Independence Bound for entropy. $X_1, X_2, \cdots, X_n$ drawn according to $p(x_1, x_2, \cdots, x_n)$. Then $H(X_1, X_2, \cdots, X_n) \le \sum_{i=1}^{n} H(X_i)$ = iff X_i are independent.

Proof: Use chain rule for entropy & fact that conditioning reduces entropy

1 nat = $\frac{1 \text{ bit}}{\ln 2}$

(1) Entropy

→ X: discrete R.V. with alphabet $\mathcal{X}$ and pmf $p(x) = Pr\{X = x\}$, $x \in \mathcal{X}$

→ $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log_2 p(x)$

→ $\lim_{\varepsilon \to 0} \varepsilon \log \varepsilon = 0$

→ $H(X) = E_p \log \frac{1}{p(X)}$

→ $H(X) \ge 0$

Concave Fⁿ ← $H(p) \overset{\text{def}}{=} -p \lg p - (1-p) \lg (1-p)$

→ Min. expected binary questions to determine X $\in (H(X), H(X)+1)$

(2) Joint Entropy $H(X,Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(x,y)$

(3) Conditional Entropy

$H(Y|X) = \sum_{x \in \mathcal{X}} p(x) H(Y|X=x)$

$= -E \log p(Y|X)$

$= -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x,y) \log p(y|x)$

(4) Chain Rule

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

(5) Relative Entropy or Kullback-Leibler divergence

b/w 2 pmf $p(x)$ and $q(x)$

$D(p\|q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} = E_p \log \frac{p(X)}{q(X)}$

convex in pair (p, q)

(6) Mutual Information

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

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

$= E_{p(x,y)} \log \frac{p(X,Y)}{p(X)p(Y)}$

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

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

$I(X;Y) = I(Y;X)$

$I(X;X) = H(X)$

(7) Chain Rule for Entropy

Let $X_1, X_2, \cdots, X_n$ be drawn according to $p(x_1, x_2, \cdots, x_n)$. Then.

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

proof: $p(x_1, \cdots, x_n) = \prod_{i=1}^{n} p(x_i | x_{i-1}, \cdots, x_1)$

(8) Conditional Mutual Information of RV X and Y given Z is defined by

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

$= E_{p(x,y,z)} \log \frac{p(X,Y|Z)}{p(X|Z)p(Y|Z)}$

(9) Chain Rule for Information

$I(X_1, X_2, \cdots, X_n; Y) = \sum_{i=1}^{n} I(X_i; Y | X_{i-1}, X_{i-2}, \cdots, X_1)$

$I(X;Y|Z) = \sum_z p(z) I(X;Y|Z=z)$

(10) Chain Rule for relative entropy.

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

Conditional Relative Entropy.

$D(p(y|x)\|q(y|x)) = \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} = E_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}$

(11) Jensen's Inequality. $E(f(X)) \ge f(E(X))$ if f is convex

Let $p(x), q(x)$ $x \in \mathcal{X}$ be 2 pmf

Information Inequality. $D(p\|q) \ge 0$

(12) Consequences

1. Non negativity of mutual information $I(X;Y) \ge 0$
2. $I(X;Y|Z) \ge 0$ $=0$ if $X \perp Y$ conditionally independent given Z
3. $H(X) \le \log |\mathcal{X}|$
4. $H(X|Y) \le H(X)$. Conditioning reduces entropy. Equality if independent
5. $H(X_1, X_2, \cdots, X_n) \le \sum_{i=1}^{n} H(X_i)$ = if X_i independent

(13) Log Sum Inequality For nonnegative nos $a_1, a_2, \cdots, a_n$ & $b_1, b_2, \cdots, b_n$

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

equality iff $\frac{a_i}{b_i} = \text{const} \; \forall i$

Consequences $D(p\|q) \ge 0$

(14) Convexity of Relative Entropy. $D(p\|q)$ is convex in pair (p, q)

i.e if (p_1, q_1) and (p_2, q_2) are 2 pairs of pmf then

$D(\lambda p_1 + (1-\lambda) p_2 \| \lambda q_1 + (1-\lambda) q_2) \le \lambda D(p_1\|q_1) + (1-\lambda) D(p_2\|q_2)$

(15) Concavity of Entropy. $H(p)$ is a concave function of p.

Save images as pdf
Screenshot JPEG loses detail.

classmate
Date
Page

Let $(X,Y) \sim p(x,y)$
The mutual information $I(X;Y)$ is a concave f^n of $p(x)$ for fixed $p(y|x)$ & convex f^n of $p(y|x)$ for fixed $p(x)$

16

Data Processing Inequality.

$X \to Y \to Z$

$p(x,y,z) = p(x)p(y|x)p(z|y)$

$\Rightarrow p(x,z|y) = p(x|y)p(z|y)$

If $X \to Y \to Z$

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

Sufficient Statistic

Sufficient Statistic

PMF $\{f_\theta(x)\}$ indexed by θ, X sample.

$\theta \to X \to T(X)$

$I(\theta;T(X)) \leq I(\theta;X)$.

A statistic $T(X)$ is called sufficient for θ if it contains all info in X about θ

A f^n $T(X)$ is sufficient statistic relative to $\{f_\theta(x)\}$ if $\theta \to T(X) \to X$

Minimal Sufficient Statistic (maximally compresses info about θ)

A statistic $T(X)$ is minimal sufficient statistic relative to $\{f_\theta(x)\}$ if it is a f^n of every other statistic U

$\theta \to T(X) \to U(X) \to X$

Fano's Inequality For any estimator $\hat{X}$ such that $X \to Y \to \hat{X}$, with $P_e = Pr(X \neq \hat{X})$

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

weakened to

$1 + P_e \log|\mathcal{X}| \geq H(X|Y)$

$D(p\|q) \geq 2 d_{TV}(p,q)^2 \log_2 e \geq \frac{\log_2 e}{2} \|p-q\|_1^2$

$P(Y \geq p+\varepsilon) \approx 2^{-nD(p+\varepsilon\|p)}$

Ch5 Data Compression

Nonsingular. $x \neq x' \Rightarrow C(x) \neq C(x')$

Uniquely Decodable

Prefix Code / Instantaneous Code: if no CW is prefix of any other codeword
↳ can be decoded without reference to future code-words

Kraft Inequality For any instantaneous code over an alphabet size of length D, the codeword lengths $l_1, l_2, \cdots, l_m$ must satisfy $\sum_i D^{-l_i} \leq 1$.

Conversely given a set of codeword lengths that satisfy this inequality, $\exists$ an instantaneous code with these word lengths.

Hoeffding Bound X_i : i.i.d. $\in \{0,1\}$ $Pr(X_i=1) = p$

$Pr(Y \geq p+\varepsilon) \leq 2^{-nD(p+\varepsilon\|p)}$ $Pr(Y \leq p-\varepsilon) \leq 2^{-nD(p-\varepsilon\|p)}$

→ For $\lambda \in (0,1)$ and large n $\binom{n}{n\lambda} \approx 2^{nH(\lambda)}$

→ For $(\alpha \leq \frac{1}{2})$ $\sum_{i \leq \alpha n} \binom{n}{i} \leq 2^{nH(\alpha)}$

Strongly Typical Sequences

A sequence $x^n \in \mathcal{X}^n$ is said to be ε-strongly typical wrt a distribution $p(x)$ on $\mathcal{X}$ if

$\forall a \in \mathcal{X} : |P_{x^n}(a) - p(a)| \leq p(a)\varepsilon$

Typical Set: $T_\varepsilon^{(n)}$:= set of all ε-typical vectors of length n.

1. a sequence chosen at random will be in typical set with probability almost one.
2. All elements of typical set have (almost) equal probabilities.

Typical average lemma Suppose $\vec{x}$ is a typical sequence. Then for any non-ve f^n $g(x)$ on $\mathcal{X}$

$(1-\varepsilon)\mathbb{E}[g(X)] \leq \frac{1}{n}\sum_{i=1}^n g(x_i) \leq (1+\varepsilon)\mathbb{E}[g(X)]$

$2^{-nH(X)(1+\varepsilon)} \leq \prod_{x \in \mathcal{X}} p(x)^{nT_{\vec{x}}(x)} \leq 2^{-nH(X)(1-\varepsilon)}$

For any $a \in \mathcal{X}$, let $N_{\vec{x}}(a)$ denote number of a in $\vec{x} \in \mathcal{X}^n$, & $T_{\vec{x}}(a) = \frac{N_{\vec{x}}(a)}{n}$

Sanov's Theorem

The vector $\left(T_{\vec{x}}(a)\right)_{a \in \mathcal{X}}$ is called the type of $\vec{x}$.

No. of types $\leq (n+1)^{|\mathcal{X}|}$

Let X_i be i.i.d R.V with pmf $\sim p$ over alphabet $\mathcal{X}$. Let q be a valid type for n-length vectors in $\mathcal{X}$. Then

$P(T_{\vec{x}} = q) \simeq 2^{-nD(q\|p)}$

Let X_i be i.i.d $\sim Q(x)$. Let E be a subset of all probability distributions on $\mathcal{X}$. Then

$Pr(T_{\vec{x}} \in E) \leq (n+1)^{|\mathcal{X}|} 2^{-nD(P^*\|Q)}$

where $P^* = \underset{P \in E}{\operatorname{argmin}}\, D(P\|Q)$

1-0.5
0.5

Ch11 Info Theory & Statistics

$X_1, X_2, \cdots, X_n$ be a sequence of n symbols from an alphabet $\mathcal{X} = \{a_1, a_2, \cdots, a_{|\mathcal{X}|}\}$.

$x^n \leftrightarrow \vec{x} \Leftrightarrow x_1, x_2, \cdots, x_n$

The type $P_{\vec{x}}$ (or empirical prob. distribution) of a sequence $\vec{x}$ is the relative proportion of occurrences of each symbol of $\mathcal{X}$, i.e. $P_{\vec{x}}(a) = N(a|\vec{x})/n$ (no. of x symbol a appears in $\vec{x} \in \mathcal{X}^n$)

Use capital letters to denote types & distributions

If $P \in \mathcal{P}_n$, the set of sequences of length n & type P is called type class of P, denoted $T(P)$

$T(P) = \{\vec{x} \in \mathcal{X}^n : P_{\vec{x}} = P\}$

$|\mathcal{P}_n| \le (n+1)^{|\mathcal{X}|}$

Let $X_1, \cdots, X_n$ drawn i.i.d. according to $Q(x)$

$Q^n(x^n) = \prod_{i=1}^n Q(x_i)$

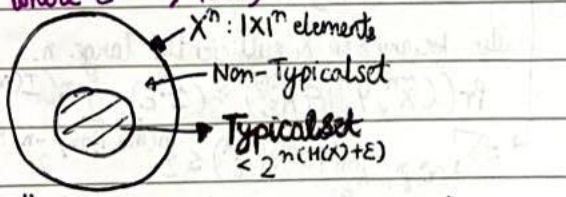
Assuming $\min_{a \in \mathcal{X}} P(a) = P_{min} > 0$

and $0 < \varepsilon < P_{min}$

Probability of Typical Set:

$\Pr(A_\varepsilon^{(n)}) \ge 1 - 2|\mathcal{X}|2^{-n\varepsilon'^2 \log e}$

where $\varepsilon' = \varepsilon / H(X)$



"The smallest high probability set"

→ Typical sequences are not necessarily high-probability sequences.

$X_i \sim Ber(\frac{3}{4})$. The highest p sequence is all ones which is not typical.

AEP (follows from WLLN)

If $X_i \sim$ i.i.d $p(x)$ then

$-\frac{1}{n} \log p(X_1, X_2, \cdots, X_n) \to H(X)$ i.p.

(Weakly Typical set) The typical set $A_\varepsilon^{(n)}$ wrt distribution $p(x)$ is the set of all sequences $(x_1, x_2, \cdots, x_n) \in \mathcal{X}^n$ with property

$2^{-n(H(X)-\varepsilon)} \ge p(x_1, x_2, \cdots, x_n) \ge 2^{-n(H(X)+\varepsilon)}$

A strongly ε-typical sequence is also weakly $H(X)\varepsilon$ typical.

Weak Typicality $\not\Rightarrow$ Strong Typicality.

Weakly Typical sequence may not 'look' typical. Take $Ber(\frac{1}{2})$ all sequences are weakly typical.

Properties of typical set

1. If $(x_1, x_2, \cdots, x_n) \in A_\varepsilon^{(n)}$, then

$H(X) - \varepsilon \le -\frac{1}{n} \log p(x_1, x_2, \cdots, x_n) \le H(X) + \varepsilon$

2. For any $\delta > 0$ $\Pr\{A_\varepsilon^{(n)}\} > 1 - \delta$ for sufficiently large n. (strong)

In fact we know result $\Pr(A_\varepsilon^{(n)}) \ge 1 - 2|\mathcal{X}|2^{-n\varepsilon^2 \log e \frac{1}{2}}$

3. $|A_\varepsilon^{(n)}| \le 2^{n(H(X)+\varepsilon)}$

4. For any $\delta > 0$ $|A_\varepsilon^{(n)}| \ge (1-\delta)2^{n(H(X)-\varepsilon)}$ for large n

3.3 High-Probability Sets and Typical sets.

$A_\varepsilon^{(n)}$ is a fairly small set that contains most of prob.

But from definition, it is not clear whether it is smallest such set.

Typical set has essentially the same number of elements as smallest set, to first order in exponent.

Def For each $n = 1, 2, \cdots$, let $B_\delta^{(n)} \subset \mathcal{X}^n$ be the smallest set with $\Pr\{B_\delta^{(n)}\} \ge 1 - \delta$

Let X_i iid $\sim p(x)$ for $\delta < \frac{1}{2}$ and any $\delta' > 0$, if $\Pr(B_\delta^{(n)}) > 1 - \delta$ then

$\frac{1}{n} \log |B_\delta^{(n)}| > H - \delta'$ for n sufficiently large.

Source Coding Theorem

$\inf\{R : R \text{ is achievable}\} = H(X)$

Channel Coding Theorem.

Capacity of a channel = supremum of all achievable rates

$C = \max_{p(x)} I(X;Y)$

For channels with non overlapping o/p $C = \log |\mathcal{X}|$

Gaussian Channel

Def Information Capacity of a Gaussian Channel with power constraint P is

$$C = \max_{f(x):\mathbb{E}[X^2]\le P} I(X;Y)$$

For any codeword $(x_1, x_2, \cdots, x_n)$ say constraint $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(Y) - h(Z) \qquad E[Y^2] = P+N \\ &\le \tfrac{1}{2}\log 2\pi e(P+N) - \tfrac{1}{2}\log 2\pi e N \\ &= \tfrac{1}{2}\log\left(1+\frac{P}{N}\right) \end{aligned}$$

→ An (M,n) code for Gaussian Channel with Power Constraint P: $\sum_{i=1}^{n} x_i^2(\omega) \le nP$, $\omega = 1,2,\cdots,M$

an index set $\{1,2,\cdots,M\}$

An encoding f^n $x:\{1,2,\cdots,M\} \to \mathcal{X}^n$ codewords $x^n(1), x^n(2), \cdots, x^n(M)$

A decoding f^n $g: \mathcal{Y}^n \to \{1,2,\cdots M\}$

$$P_e^{(n)} = \frac{1}{2^{nR}}\sum \lambda_i$$

Def: A rate R is said to be achievable for a Gaussian channel with power constraint P if $\exists$ sequence of $(2^{nR}, n)$ codes with codewords satisfying Power constraint such that maximal probability of error $\lambda^{(n)} \to 0$. The Capacity of channel is supremum of the achievable rates.

Theorem 9.1.1 Capacity of a Gaussian Channel with power constraint P & noise variance N is

$$C = \tfrac{1}{2}\log\left(1+\frac{P}{N}\right) \text{ bits/Tx}$$

Achievability:

Generation of Codebook: Codewords iid $\sim N(0, P-\varepsilon)$ $\frac{1}{n}\sum X_i^2 \to P-\varepsilon$

Encoding: Codebook revealed to both sender & receiver. To send message index w, tx sends w^{th} codeword $X^n(w)$

Decoding: Receiver searches in codebook, codeword that is jointly typical with received vector. If only one o/p it else declare error.

$$E_0 = \left\{\frac{1}{n}\sum_{j=1}^{n} X_j^2(1) > P\right\} \qquad E_i = \left\{(X^n(i), Y^n) \text{ is in } A_\varepsilon^{(n)}\right\}$$

$$\begin{aligned} P(\mathcal{E}|\omega=1) = P(\mathcal{E}) &= P(E_0 \cup E_1^c \cup E_2 \cup E_3 \cup \cdots \cup E_{2^{nR}}) \\ &\le P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \end{aligned}$$

by LLN $\to 0$ as $n\to\infty$; by joint AEP $P(E_i^c) \to 0$ hence $P(E_1^c) \le \varepsilon$ for large n; by codebook generation $X^n(i)$ & $X^n(1)$ independent

$$\begin{aligned} P_e^{(n)} = Pr(\mathcal{E}) &= Pr(\mathcal{E}|\omega=1) \\ &\le P(E_0) + P(E_1^c) + \sum_{i=2}^{2^{nR}} P(E_i) \\ &\le \varepsilon + \varepsilon + (2^{nR}-1)\, 2^{-n(I(X;Y)-3\varepsilon)} \\ &\le 2\varepsilon + 2^{3n\varepsilon}\, 2^{-n(I(X;Y)-R)} \end{aligned}$$

→ deleting worst half.

Converse to the coding Theorem for Gaussian Channels.

Proof that capacity of a Gaussian channel is $C = \frac{1}{2}\log(1+\frac{P}{N})$ by proving rates $R > C$ are not achievable.

Proof: Let W be distributed uniformly over $\{1,2,\cdots,2^{nR}\}$.

$$W \to X^{(n)}(W) \to Y^n \to \hat{W}$$

$$H(W|\hat{W}) \le 1 + nR P_e^{(n)} = n\varepsilon_n \qquad \varepsilon_n \to 0 \text{ as } P_e^{(n)} \to 0$$

$$\begin{aligned} nR = H(W) &= I(W;\hat{W}) + H(W|\hat{W}) \\ &= I(W;\hat{W}) + n\varepsilon_n \\ &\le I(X^n, Y^n) + n\varepsilon_n \\ &= h(Y^n) - h(Y^n|X^n) + n\varepsilon_n \\ &= h(Y^n) - h(Z^n) + n\varepsilon_n \\ &\le \sum_{i=1}^{n} h(Y_i) - h(Z^n) + n\varepsilon_n \\ &= \sum_{i=1}^{n} h(Y_i) - \sum_{i=1}^{n} h(Z_i) + n\varepsilon_n \\ &= \sum_{i=1}^{n} I(X_i;Y_i) + n\varepsilon_n \end{aligned}$$

$X_i = x_i(W)$, $W \in$ uniform $\{1,2,\cdots,2^{nR}\}$

P_i: average power of i^{th} column of codebook, that is $P_i = \frac{1}{2^{nR}}\sum_{w=1}^{M} x_i^2(w)$

$Y_i = X_i + Z_i$, $E[Y_i^2] = P_i + N$. Hence since entropy is maximized by the normal distribution $h(Y_i) \le \frac{1}{2}\log 2\pi e(P_i+N)$

$$\begin{aligned} nR &\le \sum_i (h(Y_i) - h(Z_i)) + n\varepsilon_n \\ &\le \sum\left(\frac{1}{2}\log 2\pi e(P_i+N) - \frac{1}{2}\log 2\pi e N\right) + n\varepsilon_n \\ &= \sum \frac{1}{2}\log\left(1+\frac{P_i}{N}\right) + n\varepsilon_n \end{aligned}$$

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\left(1+\frac{P_i}{N}\right) \le \frac{1}{2}\log\left(1+\frac{1}{n}\sum\frac{P_i}{N}\right) \quad \text{subject to } \frac{1}{n}\sum_i P_i \le P$$

$$\le \frac{1}{2}\log\left(1+\frac{P}{N}\right)$$

$$R \le \frac{1}{2}\log\left(1+\frac{P}{N}\right) + \varepsilon_n, \quad \varepsilon_n \to 0$$

9.4 Parallel Gaussian Channels

Considering k independent Gaussian channels in parallel with a common power constraint. Distribute the total power among the channels so as to maximize the capacity.

Non White Additive Gaussian Noise

For channels $Y_j = X_j + Z_j \quad j = 1, 2, \cdots, k$

$Z_j \sim \mathcal{N}(0, N_j) \qquad E\sum_{j=1}^{k} X_j^2 \le P$

$$C = \max_{f(x_1^k):\sum EX_i^2 \le P} I(X_1^k; Y_1^k)$$

$$I(X_1^k; Y_1^k) = h(Y_1^k) - h(Y_1^k|X_1^k)$$
$$= h(Y_1^k) - \sum_i h(Z_i)$$
$$< \sum_i h(Y_i) - h(Z_i)$$
$$\le \sum_i \frac{1}{2}\log\left(1 + \frac{P_i}{N_i}\right)$$

$P_i = E[X_i^2] \Big/ \sum P_i = P$

= achieved when

$$(X_1^k) \sim \mathcal{N}\left(0, \begin{bmatrix} P_1 & & 0 \\ & P_2 & \\ & & \ddots \\ 0 & & P_k \end{bmatrix}\right)$$

$$J(P_i^k) = \sum_{i=1}^{k} \frac{1}{2}\log\left(1 + \frac{P_i}{N_i}\right) + \lambda\left(\sum P_i - P\right)$$

$$\nabla_{P_i} J(P_i^k) = 0 \Rightarrow \frac{1}{2}\frac{1}{P_i + N_i} + \lambda = 0 \Rightarrow P_i = \nu - N_i$$

$$P_i = (\nu - N_i)^+ \qquad \sum (\nu - N_i)^+ = P$$

→ Channels with colored Gaussian Noise.

15.1 Gaussian Multiple-User Channels.

15.1.2 Gaussian Multiple-Access Channel with m users:

$$Y = \sum_{i=1}^{m} X_i + Z \qquad \text{Let } C\left(\frac{P}{N}\right) = \frac{1}{2}\log\left(1 + \frac{P}{N}\right)$$

$R_1 < C\left(\frac{P}{N}\right)$

$R_1 + R_2 < C\left(\frac{2P}{N}\right)$

$\vdots$

$$\sum_{i=1}^{m} R_i < C\left(\frac{mP}{N}\right)$$

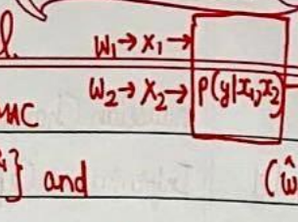
→ when all the rates same, the last inequality dominates the others.

m codebooks, i^{th} codebook having 2^{nR_i} codewords of power P.

Optimal decoding: looking for m codewords, one from each codebook, such that the vector sum is closest to $\vec{Y}$ in Euclidean Distance.

If $(R_1, R_2, \cdots, R_m)$ is in the capacity region, probability of error goes to 0 as $n \to \infty$.

15.3 Multiple-Access Channel.



A $((2^{nR_1}, 2^{nR_2}), n)$ code for MAC consists of $\mathcal{W}_1 = \{1, 2, \cdots, 2^{nR_1}\}$ and $\mathcal{W}_2 = \{1, 2, \cdots, 2^{nR_2}\}$ called message sets

Two encoding f^n $X_1 : \mathcal{W}_1 \to \mathcal{X}_1^n$, $X_2 : \mathcal{W}_2 \to \mathcal{X}_2^n$

& a decoding f^n $g : \mathcal{Y}^n \to \mathcal{W}_1 \times \mathcal{W}_2$

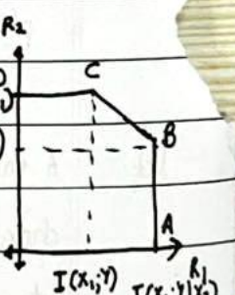
$$P_e^{(n)} = \frac{1}{2^{n(R_1+R_2)}} \sum_{(w_1,w_2) \in \mathcal{W}_1 \times \mathcal{W}_2} P_r\{g(Y^n) \neq (w_1, w_2) \,|\, (w_1, w_2) \text{ sent}\}$$

Def^n: A rate pair (R_1, R_2) is said to be achievable for MAC if $\exists$ a sequence of $((2^{nR_1}, 2^{nR_2}), n)$ codes with $P_e^{(n)} \to 0$

Def^n: The capacity region of MAC is closure of set of achievable (R_1, R_2) rate pairs

Theorem 15.3.1 MAC Capacity: The capacity of a MAC $(\mathcal{X}_1 \times \mathcal{X}_2, p(y|x_1, x_2), \mathcal{Y})$ is closure of convex hull of (R_1, R_2) satisfying $R_1 < I(X_1; Y|X_2)$ $R_2 < I(X_2; Y|X_1)$ $R_1 + R_2 < I(X_1, X_2; Y)$ for some product distribution $p_1(x_1)p_2(x_2)$ on $\mathcal{X}_1 \times \mathcal{X}_2$

→ Comments on capacity region for MAC



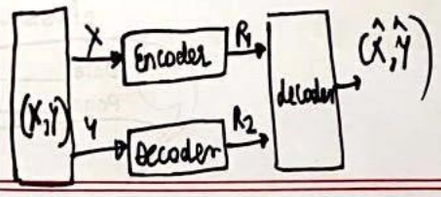
Point A $I(X_1; Y|X_2) = \sum_x p(x_2) I(X_1; Y|X_2 = x_2)$

$\le \max_{x_2} I(X_1; Y|X_2 = x_2)$

X_2 must facilitate transmission of X_1 by setting $X_2 = x_2$

Point B rate that is obtained if X_1 is considered noise for the channel.

X_2 can send at a rate $I(X_2; Y)$. The receiver now knows which X_2 codeword was used & can subtract from the channel



15.4 Encoding of Correlated Sources.

Data Compression dual to Multiple-Access Channel Problem

$(X,Y) \sim p(x,y)$. A rate $H(X,Y)$ is sufficient if we are encoding them together, but what if X & Y sources must be described separately

By separately encoding X and Y, it is seen that a rate $R = R_x + R_y > H(X) + H(Y)$ is sufficient.

It is shown that $R = H(X,Y)$ is sufficient even for separate encoding of correlated sources.

Def: A $((2^{nR_1}, 2^{nR_2}), n)$ distributed source code for joint source (X,Y) consists of 2 encoder maps $f_1 : \mathcal{X}^n \to \{1, 2, \cdots, 2^{nR_1}\}$

$$f_2 : \mathcal{Y}^n \to \{1, 2, \cdots, 2^{nR_2}\}$$

$$g : [2^{nR_1}] \times [2^{nR_2}] \to \mathcal{X}^n \times \mathcal{Y}^n$$

$$P_e^{(n)} = P\left(g\{f_1(X^n), f_2(Y^n)\} \neq (X^n, Y^n)\right)$$

A rate pair (R_1, R_2) is said to be achievable for a distributed source if $\exists$ $((2^{nR_1}, 2^{nR_2}), n)$ distributed source codes with $P_e^{(n)} \to 0$.

Theorem Slepian-Wolf: For the distributed source-coding problem for the source (X,Y) drawn i.i.d $\sim p(x,y)$, the achievable rate region is given by.

$$R_1 \geq H(X|Y), \quad R_2 \geq H(Y|X)$$

$$R_1 + R_2 \geq H(X,Y)$$

A

Channel capacity of a discrete Memoryless Channel

$C = \max_{p(x)} I(X;Y)$

BSC $C = 1 - H(p)$ bits

BEC $C = 1 - \alpha$

Properties of Channel Capacity

1. $C \geq 0$
2. $C \leq \log|\mathcal{X}|$ & $C \leq \log|\mathcal{Y}|$
3. $I(X;Y)$ is a concave f^n of $p(x)$

Memorylessness $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$

No Feedback $p(x_k|x^{k-1}, y^{k-1}) = p(x_k|x^{k-1})$

An (M,n) code for channel $(\mathcal{X}, p(y|x), \mathcal{Y})$

1. Index set $\{1, 2, \cdots, M\}$
2. An encoding f^n $X^n: \{1,2,\cdots,M\} \to \mathcal{X}^n$ yielding codewords $x^n(1), x^n(2), \cdots, x^n(M)$. Set of codewords: Codebook.
3. A decoding f^n $g: \mathcal{Y}^n \to \{1,2,\cdots,M\}$

Rate R of an (M,n) code is $R = \frac{\log M}{n}$ bits per Tx

Def: A rate R is said to be achievable if $\exists$ a sequence of ~~codewords~~ $(\lceil 2^{nR} \rceil, n)$ codes such that maximal probability of error $\lambda^{(n)}$ tends to 0 as $n \to \infty$

$C = \sup_{R: \text{achievable}} R$

$p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$

Jointly Typical Sequences The set $A_\varepsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ wrt $p(x,y)$ is

$$A_\varepsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left|-\tfrac{1}{n}\log p(x^n) - H(X)\right| < \varepsilon, \left|-\tfrac{1}{n}\log p(y^n) - H(Y)\right| < \varepsilon, \left|-\tfrac{1}{n}\log p(x^n, y^n) - H(X,Y)\right| < \varepsilon\}$$

7.6.1 Joint AEP. Let (X^n, Y^n) be sequences of length n drawn iid according to $p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)$

Then 1. $\Pr((X^n, Y^n) \in A_\varepsilon^{(n)}) \to 1$ as $n \to \infty$

2. $|A_\varepsilon^{(n)}| \leq 2^{n(H(X,Y)+\varepsilon)}$

3. If $(\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)$ [i.e. $\tilde{X}^n$ and $\tilde{Y}^n$ are independent with the same marginals as $p(x^n, y^n)$] then

$\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \leq 2^{-n(I(X;Y)-3\varepsilon)}$

also for large n $\Pr((\tilde{X}^n, \tilde{Y}^n) \in A_\varepsilon^{(n)}) \geq (1-\varepsilon)\, 2^{-n(I(X;Y)+3\varepsilon)}$

Channel Coding Theorem.

For DMC all rates below C are achievable.

Specifically for every $R < C$, $\exists$ a sequence $(2^{nR}, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$

Conversely any sequence of $(2^{nR}, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \leq C$.

Definition A channel is said to be symmetric if rows of channel Tx matrix $p(y|x)$ are permutations of each other & columns are permutation of each other

Weakly Symmetric: Every row is a permutation of every other row & column sums $\sum_x p(y|x)$ are equal.

Theorem For weakly symmetric Channel $C = \log|\mathcal{Y}| - H(\vec{r})$ (row)

CCT: Achievability Proof

Choose $R < C$, there is a $p(x)$ s.t. $I(X;Y) > R$

Choose $(\lceil 2^{nR} \rceil, n)$ code by picking all $n \cdot \lceil 2^{nR} \rceil$ code components iid $\sim p(x)$. Denote codewords as $X^n(1), X^n(2), \cdots, X^n(\lceil 2^{nR} \rceil)$

This code is known to Encoder & Decoder.

$w \to X^n(w)$

Typicality decoding: After receiving Y^n choose a codeword jtly typical with Y^n. If none or more declare error.

Probability of Error (over all messages & codes)

$$P_e^n \triangleq \sum_C \Pr(C_n) P_{e,C_n}^{(n)}$$

Clearly $\exists\, C_n$ s.t. $P_{e,C_n}^{(n)} \le P_e^n$

Achievability under maximum error probability.

By expurgating worst half we get

$R, 2\varepsilon \to R - \frac{1}{n}, 4\varepsilon$ (max error)

$\exists$ sequence of codes C_n of rate $R_n \to R < C$ of increasing lengths so that max. error probability $P_{e,C_n,m} \to 0$ as $n \to \infty$

Fano's Inequality: $H(W|\hat{W}) \le H(P_e) + P_e \log(2^{nR}) \le 1 + nRP_e$

Lemma: For any $p(x^n)$, $I(X^n;Y^n) \le nC$

For memoryless & feedback free channels

$p(y_i|y_1, y_2, \cdots, y_{i-1}, x^n) = p(y_i|x_i)$

so $X^n, Y^{i-1} \leftrightarrow X_i \leftrightarrow Y_i$ M.C.

$H(Y_i|Y_1^{i-1}, X^n) = H(Y_i|X_i)$

(i.e. if I choose some thing also MC X_i)

$nR = H(W)$

$I(W, Y^n)$

$W, X^{i-1}, Y^{i-1} \leftrightarrow X_i \leftrightarrow Y_i$

$H(X^\Delta) = -\log\Delta + h(X)$

$H(X^\Delta) = -\log\Delta + h(X)$

Differential Entropy. $h(X)$ of a continuous R.V. X with density $f(x)$ is: $h(X) := -\int_S f(x)\log f(x)\,dx$

$h(X+c) = h(X)$

$h(aX) = h(X)$ pdf change $+\log|a|$

$h(AX) = h(X) + \log|A|$

S: support set of R.V.

Ex $X \sim \phi(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-x^2}{2\sigma^2}}$

$h(\phi) = \frac{1}{2}\ln(2\pi e\sigma^2)$ nats

$= \frac{1}{2}\log(2\pi e\sigma^2)$ bits

AEP for Continuous R.V.

8.2.1 Let $X_i \sim iid\ f(x)$ then

$-\frac{1}{n}\log f(X_1, X_2, \cdots, X_n) \to E[-\log f(X)] = h(X)$ i.p.

Definition $A_\varepsilon^{(n)} = \left\{(x_1, x_2, \cdots, x_n) \in S^n : \left|-\frac{1}{n}\log f(x_1, x_2, \cdots x_n) - h(X)\right| \le \varepsilon\right\}$

Properties

1. $Pr(A_\varepsilon^{(n)}) \to 1-\varepsilon$
2. $Vol(A_\varepsilon^{(n)}) \le 2^{n(h(X)+\varepsilon)}$
3. $Vol(A_\varepsilon^{(n)}) \ge (1-\varepsilon)2^{n(h(X)-\varepsilon)}$

Theorem The set $A_\varepsilon^{(n)}$ is the smallest volume set w.p. $\ge 1-\varepsilon$ to first order in exponent.

8.3 **Rel^n of Diff. Entropy to Discrete Entropy.**

$f(x_i)\Delta = \int_{i\Delta}^{(i+1)\Delta} f(x)\,dx$

$X^\Delta = x_i$ if $i\Delta \le X < (i+1)\Delta$

$P(X^\Delta = x_i) \approx f(x_i)\Delta$

$H(X^\Delta) = -\sum(f(x_i)\Delta)\log(f(x_i)\Delta)$

$= -\log\Delta + h(X)$

8.3.1 If density $f(x)$ is Riemann Integrable then $H(X^\Delta) + \log\Delta \to h(f) = h(X)$ as $\Delta \to 0$

Joint and Conditional Differential Entropy

Theorem 8.4.1 Entropy of a multivariate normal distribution

Let $X_1, X_2, \cdots, X_n$ have a multivariate normal distribution with mean μ & covariance matrix K. Then.

$h(X_1, X_2, \cdots, X_n) = h(\mathcal{N}(\mu, K)) = \frac{1}{2}\log\left[(2\pi e)^n |K|\right]$ bits

$I(X^\Delta; Y^\Delta) = H(X^\Delta) - H(X^\Delta|Y^\Delta)$

$\approx (h(X) - \log\Delta) - (h(X|Y) - \log\Delta)$

$= I(X;Y)$

For two R.V. X and Y with Partitions P and Q of their range

$Pr([X]_P = i) = Pr(X \in P_i) = \int_{P_i} dF(x)$

$I(X;Y) = \sup_{P,Q} I([X]_P; [Y]_Q)$

Supremum over all finite partitions P, Q.

Hadamard's Inequality

If we let $\vec{X} \sim \mathcal{N}(0, K)$ be a multiplicative normal random variable,

$|K| \le \prod_{i=1}^{n} K_{ii}$

$h(aX) = h(X) + \log|a|$

Theorem Let $\vec{X} \in R^n$ have zero mean and covariance $K = E[XX^t]$ (i.e. $K_{ij} = E[X_i X_j]$)

Then $h(\vec{X}) \le \frac{1}{2}\log\left((2\pi e)^n |K|\right)$, with equality iff $\vec{X} \sim \mathcal{N}(0, K)$