low.

Improving over 1-bit: Encode blocks of N symbols. Even if we assume all symbols are independent, then we have $NH(X) \le L(C) \le NH(X) + 1$, or $H(X) \le$ average length per symbol $\le H(X) + \frac{1}{N}$. Disadvantage is that building coding complexity gets exponentially more complex with N.

2.3 Huffman Code

Keep merging two symbols with lowest probability. Proof outline: For any distribution there is an optimal code where:

- $p_i > p_j$ implies $l_i \le l_j$. (Proof: exchange argument)
- Longest two codes have same length. (Proof: trim if not)
- Longest two code words only differ in last bit, and correspond to two least likely symbols. (Proof: trim if not and rearrange)

3 Block (Fixed Length) Source Coding

Problem: $(X_1, \ldots, X_n)$ mapped to one of $[M]$ and decoded to $(\hat{X}_1, \ldots, \hat{X}_n)$ with some probability of error P_e. Discrete memory less assumption: Each $X_i \in \mathcal{X}$ and i.i.d $\sim P_X$. The (compression) rate is $R = \frac{1}{n}\log M$. The lower the rate, the more compressed.

Asymptotic Equipartition Property: For $X_1, \ldots, X_n \sim p(x)$, $\Pr[\left|\frac{-1}{n}\log p(X_1, \ldots, X_n) - H(x)\right| > \epsilon] \to 0$ as $n \to \infty$ for any $\epsilon > 0$.

Typical Set:

$$T_n(\epsilon) = \{\mathbf{x} \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \le P_{\mathbf{X}}(\mathbf{x}) \le 2^{-n(H(X)-\epsilon)}\}$$

Typical Set Properties

- $H(X) - \epsilon \le \frac{1}{n}\sum \log \frac{1}{P_X(x_i)} \le H(X) + \epsilon$
- $\Pr[\mathbf{X} \in T_n(\epsilon)] > 1 - \epsilon$ for large enough n.
- $|T_n(\epsilon)| \le 2^{n(H(X)+\epsilon)}$.
- $|T_n(\epsilon)| \ge (1-\epsilon)2^{n(H(X)-\epsilon)}$ for large enough n.

Fano's Inequality: $H(X \mid \hat{X}) \le H_2(P_e) + P_e \log(|\mathcal{X}| - 1) \le 1 + P_e \log|\mathcal{X}|$.

Fixed-Length Source Coding Theorem:

- (Achievability) If $R > H(X)$, then for any $\epsilon > 0$, $\exists$ a code of rate R with $P_e \le \epsilon$ (for sufficiently large n).
- (Converse) If $R < H(X)$, then $\exists \epsilon > 0$ such any code with rate R has $P_e > \epsilon$ for any n.

Achievability Proof: Let $M = |T_n(\epsilon)| + 1$. Map typical set elements to $[M-1]$ and everything else to M.

Converse Proof: $nR \ge I(\mathbf{X}; \hat{\mathbf{X}}) \ge H(\mathbf{X}) - H(\mathbf{X} \mid \hat{\mathbf{X}}) \ge nH(X) - 1 - nP_e \log|\mathcal{X}| \Rightarrow P_e \ge \frac{1}{\log|\mathcal{X}|}\left(H(X) - R - \frac{1}{n}\right)$. As $n \to \infty$, P_e cannot be arbitrarily small.

4 Channel Coding

Problem: Encoder takes in one of $[M]$ messages and sends $X_1, \ldots, X_n$ via a channel, which outputs $Y_1, \ldots Y_n$ using distribution $P_{Y|X}$, then decoder decodes that into $\hat{m}$. Rate $R = \frac{1}{n}\log M$ measured in bits per channel use. High rate better.

Channel Capacity: Maximum rate for which there is some n block code achieving arbitrarily small decoding error. $C = \max_{P_X} I(X;Y)$.

Example Channels:

- Noiseless binary channel: $C = 1$.
- Binary Symmetric Channel: $C = 1 - H_2(p)$.
- Binary Erasure Channel: $C = 1 - \alpha$.
- Symmetric Channel (rows are permutations of each other, sum of columns same): $C = \log|\mathcal{Y}| - H(row)$.

Jointly Typical Sets: $T_n(\epsilon)$ is all $(\mathbf{x}, \mathbf{y})$ where $\mathbf{x}$, $\mathbf{y}$, and $(\mathbf{x}, \mathbf{y})$ are typical in $P_{\mathbf{X}}$, $P_{\mathbf{Y}}$, and $P_{\mathbf{XY}}$ respectively.

Jointly Typical Set Properties:

- $|T_n(\epsilon)| \le 2^{n(H(X,Y)+\epsilon)}$.
- If $(\mathbf{X}, \mathbf{Y}) \sim P_{\mathbf{XY}}$ then $\Pr[(\mathbf{X}, \mathbf{Y}) \in T_n(\epsilon)] > 1 - \epsilon$ for large enough n.
- If $(\mathbf{X}', \mathbf{Y}') \sim P_{\mathbf{X}} \times P_{\mathbf{Y}}$ then $\Pr[(\mathbf{X}', \mathbf{Y}') \in T_n(\epsilon)] \le 2^{-n(I(X;Y)-3\epsilon)}$. For large enough n, also $\ge (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}$.

Channel Coding Theorem:

- (Achievability) For any rate $R < C$, $\exists$ code with rate at least R with arbitrarily small error probability.
- (Converse) For any $R > C$, any code of rate R cannot have arbitrarily small error probability.

Achievability via Random Coding: Generate codebook randomly. Decode by finding a jointly typical code jointly typical with $\mathbf{Y}$. Error occurs if there is none or more than one. Due to symmetry, average P_e is same as the average probability of decoding error if message 1 was sent.

$$P_e = \Pr\left[(\mathbf{X}^{(1)}, \hat{\mathbf{X}}) \notin T_n(\epsilon) \text{ or } \bigcup_{i=2}^{M} (\mathbf{X}^{(i)}, \hat{\mathbf{X}}) \in T_n(\epsilon)\right]$$
$$\le \epsilon + 2^{-n(I(X;Y)-3\epsilon-R)}$$

If $R < I(X;Y) - 3\epsilon$ then as $n \to \infty$, $P_e \to 0$. So, in a random codebook, if $X \sim p(x)$, we can send up to rate $I(X;Y)$ with $P_e \to 0$. Maximizing over $p(x)$ gives the result.

Converse Proof: $nC \ge I(\mathbf{X}; \mathbf{Y}) \ge I(m, \hat{m}) = H(m) = H(m \mid \hat{m}) \ge nR - 1 - P_e nR \Rightarrow P_e \ge 1 - \frac{C}{R} - \frac{1}{nR}$. If $R > C$ then P_e is bounded away from 1.

Joint Source-Channel Coding: If source block size is k and channel block size is n, then $nC > kH$ is needed.

5 Continuous Alphabet Channels

Examples of Differential Entropy:

- **Uniform**: $X \sim U(a,b) \Rightarrow h(X) = \log(b-a)$.
- **Normal**: $X \sim N(0, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{x^2}{2\sigma^2}} \Rightarrow h(x) = \frac{1}{2}\log(2\pi e\sigma^2)$.
- **Exponential**: $X \sim \exp(\lambda) = \lambda e^{-\lambda x} \Rightarrow h(X) = \log\frac{e}{\lambda}$.
- **Laplace**: $X \sim \frac{1}{2b} e^{-|x|/b} \Rightarrow h(X) = \log(2eb)$.

Properties:

- Chain rule, conditioning reduces entropy, sub-additivity holds.
- Non-negativity, invariance under one-one transform doesn't hold.
- $h(X) = h(X+c)$ for constant c.
- $h(cX) = h(X) + \log|c|$. If $X \sim f_X$. Then $cX \sim \frac{1}{|c|} f_X(\frac{x}{c})$.
- I and D works as expected.
- $I(f(X); g(Y)) = I(X;Y)$ for inevitable functions f and g.
- $\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y$.

Maximum Entropy:

- **Fixed** σ^2: $h(X) \le \frac{1}{2}\log(2\pi e\sigma^2)$ with equality iff Gaussian.
- **Fixed range** $[l, r]$: $h(X) \le \log(r - l)$ with equality iff uniform.
- **Fixed** μ: $h(X) \le \log(e\mu)$ with equality iff $X \sim \exp(\mu) = \frac{1}{\mu}e^{-x/\mu}$.

Additive White Gaussian Noise Channel: Output $Y = X + Z$ where $Z \sim N(0, \sigma^2)$. Encoder need to ensure $E[\mathbf{X}^2] \le P$. The average can be either over each codeword, or over each element of each codeword.

Channel Capacity for AWGN: $C = \max_{f_X : E[X^2] \le P} I(X;Y) = \frac{1}{2}\log(1 + \frac{P}{\sigma^2})$. Capacity achieving f_X is $N(0, P)$.

6 Practical Codes

Encoder gets bit vector $\mathbf{u}_{1\times k}$ and outputs $\mathbf{x}_{1\times n}$. Channel gives $\mathbf{y} = \mathbf{x} \oplus \mathbf{z}$ for some random vector $\mathbf{z}$. Decoder decodes into $\hat{\mathbf{u}}$. Rate $R = \frac{k}{n}$.

Linear Code: Where $\mathbf{x}_{1\times n} = \mathbf{u}_{1\times k}\mathbf{G}_{k\times n}$ modulo 2. Symmetric parity-check code if first k columns of $\mathbf{G}$ is I_k, i.e. $x_i = u_i$ for $i \le k$. Codewords are therefore linear combination of rows of $\mathbf{G}$. If $\mathbf{x}$, $\mathbf{x}'$ correspond to $\mathbf{u}$, $\mathbf{u}'$ then $\mathbf{x} \oplus \mathbf{x}'$ correspond to $\mathbf{u} \oplus \mathbf{u}'$.

Parity Check Matrix: $\mathbf{x}_{1\times n}\mathbf{H}_{n\times(n-k)} = 0$ iff $\mathbf{x}$ is a codeword. If $\mathbf{G} = [I_k \quad P_{k\times(n-k)}]$ then $\mathbf{H}$ is P stacked on top of I_k. We have $\mathbf{yH} = \mathbf{zH}$ which helps in decoding.

Hamming Distance: $d_H(\mathbf{x}, \mathbf{x}')$ is number of unmatched bits.

Minimum Distance: $d_{\min} = d_H(\mathbf{x}, \mathbf{x}')$ for any two $\mathbf{x} \ne \mathbf{x}' \in$ codebook.

- Can correct $d_{\min} - 1$ erasures (that is bit replaced with '?').
- Can correct $\frac{1}{2}(d_{\min} - 1)$ flips.
- Let $w(\mathbf{x})$ be the popcount of $\mathbf{x}$. Then $d_{\min} = \min w(\mathbf{x})$, $\mathbf{x} \ne 0$. Proof: $d_{\min} = d_H(\mathbf{x}, \mathbf{x}') = w(\mathbf{x} \oplus \mathbf{x}') = w(\mathbf{x}'')$ and $w(\mathbf{x}) = d_H(0, \mathbf{x})$.

Maximum A Posteriori Decoding (MAP): Decode to $\mathbf{x}^{(j)}$ which has maximum $P_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}^{(j)}|\mathbf{y})$. Always optimal in terms of P_e.

Maximum Likelihood Decoding: If m uniform on $[M]$ then decode to $\mathbf{x}^{(j)}$ which maximizes $P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(j)})$.

Minimum Distance Decoding (on BSC): Maximum Likelihood Decoding is equivalent to decoding to $X^{(j)}$ with minimum d_H.

Syndrome Decoding (Linear Code on BSC): Syndrome associated with $\mathbf{y}$ is $\mathbf{S}_{1\times(n-k)} = \mathbf{y}_{1\times n}\mathbf{H}_{n\times(n-k)}$. Find $\hat{\mathbf{z}}$ that minimizes $w(\hat{\mathbf{z}})$ and satisfies $\hat{\mathbf{z}}\mathbf{H} = S$. Decode $\mathbf{x} = \mathbf{y} \oplus \hat{\mathbf{z}}$. Equivalent to Minimum Distance Decoding.