CS3236

AY22/23 SEM 2

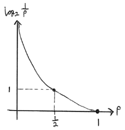
github/jovyntls

01. INFORMATION MEASURES

X is a d.r.v. with pmf P_X over an alphabet $\mathcal{X}$ (set of symbols)

- speed: **rate** $\rightarrow \frac{k}{n}$ (mapping k bits to n bits)

information of an event: $\psi(\cdot)$



$\psi(p) = \log_b \frac{1}{p}$ (for some $b > 0$)

- all choices of b are equivalent up to scaling by a universal constant
  - e.g. # of nats $= \log_e 2 \times$ # of bits

1. $\psi(p) \geq 0$ (**non-negativity**)
2. $\psi(1) = 0$ (**zero for definite events**)
3. if $p \leq p'$, then $\psi(p) \geq \psi(p')$ (**monotonicity**)
4. $\psi(p)$ in continuous in p (**continuity**)
5. $\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$ (**additivity under indep**)

if X takes N values with $\mathbf{p} = (p_1, \dots, p_N)$, only $\Phi(\mathbf{p}) = constant \times H(X)$ satisfies

1. if $p_i = \frac{1}{N}$, then $\Psi(\mathbf{p})$ is increasing in N (**uniform case**)
2. (**successive decisions**) $\Psi(p_1, \dots, p_N) = \Psi(p_1 + p_2, p_3, \dots, p_N) + (p_1 + p_2)\Psi(\frac{p_1}{p_1+p_2}, \frac{p_2}{p_1+p_2})$

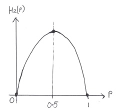
information of a random variable: $H(X)$

(Shannon) entropy $\rightarrow$ average information/uncertainty

$$H(X) = \mathbb{E}_{X \sim P_X}\left[\log_2 \frac{1}{P_X(X)}\right]$$
$$= \sum_x P_X(x) \log_2 \frac{1}{P_X(x)}$$

binary entropy function

$$H_2(p) = p \log_2 \frac{1}{p} + (1-p) \log_2 \frac{1}{1-p}$$



- binary source: $X \sim Bernoulli(p)$
  $\Rightarrow H(X) = H_2(p)$
- uniform source ($P_X(x) = \frac{1}{|\mathcal{X}|}$):
  $\Rightarrow H(X) = \mathbb{E}\left[\log_2 \frac{1}{1/|\mathcal{X}|}\right] = \log_2 |\mathcal{X}|$

variations

- **joint entropy** of two random variables (X, Y) $\rightarrow$

$$H(X,Y) = \mathbb{E}_{(X,Y) \sim P_{XY}}\left[\log_2 \frac{1}{P_{XY}(X,Y)}\right]$$
$$= \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{XY}(x,y)}$$

- **conditional entropy** of Y given X $\rightarrow$

$$H(Y|X) = \mathbb{E}_{(X,Y) \sim P_{XY}}\left[\log_2 \frac{1}{P_{Y|X}(Y|X)}\right]$$
$$= \sum_{x,y} P_{XY}(x,y) \log_2 \frac{1}{P_{Y|X}(y|x)}$$
$$= \sum_x P_X(x) H(Y|X = x)$$

- on average, $H(Y|X) \leq H(Y)$ but a *specific* outcome of X may increase uncertainty ($H(Y|X = i) > H(Y)$)

properties of entropy

1. $H(X) \geq 0$ (**non-negativity**) equality $\Leftrightarrow$ deterministic
2. $H(X) \leq \log_2 |\mathcal{X}|$ (**upper bound**)
   - equality $\iff X \sim Uniform(\mathcal{X})$
3. $H(X,Y) = H(X) + H(Y|X)$ (**chain rule**)
   $H(X,Y) = H(Y) + H(X|Y)$
   - conditioning: $H(X,Y|Z) = H(X|Z) + H(Y|X,Z)$
   - general chain rule:
     $H(X_1, \dots, X_n) = \sum_{i=1}^n H(X_i | X_1, \dots, X_{i-1})$
4. $H(X|Y) \leq H(X)$ (**conditioning reduces entropy**)
   - equality $\iff X$ and Y are independent
5. $H(X_1, \dots, X_n) \leq \sum_{i=1}^n H(X_i)$ (**sub-additivity**)
   - equality $\iff X$ and Y are independent

KL Divergence

Kullback-Leibler (KL) divergence or **relative entropy** is

$$D(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$$
$$= \mathbb{E}_{X \sim P}\left[\log_2 \frac{P(X)}{Q(X)}\right]$$

- $D(P||Q) \neq D(Q||P)$
- $D(P||Q) \geq 0$, equality $\iff P = Q$
  - *Proof.* $-D(P||Q) = -\sum_x P(x) \log_2 \frac{P(x)}{Q(x)}$
    $\leq \sum_x P(x)(\frac{Q(x)}{P(x)} - 1) = \sum_x Q(x) - \sum_x P(x) = 0$
    (using property that $\log \alpha \leq \alpha - 1$, equality iff $\alpha = 1$)
- $D(P_{XY}||P_X P_Y)$ = how far X, Y are from independent

Mutual Information

$$I(X;Y) = H(Y) - H(Y|X)$$
$$= H(X) - H(X|Y)$$
$$= H(X) + H(Y) - H(X,Y)$$
$$= D(P_{XY}||P_X \times P_Y)$$

- **mutual information**, $I(X;Y) \rightarrow$ the amount of information we learn about Y by observing X (on avg)
- **joint mutual information** $\rightarrow$
  $$I(X_1, X_2; Y_1, Y_2) = H(Y_1, Y_2) - H(Y_1, Y_2 | X_1, X_2)$$
- **conditional mutual information** $\rightarrow$
  $$I(X;Y|Z) = H(Y|Z) - H(Y|X,Z)$$
- if $X = Y$, then $I(X;Y) = H(X) = H(Y)$

properties of mutual information

1. $I(X;Y) = I(Y;X)$ (**symmetry**)
2. $I(X;Y) \geq 0$ (**non-negativity**)
   - equality $\iff X \perp Y$
3. $I(X;Y) \leq H(X) \leq \log_2 |\mathcal{X}|$ (**upper bounds**)
   $I(X;Y) \leq H(Y) \leq \log_2 |\mathcal{Y}|$
4. $I(X,Y;Z) = I(X;Z) + I(Y;Z|X)$ (**chain rule**)
   $$I(X_1, \dots, X_n; Y) = \sum_{i=1}^n I(X_i; Y | X_1, \dots, X_{i-1})$$
   $$= I(X_1;Y) + I(X_2;Y|X_1) + \dots$$
5. (**partial sub-additivity**)
   $$I(X_1, \dots, X_n; Y_1, \dots, Y_n) \leq \sum_{i=1}^n I(X_i; Y_i)$$
   if $(Y_1, \dots, Y_n)$ are conditionally indep given $(X_1, \dots, X_n)$, and Y_i depends on $(X_1, \dots, X_n)$ only through X_i
6. (**data-processing inequality**)
   $I(X;Z) \leq I(X;Y)$ if $X \rightarrow Y \rightarrow Z$
   variation: $I(X;Z) \leq I(Y;Z)$ if $X \rightarrow Y \rightarrow Z$
   $I(W;Z) \leq I(X;Y)$ if $W \rightarrow X \rightarrow Y \rightarrow Z$
   - holds if Z depends on (X,Y) only through Y (i.e. $X \rightarrow Y \rightarrow Z$ forms a **Markov chain** / X and Z are conditionally indep given Y)

02. SYMBOL-WISE SOURCE CODING

maps $x \in \mathcal{X}$ to binary sequence $C(x)$ of length $\ell(x)$.

average length of a code $C(\cdot)$,
$$L(C) = \sum_{x \in \mathcal{X}} P_X(x) \ell(x)$$

decodability conditions of $C(\cdot)$

- **nonsingular property** $\rightarrow C(x) \neq C(x') \iff x \neq x'$
- **uniquely decodable** $\rightarrow$ no 2 sequences of symbols in $\mathcal{X}$ are coded to the same sequence. $\Rightarrow x_1, \dots, x_n$ can be always uniquely identified from $C(x_1) \dots C(x_n)$
- **prefix-free** (instantaneous) $\rightarrow$ no codeword is prefix of other

Kraft's Inequality

Kraft's inequality

if $C(\cdot)$ is *prefix-free*, then $\displaystyle\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$

- *Proof.* represent the codewords by a binary tree. If there is a codeword at some point in the tree, there are no codewords further down the tree. probability of branching to a codeword $= 2^{-\ell(x)}$ and sum of probabilities cannot exceed 1
- **existence property** $\rightarrow$ if a given set of integers $\{\ell(x)\}_{x \in \mathcal{X}}$ satisfies $\sum_{x \in \mathcal{X}} 2^{-\ell(x)} \leq 1$, we can construct a *prefix-free* code that maps each $x \in \mathcal{X}$ to a codeword of length $\ell(x)$.

entropy bound

entropy bound (fundamental compression limit)
expected length, $L(C) \geq H(X)$
with equality $\iff P_X(x) = 2^{-\ell(x)} \quad \forall x \in \mathcal{X}$

- if all probabilities are negative powers of 2, optimal code

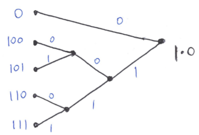
Shannon-Fano Code

$$\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$$

- $L(C)$ satisfies $H(X) \leq L(C) < H(X) + 1$
- **Kraft's inequality** holds - hence we can construct a prefix-free code with these lengths (**Existence property**)

mismatched case: if the true distribution is P_X, but lengths are chosen by Q_X, then the Shannon-Fano code satisfies
$$H(X) + D(P_X||Q_X) \leq L(C) \leq H(X) + D(P_X||Q_X) + 1$$

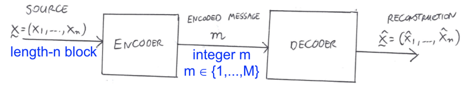
Huffman Code



- no uniquely decodable symbol code can achieve a smaller length $L(C)$ than the Huffman code.
  - always prefix-free
  - satisfies average length bound: $H(X) \leq L(C) < H(X) + 1$
- extension: using blocks of n letters; Huffman coding with $\mathcal{X}^n$
  $nH(X) \leq L(C) < nH(X) + 1$
  $\Rightarrow H(X) \leq$ avg. length per symbol $\leq H(X) + \frac{1}{n}$
  - ✓ exploits *memory*, better guarantee (even independent)
  - × but it's harder to accurately know $P_{X_1 \dots X_n}$
  - × alphabet size increases to $|\mathcal{X}|^n \Rightarrow$ expensive to sort

03. BLOCK-WISE SOURCE CODING

- **discrete memoryless source**
  - i.i.d. sequence $\mathbf{X} = (X_1, \dots, X_n)$
  - $\mathbf{X}$ has **pmf** $P_{\mathbf{X}}(\mathbf{x}) = \Pi_{i=1}^n P_X(x_i)$ (*memoryless*)
- length-n block $\mathbf{X} \Rightarrow$ integer $m \in \{1, \dots, M\}$



- **error** $\rightarrow P_e = \mathbb{P}[\hat{\mathbf{X}} \neq \mathbf{X}] = \sum_{\mathbf{x}: \text{DEC}(\text{ENC}(\mathbf{x})) \neq \mathbf{x}} P_{\mathbf{X}}(\mathbf{x})$
- **rate** $\rightarrow R = \frac{1}{n} \log_2 M$ (compressed length $k = \log_2 M$)
  - lower rate = more compression ($M = 2^{nR}$)
  - $R \leq H(X) + \epsilon$
- **fixed length source coding theorem** $\rightarrow n, R, P_e$ tradeoff
  - (**achievability**) if $R > H(X)$, then for any $\epsilon > 0$, we can get $P_e \leq \epsilon$ for large enough n
  - (**converse**) if $R < H(X)$, then $\exists \epsilon > 0$ s.t. $\forall n, P_e > \epsilon$

Typical Sequences

typical set, $\mathcal{T}_n(\epsilon) =$
$$\left\{\mathbf{x} \in \mathcal{X}^n : 2^{-n(H(X)+\epsilon)} \leq P_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}\right\}$$
where $\epsilon > 0$ is a (small) fixed constant
i.e. $P_{\mathbf{X}}(\mathbf{x}) \simeq 2^{-nH(\mathbf{X})}$

- only assign a (unique) $m \in \{1, \dots, M-1\}$ if $\mathbf{x} \in \mathcal{T}_n(\epsilon)$
  - choose $\mathbf{x}$ such that $\mathbb{P}[\mathbf{x} \in \mathcal{T}_n(\epsilon)] \simeq 1$
  - map $\mathbf{x} \notin \mathcal{T}_n(\epsilon)$ to dummy value M: $P_e = \mathbb{P}[\mathbf{X} \notin \mathcal{T}_X]$

properties of a typical set

1. (**equivalent definition**) $\mathbf{x} \in \mathcal{T}_n(\epsilon) \iff$
   $$H(X) - \epsilon \leq \frac{1}{n} \sum_{i=1}^n \log_2 \frac{1}{P_X(x_i)} \leq H(X) + \epsilon$$
   - $\mathbb{E}[\log P_X(x_i)] = H(X_i) = H(X)$
2. $\mathbb{P}[X \in \mathcal{T}_n(\epsilon)] \rightarrow 1$ as $n \rightarrow \infty$ (**high probability**)
3. $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X)+\epsilon)}$ (**cardinality upper bound**)
4. $|\mathcal{T}_n(\epsilon)| \geq (1 - o(1)) 2^{n(H(X)+\epsilon)}$
   where $o(1) \rightarrow 0$ as $n \rightarrow \infty$ (**cardinality lower bound**)

asymptotic equipartition property
as $n \rightarrow \infty$, the distribution is roughly uniform over $\mathcal{T}_n(\epsilon)$

- with *high probability* (2), a randomly drawn i.i.d. sequence $\mathbf{X}$ will be one of $\approx 2^{n(H(X))}$ sequences (3)(4), each of which has probability of $\approx 2^{-nH(X)}$ (definition of typical set)

Fano's Inequality

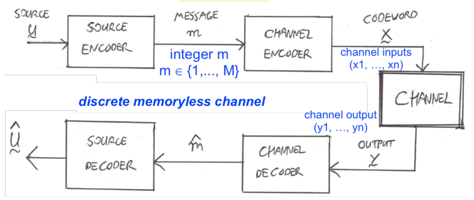
Fano's Inequality

$$H(X|\hat{X}) \leq H_2(P_e) + P_e \log_2(|\mathcal{X}| - 1)$$
$$\leq 1 + P_e \log_2 |\mathcal{X}|$$

- intuition: if estimate $\hat{\mathbf{X}}$ is accurate (small P_e), then
  $I(\mathbf{X}; \hat{\mathbf{X}}) \approx H(\mathbf{X}) = nH(X) \quad \Rightarrow H(\mathbf{X}|\hat{\mathbf{X}}) \approx 0$
  - $H_2(P_e)$ = uncertainty in "is $X = \hat{X}$"
  - $\log_2(|\mathcal{X}| - 1)$ = max uncertainty in the no case
- proves *converse* of **fixed length source coding theorem**
  $\Rightarrow P_e \geq \frac{1}{\log_2 |\mathcal{X}|}(H(X) - R - \frac{1}{n})$

04. CHANNEL CODING

- transmit $m \in \{1, \dots, M\}$ $\quad (M = 2^k = 2^{nR}$ for length-$k)$
- **codeword** $\mathbf{x}^{(m)} = (x_1^{(m)}, \dots, x_n^{(m)})$ transmitted over the channel in n uses; **codebook** $\mathcal{C} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$



- **memoryless** $\to$ outputs are (conditionally) independent: $\mathbb{P}[Y = y|X = x] = \Pi_{i=1}^n P_{Y|X}(y_i|x_i)$
- **error probability** $\to P_e = \mathbb{P}[\hat{m} \neq m]$
- **rate** $\to R = \frac{1}{n}\log_2 M$ $\quad (R \leq 1$ for binary channels)
  - higher rate = sending faster (vs source coding: lower)
- channel $P_{X|Y}$ is fixed; choose P_X by codebook generation

Channel Capacity

- **channel capacity**, $C \to$ maximum of all rates R such that, for any target error probability $\epsilon > 0$, $\exists$ block length n, codebook $\mathcal{C} = \{x^{(1)}, \dots, x^{(M)}\}$, such that $P_e \leq \epsilon$

  channel coding theorem $\to \mathbb{P}_e \leq \epsilon \Leftrightarrow$ rate $< C$
  where the capacity $C = \max_{P_x} I(X;Y)$
- (**achievability**) for any $R < C$, there exists a code of rate $\geq R$ with arbitrarily small P_e
- (**converse**) for any $R > C$, any code rate $\geq R$ cannot have arbitrarily small P_e (for any codebook)
- noiseless/deterministic channel: $C = \max_{P_X} H(X) = 1$
- binary symmetric channel: $C = 1 - H_2(\delta)$
- binary erasure channel ($\mathcal{Y} = \{0, 1, e\}$, $\mathbb{P}[\text{erasure}] = \epsilon$): $C = \max_{P_X}(H(X) - \epsilon H(X)) = 1 - \epsilon$

Jointly Typical Sequences

a pair of $(\mathbf{x}, \mathbf{y})$ of length-n input and output sequences is **jointly typical** wrt a joint distribution P_{XY} if

$$2^{-n(H(X)+\epsilon)} \leq P_{\mathbf{X}}(\mathbf{x}) \leq 2^{-n(H(X)-\epsilon)}$$
$$2^{-n(H(Y)+\epsilon)} \leq P_{\mathbf{Y}}(\mathbf{y}) \leq 2^{-n(H(Y)-\epsilon)}$$
$$2^{-n(H(X,Y)+\epsilon)} \leq P_{\mathbf{XY}}(\mathbf{x}, \mathbf{y}) \leq 2^{-n(H(X,Y)-\epsilon)}$$

- aka: the X seq, Y seq, and joint (X, Y) seq are all typical
- **jointly typical set**, $\mathcal{T}_n(\epsilon) \to$ set of all jointly typical seqs

properties

1. (**equivalent definition**) $\quad (\mathbf{x}, \mathbf{y}) \in \mathcal{T}_n(\epsilon) \iff$
$$H(X) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n \log_2 \frac{1}{P_X(x_i)} \leq H(X) + \epsilon$$
$$H(Y) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n \log_2 \frac{1}{P_Y(y_i)} \leq H(Y) + \epsilon$$
$$H(X,Y) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n \log_2 \frac{1}{P_Y(x_i, y_i)} \leq H(X,Y) + \epsilon$$
2. (**high probability**) $\mathbb{P}[(\mathbf{X}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$
3. (**cardinality upper bound**) $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X,Y)+\epsilon)}$
4. (**probability for independent sequences**) if $(\mathbf{X}', \mathbf{Y}') \sim P_X(\mathbf{x}')P_Y(\mathbf{y}')$ are independent copies of $(\mathbf{X}, \mathbf{Y})$, then the probability of joint typicality is $\mathbb{P}[(\mathbf{X}', \mathbf{Y}') \in \mathcal{T}_n(\epsilon)] \leq 2^{-n(I(X;Y)-3\epsilon)}$
   - X and Y drawn independently (instead of joint distribution) $\Rightarrow$ much lower probability of being typical

Achievability via Random Coding

- for a random $\mathcal{C}$, show $\mathbb{E}[P_e(\mathcal{C})] \leq \epsilon$ (thus $\exists \mathcal{C}$ with $P_e \leq \epsilon$)
  - if $!\exists m'$ s.t. $(\mathbf{X}^{(m')}, \mathbf{Y}) \in \mathcal{T}_n(\epsilon)$, set $\hat{m} = m'$
- $P_e \leq \delta_n + M \times 2^{-n(I(X;Y)-3\epsilon)}$
- arbitrarily small P_e for any R close to $I(X;Y)$ (close to C)

Converse via Fano's Inequality

- note that $m \to \mathbf{X} \to \mathbf{Y} \to \hat{m}$ forms a **Markov chain**

$I(m; \hat{m}) \leq I(\mathbf{X}; \mathbf{Y}) \leq nC \quad \Rightarrow P_e \geq 1 - \frac{nC + 1}{nR}$

05. CONTINUOUS-ALPHABET CH

Differential Entropy

differential entropy of a continuous r.v. X with pdf f_X

$$h(X) = \mathbb{E}_{f_X}\left[\log_2 \frac{1}{f_X(X)}\right]$$
$$= \int_{\mathbb{R}} f_X(x) \log_2 \frac{1}{f_X(x)}\, dx$$

joint version, $h(X, Y) = \mathbb{E}\left[\log_2 \frac{1}{f_{XY}(x,y)}\right]$

conditional version,

$$h(Y|X) = \mathbb{E}_{(X,Y)\sim f_{XY}}\left[\log_2 \frac{1}{f_{Y|X}(Y|X)}\right]$$
$$= \int_{\mathbb{R}} f_X(x) H(Y|X = x)\, dx$$

where (X, Y) have a joint density function
$f_{XY}(x, y) = f_X(x) f_{Y|X}(y|x)$

properties that still hold

- (**chain rule**)
$h(X_1, \dots, X_n) = \sum_{i=1}^n h(X_i|X_1, \dots, X_{i-1})$
- (**conditioning reduces entropy**) $\quad h(X|Y) \leq h(X)$
- (**sub-additivity**) $\quad h(X_1, \dots, X_n) \leq \sum_{i=1}^n h(X_i)$
- $h(X) = h(X + c)$ for some constant c

properties of entropy that *do not* hold

- non-negativity: we can have $h(X) < 0$
- invariance under 1-1 transformations: $h(X) \neq h(\psi(X))$
- *counterexample:* $Y = cX$. $\quad$ then $f_Y(y) = \frac{1}{|c|} f_X(\frac{y}{c})$,
  - which gives $h(Y) = \mathbb{E}[\log_2 \frac{1}{f_Y(y)}]$
  $= \mathbb{E}[\log_2 \frac{|c|}{f_X(Y/c)}] = \log_2|c| + h(X) \quad \neq h(\psi(X))$
  - violation of non-negativity: $\log_2|c| \to \infty$ as $c \to 0$

examples

- **Uniform**$(a, b) \Rightarrow h(X) = \mathbb{E}[\log_2 \frac{1}{f_X(x)}] = \log_2(b - a)$
- **gaussian** $X \sim N(\mu, \sigma^2) \Rightarrow h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$

Mutual information & KL Divergence

mutual information

$$I(X;Y) = h(Y) - h(Y|X)$$
$$= h(X) - h(X|Y)$$
$$= D(f_{XY} || f_X \times f_Y)$$
$$= \mathbb{E}_{f_{XY}}\left[\log_2 \frac{f_{XY}(x,y)}{f_X(x)f_Y(y)}\right]$$

KL divergence, $D(f||g) = \int_{\mathbb{R}} f(x)\log_2 \frac{f(x)}{g(x)}\, dx$

properties: all hold

- $I(X;Y) = I(\psi(X); \phi(Y))$ for invertible $\psi(\cdot)$ and $\phi(\cdot)$

Gaussian Random Variables

if $X \sim N(\mu, \sigma^2)$, then $h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$

maximum entropy property

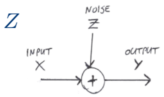
$$h(X) \leq \frac{1}{2}\log_2(2\pi e Var[X])$$

with equality $\iff X$ is Gaussian

- for a given *variance:* gaussian r.v. has highest entropy $h(\cdot)$
- for given *values* ($X \in [a, b]$): uniform maximises $h(\cdot)$

Gaussian Channel

a continuous channel is described by conditional pdf $f_{Y|X}$



- **additive noise channels** $\to Y = X + Z$
  - Z is a noise term independent of X
  - $f_{Y|X}(y|x) = f_Z(y - x)$
- **additive white Gaussian noise (AWGN) channel** $\to$ $Z \sim N(0, \sigma^2)$ for some noise variance $\sigma^2 > 0$
- **power constraint:** $\mathbb{E}[X^2] \leq P$

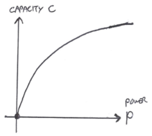
Channel Capacity

- channel capacity $C(P)$ is same as DMC, but codebooks are constrained to satisfy average power constraint
- for AWGN, capacity-achieving f_X is gaussian: $N(0, P)$

  AWGN capacity $\to C(P) = \frac{1}{2}\log_2(1 + \frac{P}{\sigma^2})$
  general $\to C(P) = \max_{f_X: \mathbb{E}_{f_X}[X^2] \leq P} I(X;Y)$

properties of Gaussian channel capacity

- depends on P, σ^2 only through *signal-to-noise ratio* $\frac{P}{\sigma^2}$
- $P = 0 \Rightarrow SNR = 0 \Rightarrow C = 0$
- as $\sigma^2 \to 0$ for fixed P, then $SNR \to \infty, C \to \infty$



diminishing returns of increasing P
- small $\frac{P}{\sigma^2}$, $C(P) \approx \frac{P}{2\sigma^2}$
  $\Rightarrow$ almost proportional to P
- large $\frac{P}{\sigma^2}$, $C(P) \approx \frac{1}{2}\log_2\frac{P}{\sigma^2}$
  $\Rightarrow$ diminishing returns

06. PRACTICAL CHANNEL CODES

$\mathbf{u}_{\in\{0,1\}^k} = m_{\in\{1,\dots,M\}} \Rightarrow \mathbf{x}^{(m)} \Rightarrow \mathbf{y}, P_e = \mathbb{P}[\hat{m} \neq m]$

- **parity check** $\to c = b_1 \oplus \dots \oplus b_m$
  - $\Rightarrow$ ensures an even number of 1's in the sequence
- channel: $\mathbf{y} = \mathbf{x} \oplus \mathbf{z}$; $\quad \mathbf{z} \in \{0, 1\}^n$ indicates flipped bits
- **rate** = $\frac{k}{n} = \frac{1}{n}\log_2(\#\text{messages})$ since $M = 2^k$

Linear Codes

- **linear code** $\to$ is comprised of parity checks
  - $\oplus$ of any 2 codewords is another valid codeword
  - if $\mathbf{u}, \mathbf{u}'$ correspond to codewords $\mathbf{x} = \mathbf{u}\mathbf{G}, \mathbf{x}' = \mathbf{u}'\mathbf{G}$, then $\mathbf{x} \oplus \mathbf{x}'$ is also a codeword

$$\mathbf{x} \oplus \mathbf{x}' = \mathbf{u}\mathbf{G} \oplus \mathbf{u}'\mathbf{G} = (\mathbf{u} \oplus \mathbf{u}')\mathbf{G}$$

- **systematic** parity-check code $\to$ the first k bits of $\mathbf{x}$ are always the original k bits; remaining $n - k$ are parity checks

$$x_i = \begin{cases} u_i & \text{if } i = 1, \dots, k, \\ \bigoplus_{j=1}^k u_j g_{j,i} & \text{if } i = k+1, \dots, n \end{cases}$$

- **general** parity-check code $\to$ all n codeword bits may be arbitrary parity checks: $\bigoplus_{j=1}^k u_j g_{j,i}$ for $i = 1, \dots, n$

generator matrix

$\mathbf{x}$ is a codeword $\iff \mathbf{x} = \mathbf{u}\mathbf{G}$ (for some $\mathbf{u}$)

generator matrix (general)

$$\mathbf{G} = \begin{bmatrix} g_{1,1} & g_{1,2} & \cdots & g_{1,n} \\ g_{2,1} & g_{2,2} & \cdots & g_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{k,1} & g_{k,2} & \cdots & g_{k,n} \end{bmatrix}$$

single-parity-check:
$$\mathbf{G}_{\text{parity}} = \begin{bmatrix} 1&0&0&0&1 \\ 0&1&0&0&1 \\ 0&0&1&0&1 \\ 0&0&0&1&1 \end{bmatrix}$$

Hamming code:
$$\mathbf{G} = \begin{bmatrix} 1&0&0&0&1&1&0 \\ 0&1&0&0&1&0&1 \\ 0&0&1&0&0&1&1 \\ 0&0&0&1&1&1&1 \end{bmatrix}$$

- **systematic**: leftmost $k \times k$ sub-matrix = identity matrix I_k
- codewords are linear combinations of the rows of $\mathbf{G}$
- $g_{j,i} = 1 \iff$ the j-th bit is used in the i-th parity check

parity-check matrix

$\mathbf{x}\mathbf{H} = \mathbf{0} \iff \mathbf{x}$ is a valid codeword

$$\mathbf{G} = [\ \mathbf{I}_k \ \ \mathbf{P}\ ] \implies \mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I}_{n-k} \end{bmatrix}$$

parity-check matrix (systematic)
an $n \times (n - k)$ matrix

$$\mathbf{H} = \begin{bmatrix} g_{1,k+1} & g_{1,k+2} & \cdots & g_{1,n} \\ g_{2,k+1} & g_{2,k+2} & \cdots & g_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ g_{k,k+1} & g_{k,k+2} & \cdots & g_{k,n} \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

single-parity-check:
$$\mathbf{H}_{\text{parity}} = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

Hamming code:
$$\mathbf{H}_{\text{Hamming}} = \begin{bmatrix} 1&1&0 \\ 1&0&1 \\ 0&1&1 \\ 1&1&1 \\ 1&0&0 \\ 0&1&0 \\ 0&0&1 \end{bmatrix}$$

- for $\mathbf{y} = \mathbf{x} \oplus \mathbf{z}$ ($\mathbf{z}$ is noise),
$\mathbf{y}\mathbf{H} = (\mathbf{x} \oplus \mathbf{z})\mathbf{H} = (\mathbf{x}\mathbf{H}) \oplus (\mathbf{z}\mathbf{H}) = \mathbf{z}\mathbf{H}$
- $\left(\bigoplus_{j=1}^k x_j g_{j,i}\right) \oplus x_i = 0$ since $x_i = \bigoplus_{j=1}^k x_j g_{j,i}$ for $i \geq k+1$

Distance Properties

- **Hamming distance** $\to$ number of differing positions
  - $d_H(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^n \mathbb{1}\{x_i \neq x_i'\}$
- **minimum distance** $\to d_{\min} = \min_{\mathbf{x} \in \mathcal{C}, \mathbf{x}' \in \mathcal{C}: \mathbf{x} \neq \mathbf{x}'} d_H(\mathbf{x}, \mathbf{x}')$
  - correct $\leq d_{\min} - 1$ erasures and $\leq \frac{d_{\min} - 1}{2}$ bit flips
- **weight** $\to w(\mathbf{x}) = \sum_{i=1}^n \mathbb{1}\{w_i = 1\}$ $\quad$ (number of 1's)
  - $w(\mathbf{x}) = \sum_{i=1}^n \mathbb{1}\{w_i = 1\}$
  - for linear codes, min distance = min weight
    - $d_{\min} = \min_{\mathbf{x} \in \mathcal{C}: \mathbf{x} \neq 0} w(\mathbf{x})$ $\quad$ for $d_{\min} > 0$

Minimum Distance Decoding

maximum likelihood decoding

for any channel $P_{\mathbf{Y}|\mathbf{X}}$ and any codebook $\{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(M)}\}$,

maximum-likelihood (ML) decoder $\to$ minimises P_e

$$\hat{m} = \arg\max_{j=1,\dots,M} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(j)})$$

for BSC, ML decoding is equivalent to
minimum (Hamming) distance decoding

$$\arg\max_{j=1,\dots,M} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(j)}) = \arg\min_{j=1,\dots,M} d_H(\mathbf{x}^{(j)}, \mathbf{y})$$

syndrome decoding

for linear codes for the BSC,

- **syndrome** $\to \mathbf{S} = \mathbf{z}\mathbf{H} = \mathbf{y}\mathbf{H}$ $\quad \Rightarrow 1 \times (n - k)$ vector
- the *minimum-distance codeword* to $\mathbf{y}$ is
  1. $\hat{\mathbf{z}} = \arg\min_{\mathbf{z}': \mathbf{z}'\mathbf{H} = \mathbf{S}} w(\mathbf{z}')$ $\quad$ (i.e. $\mathbf{z}'$ with fewest 1's)
  2. $\hat{\mathbf{x}} = \mathbf{y} \oplus \hat{\mathbf{z}}$
- *Proof.* define $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} \oplus \mathbf{y} \Rightarrow d_H(\mathbf{x}^{(i)} \oplus \mathbf{y}) = w(\mathbf{z}^{(i)})$

Hamming code

- maps length 4 $\Rightarrow$ length 7: sum of bits in each circle is even