Info of event

Probability of an event happening is $Pr(A)$. The amount of information we learn from A can be $(information)(A)$, or $\psi(Pr(A))$.

Non-negative. No such thing is negative information.

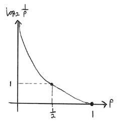
Zero for definite events. If we know an event will happen, we wont learn anything from it.

Monotone. The less likely an event is to happen, the more information we can learn from it. $p \leq p', \psi(p) \geq \psi(p')$

Continuity. Small changes in probability of event occurring does not cause big changes in info we learn.

Addivity under independence. $\psi(p_1 p_2) = \psi(p_1) + \psi(p_2)$. If events are independent, the information we learn from each event is independent of each other. So the total information learnt is the sum.

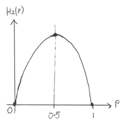
Axiom satisfaction. $\psi(p) = log_b \frac{1}{p}$ where $b > 0$ satisfies all 5. b tells us how is information measured. $b = 2$ means measured in bits. $b = e$ means measured in nats.



Entropy. Discrete random variables, then the probability mass function is $P_X(x) = Pr(X = x)$. If $X = X$, then we learnt $\psi(P_X(x))$.

Shannon Entropy. Amount of information (after X) or uncertainty (before X).

$$\psi(P_X(x)) = \sum_x P_X(x) log_2 \frac{1}{P_X(x)}$$



Binary Entropy function. If X is bernoulli, then

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

Uniform Entropy function. If $P_X(x)$ is the same for all x, then $H(X) = log_2(|\mathcal{X}|)$

Successive decisions. First draw from a distribution that doesnt resolve 2 symbols and then draw from another if we need to resolve it.

$$\psi(p_1, \ldots, p_N) = \psi(p_1 + p_2, p_3, \ldots) + (p_1 + p_2)\psi(\frac{p_1}{p_1 + p_2}, \frac{p_2}{p_1 + p_2})$$

Joint Entropy.

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

Conditional Entropy. Amount of info we get from the next event after observing another. If X uniquely determines Y, then $H(Y|X) = 0$.

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

$\sum_x P_X(x)$ is the average over x. $H(Y|X = x)$ is the amount of info we can learn from Y given that we have seen that $X = x$.

Non-negative. Equality when X is deterministic. $H(X) \geq 0$

Upper-bound. $H(X) \leq log_2(|\mathcal{X}|)$ Uniform is most uncertain since we need to randomly guess.

Chain Rule. $H(X, Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$ Overall uncertainty is the sum of the uncertainty X and the remaining uncertainty Y after seeing X. General case: $H(X_1, \ldots, X_n) = \sum_{i=1}^n H(X_i|X_1, \ldots, X_{i-1})$

Conditioning reduces entropy. $H(X|Y) \leq H(X)$. Equality holds if X, Y are independent. Upperbound is $H(X)$ since having additional information cannot increase uncertainty on average. **Possible for the following:** $H(X|Y = y) > H(X)$

Subaddivity. Equality holds if $X_1, \ldots, X_n$ is independent. $H(X_1, \ldots, X_n) \leq \sum_{i=1}^n H(X_i)$

Relative Entropy. Measures how similar the 2 distributions are. Equality holds when they are the same distribution since no differences. $D(P||Q) \neq D(Q||P)$. No triangle inequality as well. Equality holds if $P = Q$. Uses $log_e \alpha \leq \alpha - 1$

$$D(P||Q) = \sum_x P(x) \log_2 \frac{P(x)}{Q(x)} = \mathbb{E}_{X\sim P}\left[\log_2 \frac{P(X)}{Q(X)}\right] \geq 0$$

Mutual Information

How much information X gives about Y.

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

Joint. $I(X_1, X_2; Y_1, Y_2)$ Similar to above but $X \leftarrow (X_1, X_2)$ and $Y \leftarrow (Y_1, Y_2)$

Conditional. $I(X; Y|Z)$. Conditions **both** X and Y on Z. $I(X; Y|Z) = H(Y|X, Z) - \sum_z P_Z(z) I(X; Y|Z = z)$

Independence. If X, Y are independent, then $H(Y|X) = H(Y)$ so $I(X; Y) = 0$.

Equivalence. If $Y = X$, then $H(Y|X) = 0$ so $I(X; Y) = H(Y)$

Symmetry. $I(X; Y) = I(Y; X)$. X, Y reveal equal amount of information about each other.

Non-negativity. $I(X; Y) \geq 0$. Equality only when independent. 1 random variable cannot tell negative information about the other. $D(P_{XY}||P_X \times P_Y) \geq 0$ iff $P_{X,Y} = P_X \times P_Y = P_Y$ which also implies independence.

Upper bound. $I(X; Y) \leq H(X)$ equality iff $H(X|Y) = 0$ iff X is deterministic given Y. Vice versa for Y. Cannot reveal more than prior uncertainty.

Chain Rule. $I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y|X)$ Total information of X_1, X_2 is the sum of information gained from X_1 and information gained from X_2 given X_1. General case $I(X_1, \ldots, X_n; Y) = \sum_n I(X_i; Y|X_1, \ldots X_{i-1})$

Data processing inequality. If Z depends on (X, Y) **only** via Y (Markov chain, $X \to Y \to Z$) and equivalent to the statement X, Z are conditionally independent given Y, then $I(X; Z) \leq I(X; Y)$. Post processing cannot increase info about X. Equality holds when $I(X; Y|Z) = 0$. This means that all information that Y reveals about X is revealed by Z alone. Processing Y (to produce Z) cannot increase information available regarding X.

Partial sub-addivity. $I(X_1, \ldots X_n; Y_1, \ldots Y_n)$ can be smaller or larger than $\sum_n I(X_i; Y_i)$ but typically is $\leq$

Larger or equal. When $X_1 \ldots, X_n$ are mutually independent.

Smaller or equal. If $Y_1, \ldots, Y_n$ are conditionally independent given $X_1, \ldots, X_n$ and Y_i depends on $X_1, \ldots, X_n$ only through X_i.

Symbol Source

Higher $P_X(x)$ leads to a shorter encoded length.

Average code length. $L(C) = \sum_x P_X(x) l(x)$ $l(x)$ is the length of a sequence of binary code for x.

Nonsingular property. $C(x) \neq C(x')$ if $x \neq x'$.

Uniquely Decodable. No 2 sequences (of equal/unequal lengths) of symbols in $\mathcal{X}$ are coded to the same concatenated binary sequence. $x_1, \ldots, x_n$ always uniquely identified from the string $C(x_1) \ldots C(x_n)$.

Prefix-free. No $C(x)$ is a prefix of any $C(x')$ where $x \neq x'$. Its not possible to append more bits to some $C(x)$ in order to produce some other $C(x')$.

Krafts Inequality. Any prefix-free (any uniquely decodable code) code that maps each $x \in \mathcal{X}$ to a code word of length $l(x)$ must satisfy $\sum_x 2^{-l(x)} \leq 1$.

Proof. In a binary tree, each node is a code word. If there is a codeword that is used at some point in the tree, then there are no codewords further down the tree. The probability of getting any of the codeword is $2^{-l(x)}$. Since total probability of hitting codewords cannot exceed 1, so the sum of them must be ≤ 1.

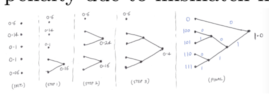
Existence Property. If there are lengths that satisfies kraft's inequality, then its possible to construct prefix-free code that maps each $x \in (X)$ to a codeword of length $l(x)$.

Entropy Bound. Fundamental compression limit(can never get an average length smaller than entropy). For $X \sim P_X$ and any prefix free code, the expected length satisfies $L(C) \geq H(X)$. Equality iff $P_X(x) = 2^{-l(x)}$ for all $x \in \mathcal{X}$

Shannon-Fano Code. $\ell(x) = \left\lceil \log_2 \frac{1}{P_X(x)} \right\rceil$ Satisfies Krafts inequality because of the ceiling. So its possible to create prefix free codes with these lengths.

Average length. Average length satisfies $H(X) \leq L(C) \leq H(X) + 1$

Unknown distribution. Apply Shannon-Fano code to Q_x but true distribution is P_x. We get a mismatch case so we have $H(X) + D(P_X||Q_X) \leq L(C) \leq H(X) + D(P_X||Q(X)) + 1$. Relative entropy is the penalty due to mismatch here.



Huffman Code. List the symbols from highest probability to lowest.

Kraft's inequality. Does not violate Kraft's inequality since its always prefix-free; always satisfies with equality.

Theorem. No uniquely decodable code can achieve a smaller average length $L(C)$ than the Huffman code. Always optimal.

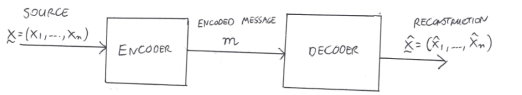
Properties. Don't exploit correlations/memory (dependence between subsequent symbols).

Solution. Code cover the blocks of letters instead. Can exploit statistics of groups of letters. Even if source has independent letters, this can help.

Shannon-Fano Guarantee. $H(X) < \frac{1}{n} L(C) < H(X) + \frac{1}{n}$ Result of normalising is the average length per letter.

Disadvantage. Determining the distribution of $P_{X_1} \ldots P_{X_n}$ accurately is hard. Sorting the probabilities become computationally difficult.

Block Source



Output of decoder is an interger.

Discrete memoryless sources. Alphabet $\mathcal{X}$ is finite and $P_X(x) = \Pi_{i=1}^n P_X(x_i)$ where source symbols are iid on some distribution P_X.

Error probability. $P_e = P[\hat{X} \neq X]$

Rate. $R = \frac{1}{n} log_2 M$ represents number of bits per source symbol used to represent the encoded value m. Lower rate means more compressed the source sequence.

Fixed-length source coding thm. For any discrete memoryless source with per-symbol distribution P_X.

Achievability. If $R > H(X)$ then for any $\epsilon > 0$ there is a (sufficiently large) block length n and a source code (encoder and decoder) of rate R st $P_e \leq \epsilon$.

Converse. If $R < H(X)$ then there is $\epsilon > 0$ such that every source code of rate R has $P_e > \epsilon$ regardless of block length. (P_e cannot be arbitrarily small).

Typical Sequences.

$$\mathcal{T}_n(\epsilon) = \{x \in \mathcal{X} : P_X(x) = 2^{-n(H(x)+\alpha)}\}$$

Where $\alpha \in [-\epsilon, \epsilon]$ and $\epsilon > 0$ is a fixed small constant. Typicality only interested in the probability of the sequence and not the sequence itself.

Equivalence. $x \in \mathcal{T}_n(\epsilon)$ iff $\frac{1}{n}\sum_n^{i=1} log_2 \frac{1}{P_X(x_i)} = H(X) + \alpha$, where x_i is ith entry of x and $\alpha \in [-\epsilon, \epsilon]$

High probability. $P[X \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$. Probability that some sequence exists in typical set increases as block length gets very large.

Cardinality upper bound. $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X)+\epsilon)}$

Cardinality lower bound. $|\mathcal{T}_n(\epsilon)| \geq (1 - o(1))2^{n(H(X)-\epsilon)}$ where $o(1) \to 0$ as $n \to \infty$.

Asymptotic Equipartition. As $n \to \infty$ distribution is roughly uniform over $\mathcal{T}_n(\epsilon)$

Interpretation. With high probability, a randomly drawn iid seqeunce X will be 1 of roughly $2^{nH(X)}$ sequences, each of which has probability $2^{-nH(X)}$

Fano Inequality. If $H(X|\hat{X})$ is large,then $\hat{X}$ does not reveal much info about X, so P_e must not be too small. Otherwise, knowing $\hat{X}$ tells us alot about X. Accurate estimation implies $H(X|\hat{X}) \approx 0$.

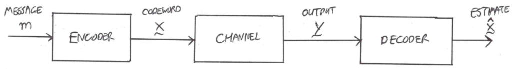
$$H(X|\hat{X}) \leq H_2(P_e) + P_e log_2(|\mathcal{X} - 1)$$

$H_2(P_e)$ is the uncertainty that $X = \hat{X}$. They differ P_e of the time and the remaining uncertainty is at most $log_2(|\mathcal{X} - 1)$ since uniform distribution maximises entropy. Proves converse of fixed-length source coding thm.

$$P_e \geq \frac{1}{log_2|\mathcal{X}}(H(X) - R - \frac{1}{n})$$

Channel Coding

Transmit a message $m \in \{1, \dots, M\}$, and if the output is k bits, then $M = 2^k$ and map each output to a unique index.



Codeword. $x^(m) = (x_1^{(m)}, \dots, x_n^{(m)})$ is the sequence when message is m. Transmitted in n uses.
Codebook. Collection of codewords. Known by encoder and decoder but only encoder knows m.
Discrete channel. Input/output alphabets are finite.
Memoryless. Transmitting several symbols in successive uses, the outputs are (conditionally) independent $P_{Y|X}(y|x) = \Pi_{i=1}^n P_{Y|X}(y_i|x_i)$
Error probability. $P_e = P[\hat{m} \neq m]$
Rate. Bits per channel use. $R = \frac{1}{n}log_2 M$. Number of messages is $M = 2^{nR}$. Higher rate means sending faster.
Channel Capacity. Maximum of all rates R, such that for any target error probability $\epsilon > 0$, there is a block length n and codebook $\mathcal{C}$ with $M = 2^{nR}$ codewords such that $P_e \leq \epsilon$. Highest rate st error probability can be made small at some (possibly large) block length.
Channel coding thm. Capacity of a dms is $C = \max_{P_x} I(X;Y)$
Achievability. For any $R < C$, there is a code of rate at least R with arbitrarily small P_e.
Converse. For any $R > C$, any code rate at least R cannot have arbitrarily small error probability.
Capacity-achieving input distribution. For a given channel $P_{Y|X}$, any input distribution P_X that maximises MI.
Noiseless channels. Output deterministically equals input. Then $C = \max_{P_x} I(X;Y) = 1$
Binary symmetric channel. Inputs are flipped with some probability δ. $C = 1 - H_2(\delta)$
Binary erasure channel. Erasure probability ϵ. Output equals input with probability $1 - \epsilon$ but erased with probability ϵ. $C = 1 - \epsilon$.
Joint typicality. Pair (x,y) of length-n input and output is joint typical wrt joint distribution P_{XY} if the following holds

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

X and Y sequence and joint (X,Y) sequences are all typical.
Jointly typical set. Set of all jointly typical sequences denoted by $\mathcal{T}_n(\epsilon)$
Equivalence. $(x,y) \in \mathcal{T}_n(\epsilon)$ iff following holds

$$H(X) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n log_2 \frac{1}{P_X(x)} \leq H(X) + \epsilon$$
$$H(Y) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n log_2 \frac{1}{P_Y(y)} \leq H(Y) + \epsilon$$
$$H(X,Y) - \epsilon \leq \frac{1}{n}\sum_{i=1}^n log_2 \frac{1}{P_{XY}(x_i, y_i)} \leq H(X,Y) + \epsilon$$

High probability. $P[(X,Y) \in \mathcal{T}_n(\epsilon)] \to 1$ as $n \to \infty$
Cardinality upper bound. $|\mathcal{T}_n(\epsilon)| \leq 2^{n(H(X,Y)+\epsilon)}$
Probability for indepenent seq. If $(X', Y')\widetilde{P}_X(x')P_Y(y')$ are independent copies of (X,Y) then probability of joint typicality is $P[(X', Y') \in \mathcal{T}_n(\epsilon)] \leq 2^{n(I(X,Y)+-3\epsilon)}$. If X', Y' are generated independently, then the further P_{XY} is from being independent, the less likely it is for those indepedent seqeuences to be jointly typical wrt P_{XY}
Achievability via random coding. Generate each symbol of each codeword randomly and independently according to some distribution.
Random-coding error probability. Calculate error probability given the message average over both randomness in the channel and random codebook.
Converse via Fano Inequality. To achieve small P_e, need amount of info that $\hat{m}$ reveals about m to be close to prior uncertainty in m. Then $P_e \geq 1 - \frac{C + \frac{1}{n}}{R}$

Continuous

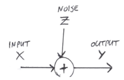
Differential Entropy.
$h(X) = \int_{\mathbb{R}} f_X(x) \log_2 \frac{1}{f_X(x)} dx$
Joint version. $h(X,Y) = \mathbb{E}\left[\log_2 \frac{1}{f_{XY}(x,y)}\right]$
Conditional Version.
$h(Y|X) = \int_{\mathbb{R}} f_X(x) h(Y|X = x) dx$
where (X,Y) have a joint density function $f_{XY}(x,y) = f_X(x) f_{Y|X}(y|x)$
Properties. Chain rule $(h(X_1, \dots, X_n) = \sum_{i=1}^n h(X_i|X_1, \dots, X_{i-1}))$, conditioning reduces entropy ($h(X|Y) \leq h(X)$), sub-additivity ($h(X_1, \dots, X_n) \leq \sum_{i=1}^n h(X_i)$), $h(X) = h(X + c)$ and $h(cX|Y) = h(X|Y) + log_2|c|$ for some constant c
Non-negativity. Possible for $h(X) < 0$
Invariance under 1-to-1 transformation. $h(X) \neq h(\psi(X))$
Counter example. If $Y = cX$ for some constant c, then density of a function gives $f_Y(y) = \frac{1}{|c|} f_X(\frac{y}{c})$. Substituting into differential entropy gives $h(Y) = h(X) + log_2|c|$. As $c \to 0$, $log_2|c| \to -\infty$.
Uniform RV. $h(X) = log_2(b - a)$ where X is RV over a **Uniform**(a,b) with $a < b$.
Univariate Guassian. $h(X) = \frac{1}{2}log_2(2\pi e \sigma^2)$. X is univariate gaussian over $N(\mu, \sigma^2)$
Relative entropy. $D(f||g) = \int_R f(x) log_2 \frac{f(x)}{g(x)} dx$
Retains all properties including non-negativity.
Mutual Info. Retains all properties including non-negativity.

$$\begin{aligned} I(X;Y) &= h(Y) - h(Y|X) \\ &= h(X) - h(X|Y) \\ &= D(f_{XY}||f_X \times f_Y) \\ &= E_{f_{XY}}[log_2 \frac{f_{XY}(x,y)}{f_X(x) f_Y(y)}] \end{aligned}$$

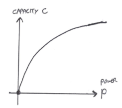
Retains all properties including non-negativity. $I(X;Y) = I(\phi(X); \varphi(Y))$ for invertible functions $\phi(\cdot), \varphi(\cdot)$
Maximum entropy property. Univariate case. For any rv X with density f_X and variance $Var[X]$, we have

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

Equality iff X is Gaussian. Gaussian maximises entropy for fixed variance. Not necessrily true if we fix other properties.
Gaussian Channel. Additive noise channels, $Y = X + Z$ where Z is a noise term **independent** of X. So $f_{Y|X}(y|x) = f_Z(y - x)$.

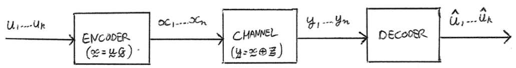


Additive white gaussian noise. Happens when $Z \sim N(0, \sigma^2)$ for some noise variance $\sigma^2 > 0$
Power constriant. $E[X^2] \leq P$ Require each codeword in a codebook to have power at most P averaged over block length. $\frac{1}{n}\sum_n^{i=1}(x_i^{(m)})^2 \leq P$ $\forall m \in \{1, \dots, M\}$ or $\frac{1}{M}\sum_{m=1}^M \frac{1}{n}\sum_n^{i=1}(x_i^{(m)})^2 \leq P$.
Channel capacity. Defined same as DMS but with codebooks constrained to satisfy average power constraint.
General noise models.
$C(P) = \max_{f_X : \mathbb{E}_{f_X}[X^2] \leq P} I(X;Y)$



Gaussian. With power constraint P and noise variance σ^2, capacity-achieving f_X is Gaussian, namely $N(0, P)$, $C(P) = \frac{1}{2}\log_2(1 + \frac{P}{\sigma^2})$
Properties. Depends on P, σ^2 through signal-to-noise ratio $\frac{P}{\sigma^2}$. Equals 0 when $P = 0$. When $\frac{P}{\sigma^2} \to 0$, we have $C(P) \approx \frac{P}{2\sigma^2}$. When $\frac{P}{\sigma^2} \to \infty$, have $C(P) \approx \frac{1}{2}log_2 \frac{P}{\sigma^2}$

Practical Channel Codes



Parity Check. For a sequence of bits, have an extra bit at the end equaling 1 if #1's is odd, 0 if even. $c = b_1 \oplus b_2 \dots b_m$
Linear Code Notation. $y = x \oplus z$ where $z \in \{0,1\}^n$ indicates which bits are flipped and $\oplus$ applied bitwise. Generic message replaced by message bits instead so $M = 2^k$ then rate will be $R = \frac{1}{n}log_2 M = \frac{k}{n}$.
Linear Code. Any code that comprises of parity checks is a linear code. Modulo sum of any 2 valid codewords is another vlaid codeword. if $\mathbf{u}, \mathbf{u}'$ correspond to codewords $\mathbf{x} = \mathbf{uG}, \mathbf{x}' = \mathbf{u}'\mathbf{G}$, then $\mathbf{x} \oplus \mathbf{x}'$ is also a codeword
Systematic code. First k bits out of n of $\mathbf{x}$ are the original k bits and the remaining $n - k$ bits 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 code. All n codeword bits may be arbitrary parity checks. Systematic code is a special case of this. $\bigoplus_{j=1}^k u_j g_{j,i}$ for $i = 1, \dots, n$
Generator matrix. $x = uG$. Rows are the index bits of u and cols are index bits of x.

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:

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

Parity Check Matrix. $xH = 0$ iff x is a valid codeword. H is used to check if x can be generated from any u.

$$\mathbf{xH} = \mathbf{0} \iff \mathbf{x} \text{ is a valid codeword}$$
$$\mathbf{G} = [\,\mathbf{I}_k \;\; \mathbf{P}\,] \implies \mathbf{H} = \begin{bmatrix} \mathbf{P} \\ \mathbf{I}_{n-k} \end{bmatrix}$$

systematic parity-check
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:

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

Rational. $yH = (x \oplus z)H = zH$. z indicates which bits got flipped. $\left(\bigoplus_{j=1}^k x_j g_{j,i}\right) \oplus x_i = 0$ since $x_i = \bigoplus_{j=1}^k x_j g_{j,i}$ for $i \geq k+1$
Hamming distance. Distance between 2 vectors is the number positions in which they differ. $d_H(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^n \mathbb{1}\{x_i \neq x_i'\}$
Minimum distance. Of a codebook of length-n codewords is $d_{\min} = \min_{\mathbf{x} \in \mathcal{C}, \mathbf{x}' \in \mathcal{C} : \mathbf{x} \neq \mathbf{x}'} d_H(\mathbf{x}, \mathbf{x}')$. Highest Correct $\leq d_{min} - 1$ erasures and $\leq \frac{d_{min} - 1}{2}$ bit flips.
Codeword weights. With $d_{min} > 0$ will have $d_{\min} = \min_{\mathbf{x} \in \mathcal{C} : \mathbf{x} \neq 0} w(\mathbf{x})$ where $w(\mathbf{x}) = \sum_{i=1}^n \mathbb{1}\{w_i = 1\}$ is the weight of the codeword (# of 1s). For linear code, minimum distance equals minimum weight.
Maximum-likelihood decoding. For any channel $P_{Y|X}$ and any codebook $\{x^{(1)}, \dots, x^{(M)}\}$ the decoding rule that minimises P_e is the maximum-likelihood decoder. $\hat{m} = \arg\max_{j=1,\dots,M} P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}^{(j)})$. For BSC, ML decoding is equivalent to minimum (Hamming) distance decoding. $\arg\min_{j=1,\dots,M} d_H(\mathbf{x}^{(j)}, \mathbf{y})$
Syndrome decoding. $S = yH = (x \oplus z)H$. Can be immediately computed from check matrix H given channel input y.
Min distance codeword. $\hat{\mathbf{z}} = \arg\min_{\mathbf{z}' : \mathbf{z}'\mathbf{H} = \mathbf{S}} w(\mathbf{z}')$ (i.e. $\mathbf{z}'$ with fewest 1's) $\hat{\mathbf{x}} = \mathbf{y} \oplus \hat{\mathbf{z}}$
Proof. Define $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} \oplus \mathbf{y} \Rightarrow$ $d_H(\mathbf{x}^{(i)} \oplus \mathbf{y}) = w(\mathbf{z}^{(i)})$