Channel Coding Introduction

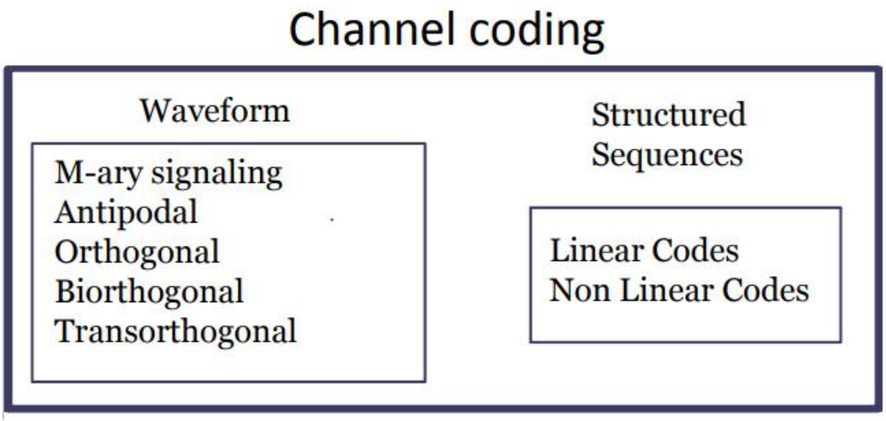
Arranged by long.a.zhang@tieto.com

有则改之，无则加勉。

——《论语·学而》

What is Channel Coding?

Amazing: No Fundamental Difference between Coding and Modulation !



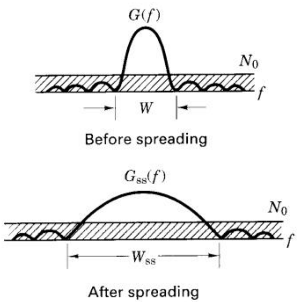
- Waveform Coding: Transforming waveforms to <u>better</u> waveforms
- Structured Sequences: Transforming data sequences into <u>better</u> sequences, having structured redundancy

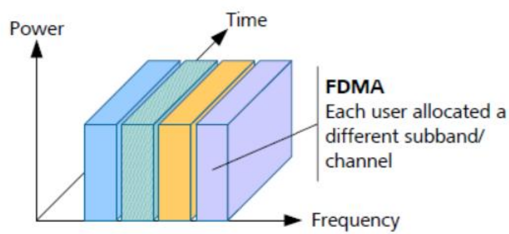
Better: in the sense of making the decision process less subject to errors

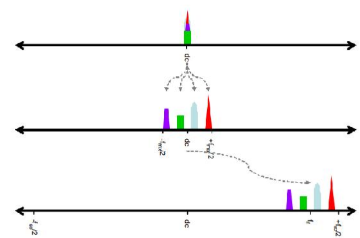
➢ Why Modulate?

Digital symbols are transformed into waveforms that are compatible with the characteristics of the channel.

- The size of the antenna depends on the wavelength λ.
- Separate the different signals: FDMA (Frequency-Division Multiplexing Access).
- Minimize the effects of interference: Spread-Spectrum Modulation: CDMA.
- Place a signal in a frequency band where design requirement (filtering, amplification) can be easily met: RF signal converted to IF (Intermediate Frequency).



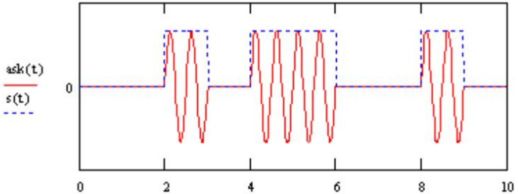




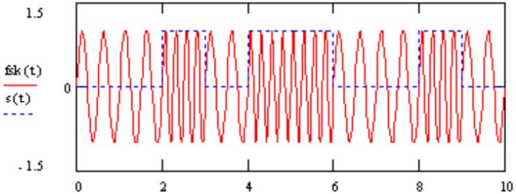
- **What is channel coding?**
  - Coding could change the input waveforms so as to improve communications performance by increasing the robustness against channel impairments (noise, interference, fading(small scale), ..).
  - Coding can be considered as the embedding of signal constellation points in a **higher dimensional** signaling space, the distance between points can be increased, which provides for better error correction and detection.
  - **Time diversity** can be obtained by **interleaving** (against large scale fading) and **coding** over symbols across different coherent time periods.
- **Note:** Channel coding was developed in the context of AWGN channels.
  - **White**: The power spectral density (PSD) is a constant (i.e., flat) over all frequencies.
  - **Gaussian**: The probability density function (pdf) of the noise amplitude at any given time follows a Gaussian distribution.

More About Modulation

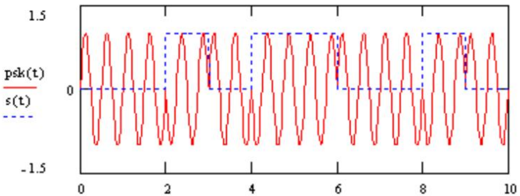
Three kinds of Memoryless Modulation



$$ASK(t) = s(t)\sin(2\pi f t)$$



$$FSK(t) = \begin{cases} \sin(2\pi f_1 t) & \textit{for bit } 1 \\ \sin(2\pi f_2 t) & \textit{for bit } 0 \end{cases}$$



$$PSK(t) = \begin{cases} \sin(2\pi f t) & \textit{for bit } 1 \\ \sin(2\pi f t + \pi) & \textit{for bit } 0 \end{cases}$$

The word modulation it is more common today used for the mapping bits to waveforms process. The mapping of baseband encoded waveform into passband waveform also called modulation for history reason.

Objective: Maximize the rate enter the encoder, retrieve the original bit with a suitably small error rate, subject to constraints on the transmitted power and bandwidth.

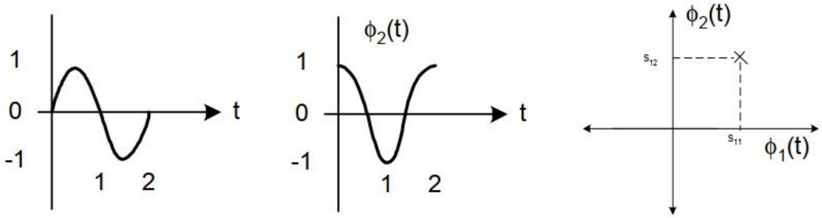
Signal Spaces and Modulation

- **Signal Spaces** provides us with a **geometric method** of conceptualizing the modulation process.

- In a physical space when we describe a vector by its coordinates (x, y); the vector is being described by a linear combination of two functions (1, 0) and (0, 1). These two functions are called **basis functions** , they have **unit energy** and **orthogonal** to each other.

$$\int_{-\infty}^{+\infty} \phi_i(t)\phi_j(t) = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

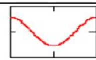
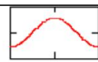
- You can think of I and Q as the x, y axis projection of a signal.



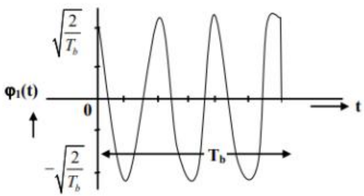
BPSK (Binary Phase Shift Keying)

- The two time-limited energy signals s1(t) and s2(t) are defined based on a single basis function ϕ1(t) as:

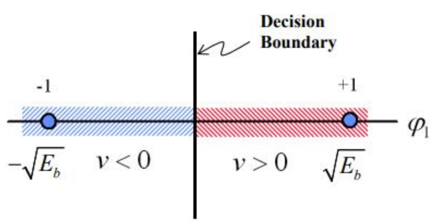
$$s_1(t) = \sqrt{\frac{2E_b}{T_b}}.\cos 2\pi f_c t \quad \text{and} \quad s_2(t) = \sqrt{\frac{2E_b}{T_b}}.\cos[2\pi f_c t + \pi] = -\sqrt{\frac{2E_b}{T_b}}.\cos 2\pi f_c t$$

Symbol	Bit	Carrier Signal	I	Q
S1	0		1	0
S2	1		-1	0

- The basis function is: $\varphi_1(t) = \sqrt{\frac{2}{\mathrm{T_b}}}.\cos 2\pi f_c t \;;\quad 0 \le \mathrm{t} < \mathrm{T_b}$

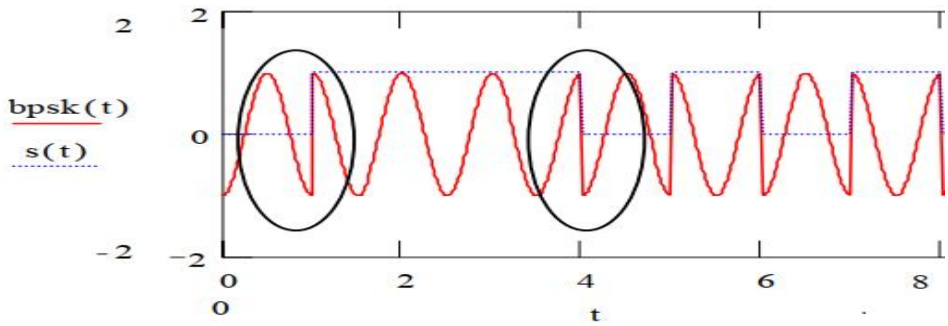


- The constellation is:

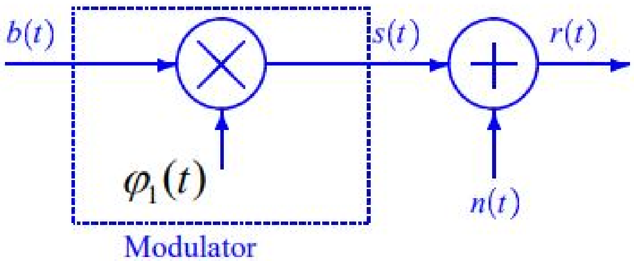


BPSK

- The picture is a sequence 0111 0101 at a carrier frequency of 1 Hz, which is not realistic.

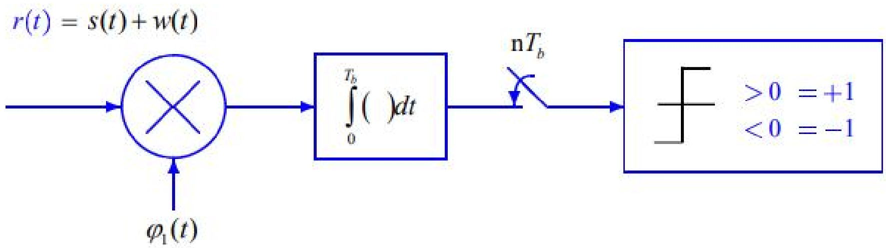


- A simple scheme for BPSK modulator.

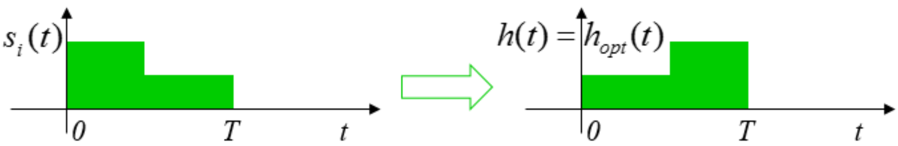


BPSK

- A simple scheme for **coherent demodulation** of BPSK is:

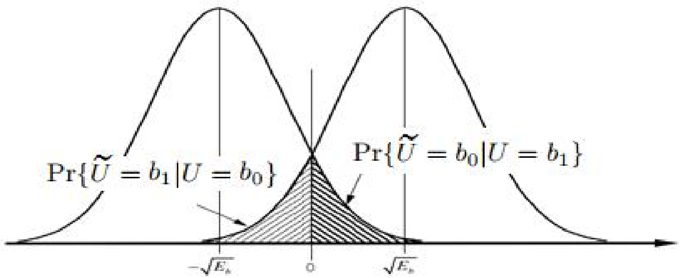


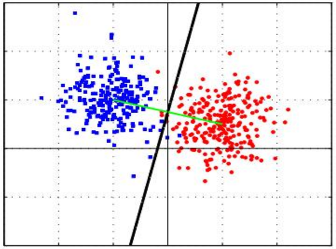
- The **matched filter** is designed for the **SNR is maximized** at the sampling time.
- The matched filter output at the sampling time, can be realized as the **correlator** output.
- The matched filter which is the time-reversed and delayed version of the conjugate of the transmitted signal.



BPSK

- The probability density of decision statistic for BPSK is :

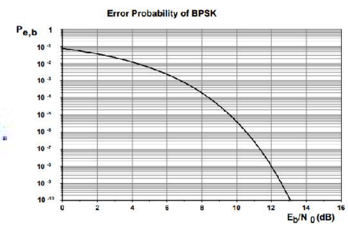




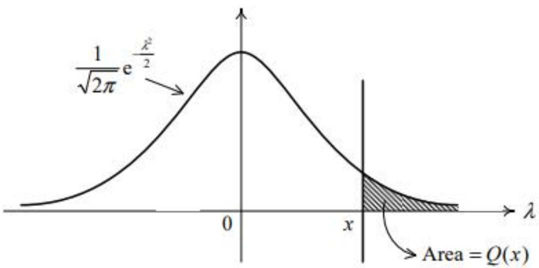
- The ***Bit Error Probability*** of BPSK is:

$$P_{e,b} = Q\left(\sqrt{\frac{2E_b}{N_0}}\right)$$

For $P_{e,b} = 10^{-5}$ we need a $E_b/N_0 = 9.6\text{dB}$.



- ***Q-Function:***

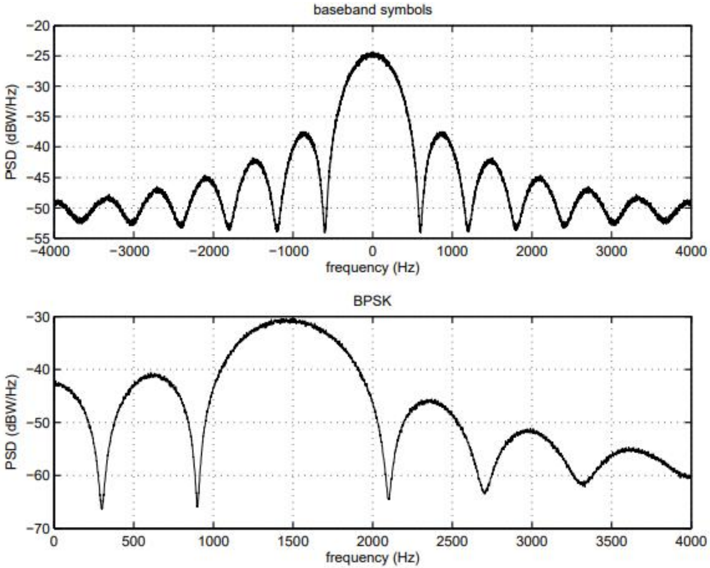


$$Q(x) \equiv \frac{1}{\sqrt{2\pi}} \int_x^{\infty} \exp\left(-\frac{\lambda^2}{2}\right) \mathrm{d}\lambda$$

$$Q(-x) = 1 - Q(x)$$

BPSK

➢ BPSK Frequency Spectrum (low spectral efficiency):

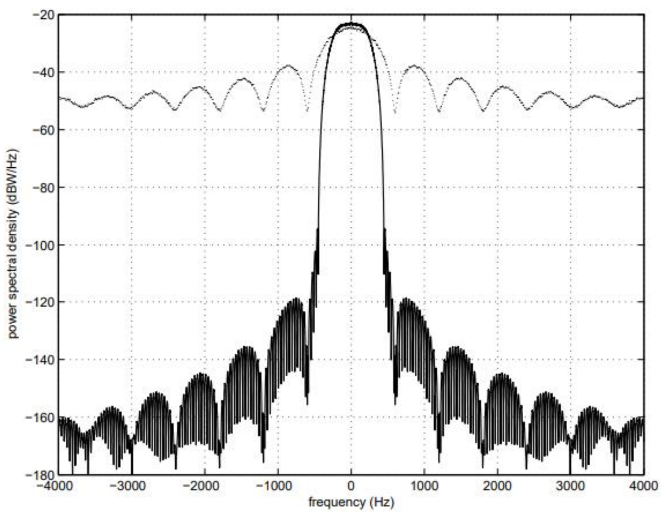


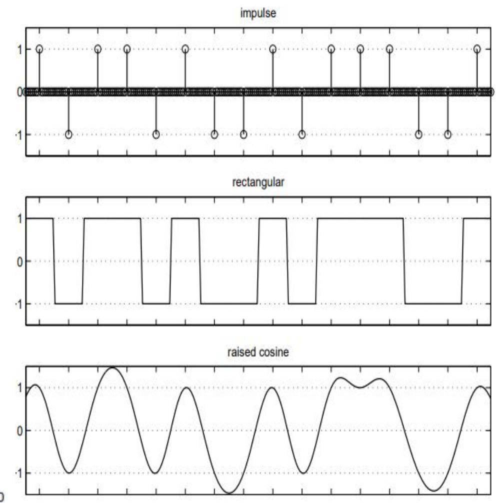
Carrier is a pure sinusoid， the spectrum is impulse. Multiplying in time domain is equivalent convolving in frequency domain. Convolution of any spectrum with a frequency impulse centers this spectrum about the frequency of the impulse.

BPSK

➢ ***Pulse Shaping*:**

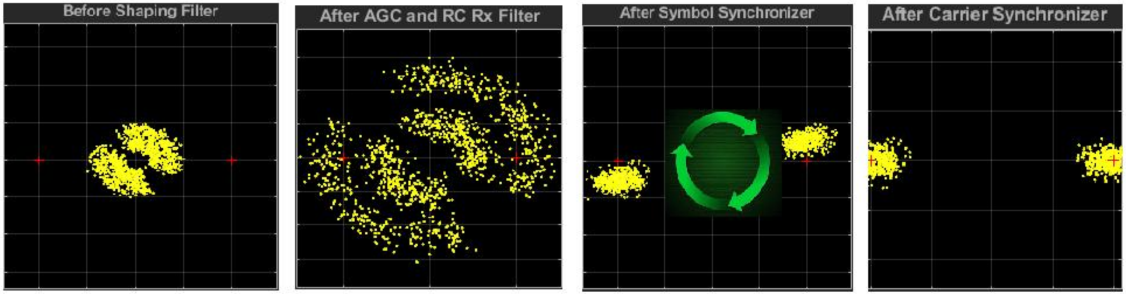
- *Goals or Why?*
  Smaller Side lobes, Robustness to timing error
  Reduce inter-symbol interference (ISI)
  Efficient bandwidth utilization
- The difference between the rectangular and raised-cosine pulse shapes in frequency and time domain are:





BPSK

- ***Other Components in Real Receiver Loops of BPSK***
  - *Automatic Gain Control (AGC)*
    Avoid Saturation errors
    Better synchronize performance
  - *Timing Recovery (symbol synchronization)*
    Sampling frequency/phase must be determined
  - *Carrier Recovery*
    Remove the frequency offset
  - *Channel Equalization*
    An adaptive filter attempts to remove ISI caused by multipath channel which can be viewed as linear filter.



Public

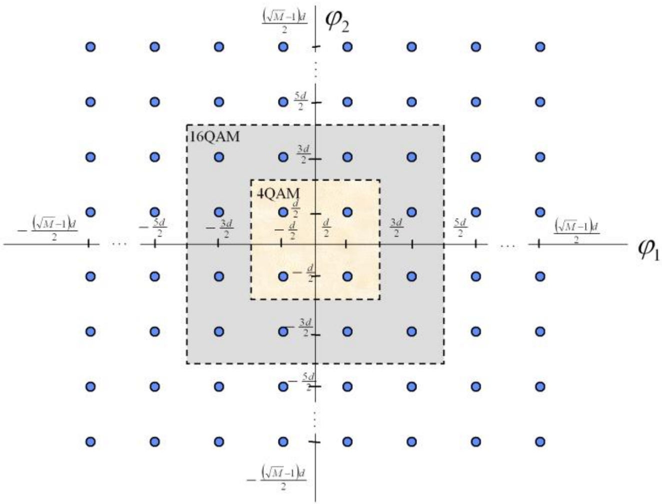
QAM (Quadrature Amplitude Modulation)

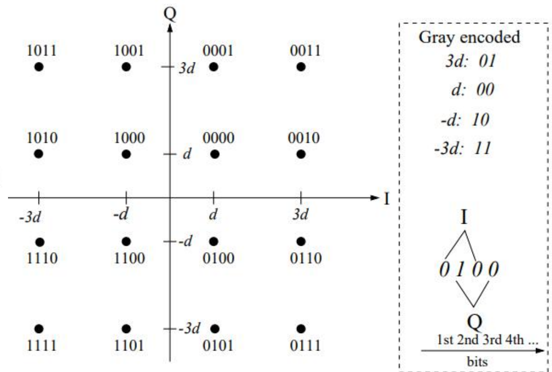
- *Can be seen as a combination of **ASK** and **PSK***
  - QAM is a two dimensional generalization of PAM, the two basis functions are usually (contain shaping pulse):

$$\varphi_1(t) = \sqrt{\frac{2}{T}}\mathrm{sinc}\left(\frac{t}{T}\right)\cos\omega_c t$$

$$\varphi_2(t) = -\sqrt{\frac{2}{T}}\mathrm{sinc}\left(\frac{t}{T}\right)\sin\omega_c t$$

The sinc(t/T) term may be replaced by any Nyquist pulse shape

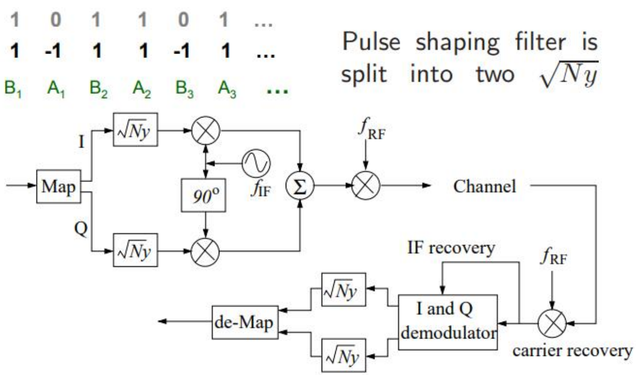


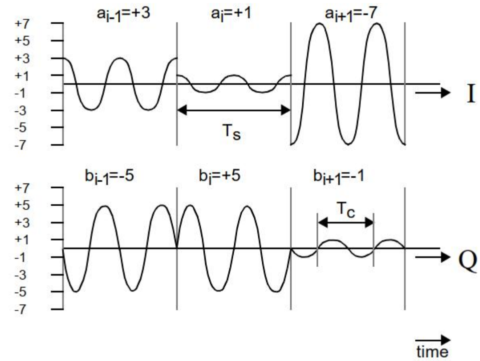


Public

QAM

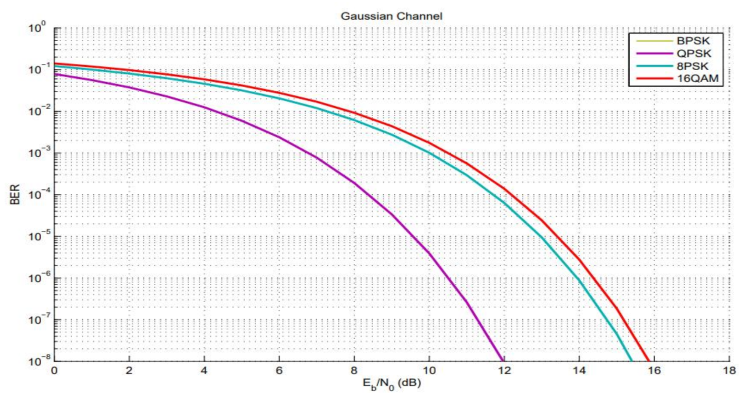
➢ Simplified QAM modem:





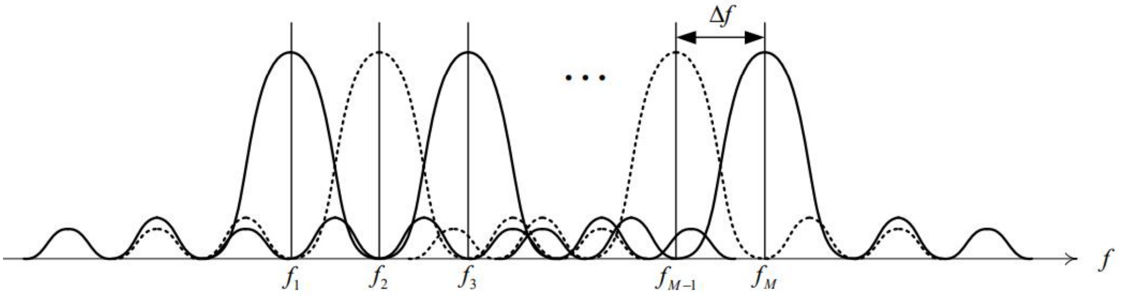
➢ *BER:*
Bandwidth efficient than PSK.

$$P_E(M) = 4\left(1 - \frac{1}{\sqrt{M}}\right) Q\left(\sqrt{\frac{3\log_2 M}{M-1}\frac{E_b}{N_0}}\right)$$



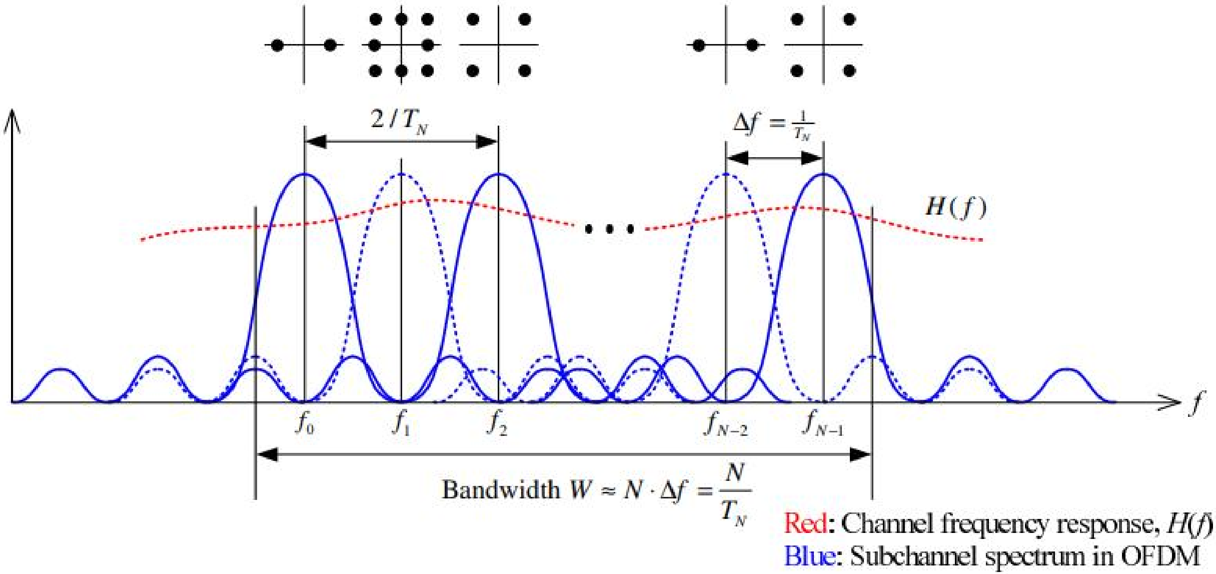
OFDM (Orthogonal Frequency-Division Multiplexing)

- Spectrum of M-FSK:



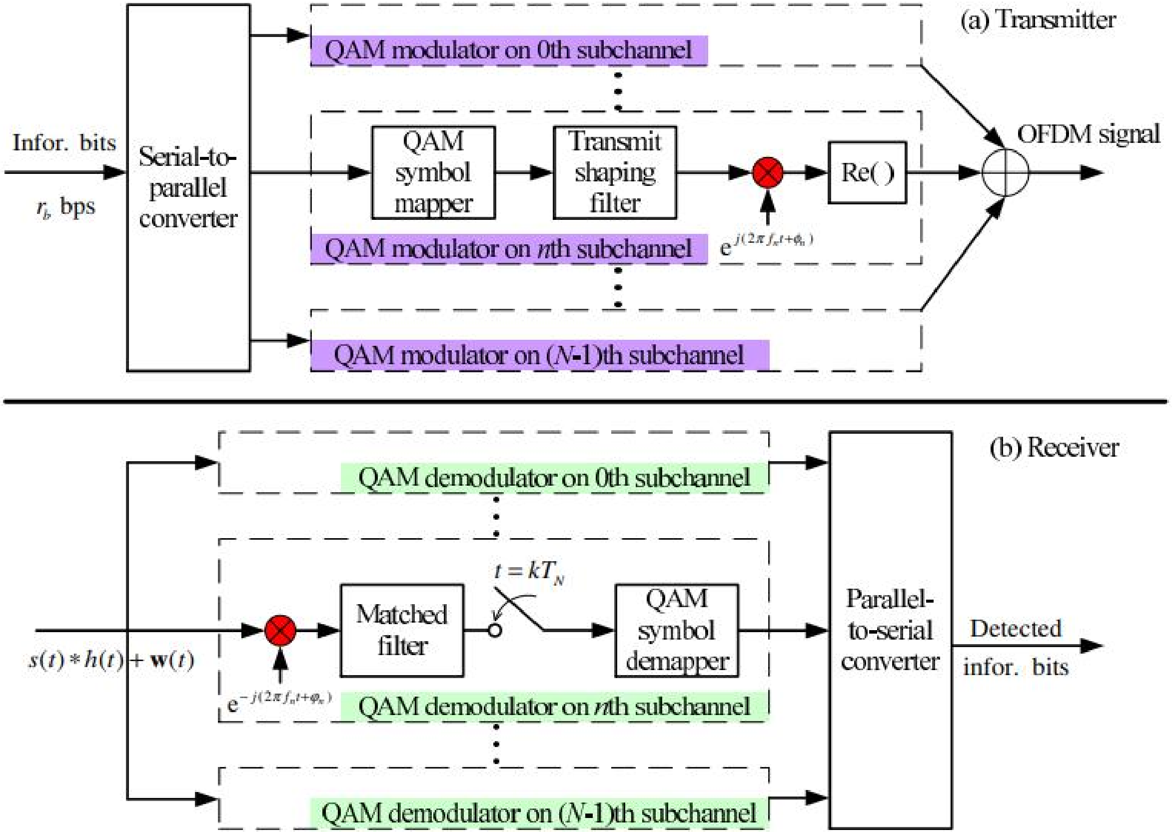
- For FSK with $N = 2^\lambda$ frequencies, only one of N frequencies is activated over one symbol duration of $T_s = \lambda T_b$, where T_b is the bit duration. What frequency that is activated over any symbol duration is determined by the mapping from λ information bits to the frequency value.
- FSK is not a spectral-efficient modulation scheme!
- Why not using all the carriers to carry information at the same time since they are orthogonal? This leads to OFDM (orthogonal frequency-division multiplexing) technique.

OFDM

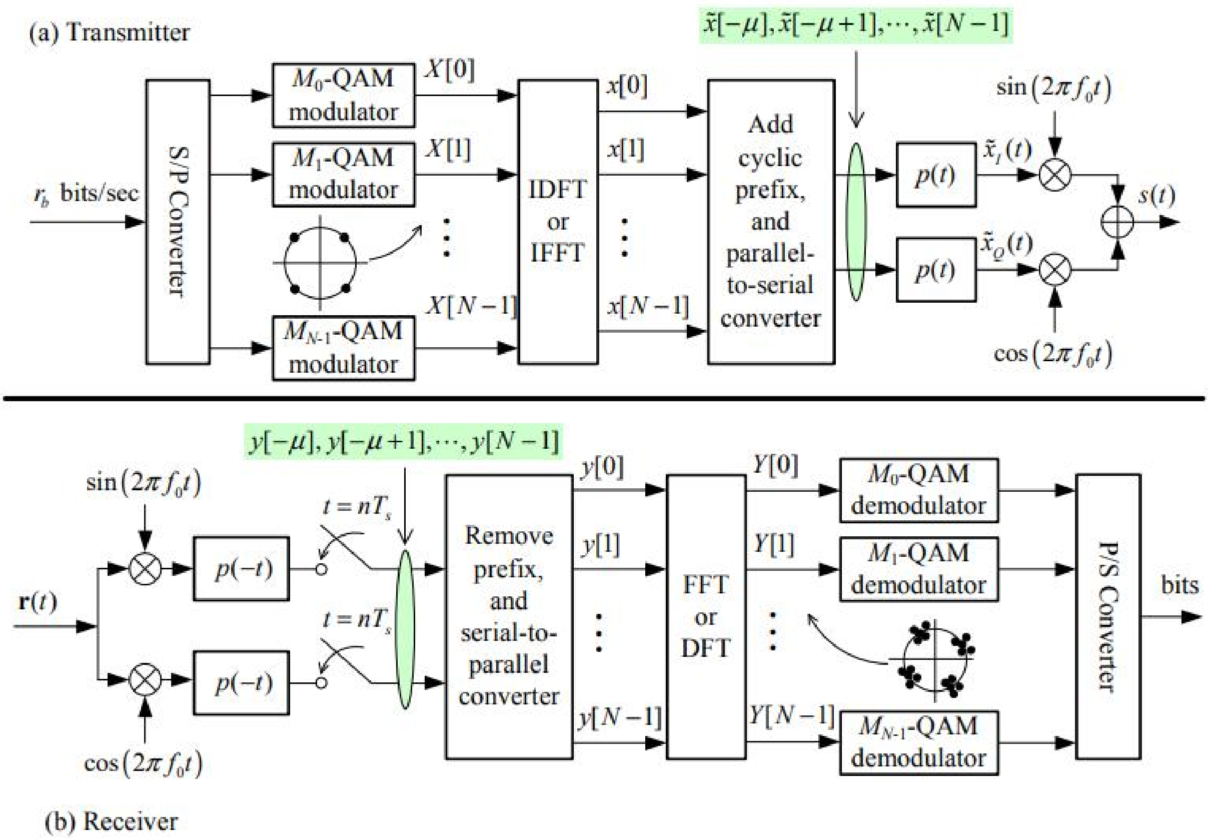


- In OFDM, the spectrum is divided into *overlapping* but *orthogonal* subcarriers. Each sub-carrier is independently modulated by QAM. The minimum subcarrier separation is $1/T_N$, where T_N is the OFDM symbol length.
- OFDM can be simply looked upon as a combination of *amplitude*, *phase* and *frequency* modulation techniques.

OFDM (Viewed as a Multiple QAM system)



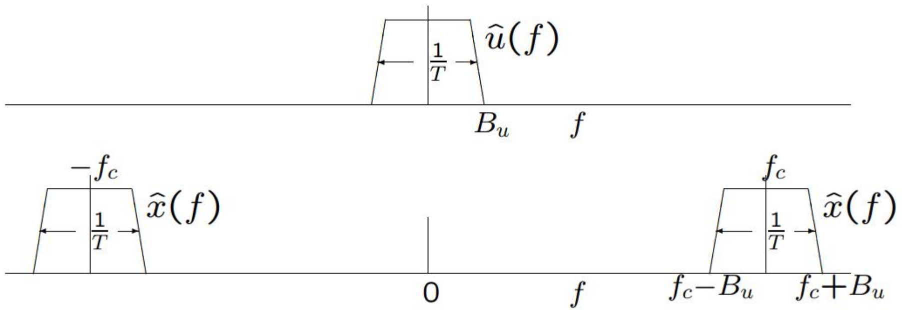
OFDM (Implemented with DFT/IDFT)



Public

Frequency Translation

$$
\begin{aligned}
x(t) &= u(t)[e^{2\pi i f_c t} + e^{-2\pi i f_c t}] = 2u(t)\cos(2\pi f_c t), \\
\hat{x}(f) &= \hat{u}(f - f_c) + \hat{u}(f + f_c).
\end{aligned}
$$



- *The bandwidth **B** is **2Bu**. The **bandwidth** is always the range of **positive frequencies** (used in transmission).*
- *The **power** at **passband** is scaled to be **twice** that at baseband.*
- *When a baseband waveform limited to B is shifted up to passband, the passband waveform becomes limited to 2B.*
- *If fc - B_u ≤ f ≤ f_c filtered out of x(t), result is single sideband.*

Why Error Correction Coding?

Error control techniques

- Automatic Repeat reQuest (ARQ)
  - Full-duplex connection, error detection codes
  - The receiver sends feedback to the transmitter, saying that if any error is detected in the received packet or not (Not-Acknowledgement (NACK) and Acknowledgement (ACK), respectively).
  - The transmitter retransmits the previously sent packet if it receives NACK.
- Forward Error Correction (FEC)
  - Simplex connection, error correction codes
  - The receiver tries to correct some errors
- Hybrid ARQ (ARQ+FEC)
  - Full-duplex, error detection and correction codes

Channel Capacity: Discrete-time AWGN Channel

- $\mathbf{x}$ is an input sequence with power constraint: $\frac{1}{n}\sum_{i=1}^{n} x_i^2 \le P$

$$\mathbf{y} = \mathbf{x} + \mathbf{z}$$

- Noise z_i is a zero-mean Gaussian random variable with variance σ^2.

for large n

$$\frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\sum_{i=1}^{n} (y_i - x_i)^2 \rightarrow \sigma^2$$

$$\| \mathbf{y} - \mathbf{x} \|^2 = n\sigma^2$$

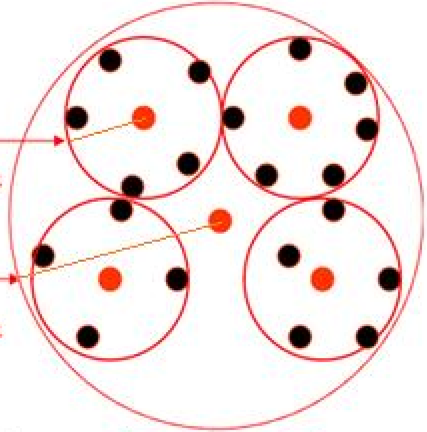
n-dimensional hyper-sphere with radius $\sqrt{n\sigma^2}$

$$\frac{1}{n}\sum_{i=1}^{n} y_i^2 \le P + \sigma^2$$

$$\| \mathbf{y} \|^2 \le n(P + \sigma^2)$$

n-dimensional hyper-sphere with radius $\sqrt{n(P+\sigma^2)}$

$$V_n(R) = \frac{\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}+1\right)} R^n$$



- How many input sequences can we transmit over this channel at most such that the hyperspheres do not overlap?

$$M = (\sqrt{P+\sigma^2})^n / (\sqrt{\sigma^2})^n$$

- The maximum rate that can be reliably communicated :

$$\boxed{C = \frac{1}{n}\log_2 M = \frac{1}{2}\log_2\left(1 + \frac{P}{\sigma^2}\right) \text{ (bit/transmission)}}$$

This is a nice geometric interpretation

Channel Capacity: Continuous-time AWGN Channel

- Capacity of discrete-time AWGN channel:

$$C = \frac{1}{n}\log_2 M = \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \quad \text{bit/transmission}$$

For continuous-time AWGN baseband channel with bandwidth W, power constraint $\boldsymbol{P}$ watts, and two-sided power spectral density of noise $N_0/2$,

- What is the average noise power per sampling symbol? N_0W
- What is the minimum sampling rate without introducing distortion? $2W$

- Capacity of continuous-time AWGN channel:

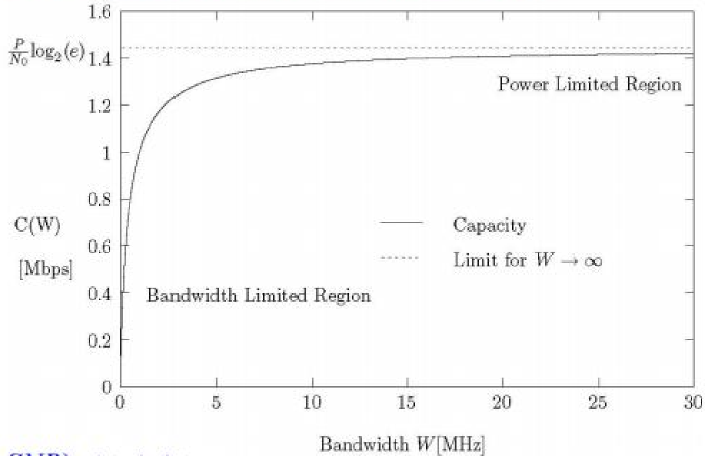
$$C = 2W \cdot \frac{1}{2}\log_2\left(1+\frac{\boldsymbol{P}}{N_0W}\right) = W\log_2\left(1+\frac{\boldsymbol{P}}{N_0W}\right) \quad \text{bit/s}$$

- Can we increase the capacity by enhancing the transmission power?

  Yes, but the capacity increases logarithmically with $\boldsymbol{P}$ when $\boldsymbol{P}$ is large.

- If we can increase the bandwidth without limit, can we get an infinitely large channel capacity?

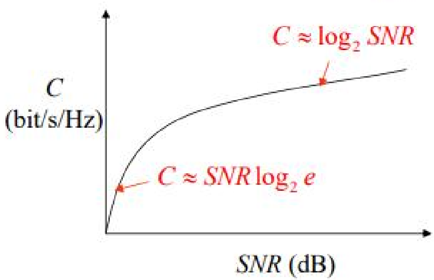
  No. $\lim_{W \to \infty} C = \frac{\boldsymbol{P}}{N_0} \log_2 e$



- How to achieve (approach) AWGN channel capacity?

  Turbo, LDPC, Polar, ..., ?

- Spectral Efficiency: $C = \log_2(1 + SNR)$ bit/s/Hz



$$SNR = \frac{\boldsymbol{P}}{N_0 W} = \frac{P}{N_0}$$

P: power per unit bandwidth

SNR and E_b/N_0

- Nominal bandwidth $W = 1/2T$ Hz
- The noise energy per real plus imaginary component is defined to be N_0 , signal power $P = E_s W$
- Rate $R = W\rho$ b/s: number of bits per symbol times the number of symbols per second.

$$R = W \log_2 |\mathcal{A}|, \quad \text{for QAM;} \qquad R = 2W \log_2 |\mathcal{A}|, \quad \text{for PAM.}$$

- Spectral Efficiency: ρ (b/s)/Hz

$$\rho = \log_2 |\mathcal{A}|, \quad \text{for QAM;} \qquad \rho = 2\log_2 |\mathcal{A}|, \quad \text{for PAM.}$$

- Energy per bit: $E_b = E_s / \rho$

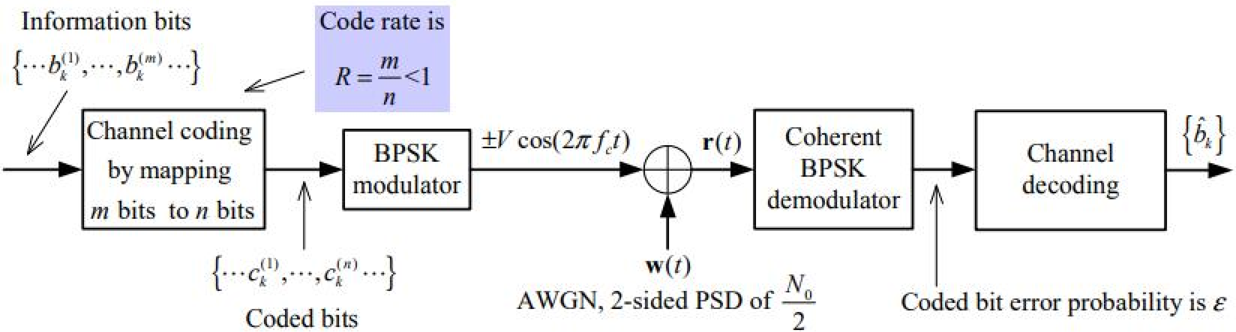
$\rho < C = \log_2(1+SNR)$
$SNR = E_s/N_0 > 2^\rho - 1$
$E_b/N_0 = E_s/\rho N_0 = SNR/\rho > (2^\rho - 1)/\rho$

Ultimate Shannon Limit on E_b/N_0 : $\rho \to 0$
ln 2 ≈ 0.69 (-1.59 dB)

Public

tieto

BPSK with (7,4) Hamming Code



$b_k^{(1)} b_k^{(2)} b_k^{(3)} b_k^{(4)}$	$c_k^{(1)} c_k^{(2)} \cdots c_k^{(7)}$
(0000)	(0000000)
(1000)	(1101000)
(0100)	(0110100)
(1100)	(1011100)
(0010)	(1110010)
(1010)	(0011010)
(0110)	(1000110)
(1110)	(0101110)
(0001)	(1010001)
(1001)	(0111001)
(0101)	(1100101)
(1101)	(0001101)
(0011)	(0100011)
(1011)	(1001011)
(0111)	(0010111)
(1111)	(1111111)

$$\mathbf{c}_k = \left[c_k^{(1)} c_k^{(2)} \cdots c_k^{(7)}\right] = \left[b_k^{(1)}, b_k^{(2)}, b_k^{(3)}, b_k^{(4)}\right] \cdot \underbrace{\begin{bmatrix} 1 & 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 1 \end{bmatrix}}_{\text{Generator matrix } \mathbf{G}}$$

Note: Multiplication and addition are modulo-2 operations.

Decoding Hamming Code

- For a (n, m) block code, there is $(n-m) \times n$ *parity-check* matrix $\mathbf{H}$ such that every valid codeword $\mathbf{c} = \left[c^{(1)}, c^{(2)}, \cdots, c^{(n)}\right]$ satisfies

$$\mathbf{c} \cdot \mathbf{H}^\top = \mathbf{0}.$$

- This parity-check matrix $\mathbf{H}$ can be used for decoding.
- The parity-check matrix of the (7,4) Hamming code is

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

- *Hamming distance is important for classical coding theory but not much useful in today.*

Decoding Hamming Code

Public

- Let $\mathbf{y}$ be the length-n vector of the decoded bits after BPSK demodulator. The decoding consists of three steps:
  1. Compute the *syndrome* of $\mathbf{y}$, namely $\mathbf{y} \cdot \mathbf{H}^\top$.
  2. Locate the coset leader $\mathbf{e}_l$ whose syndrome is equal to $\mathbf{y} \cdot \mathbf{H}^\top$. Then $\mathbf{e}_l$ is taken to be the error pattern caused by the channel.
  3. Decode the vector $\mathbf{y}$ into the codeword $\mathbf{c}^* = \mathbf{y} + \mathbf{e}_l$. From $\mathbf{c}^*$ find the corresponding k information bits by inverse mapping.

Syndrome	Coset leaders
(100)	(1000000)
(010)	(0100000)
(001)	(0010000)
(110)	(0001000)
(011)	(0000100)
(111)	(0000010)
(101)	(0000001)

y = (1 0 0 1 0 0 1)

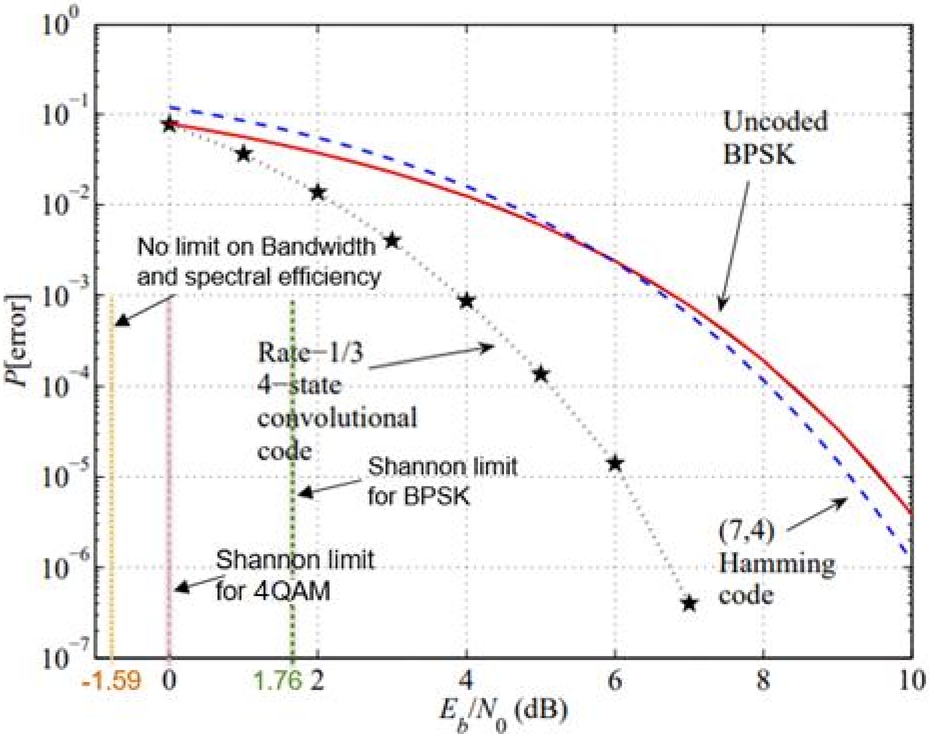
s = y$\cdot H^T$ = (1 1 1)

e_l = (0 0 0 0 0 1 0)

c^* = y + e_l = (1 0 0 1 0 1 1)

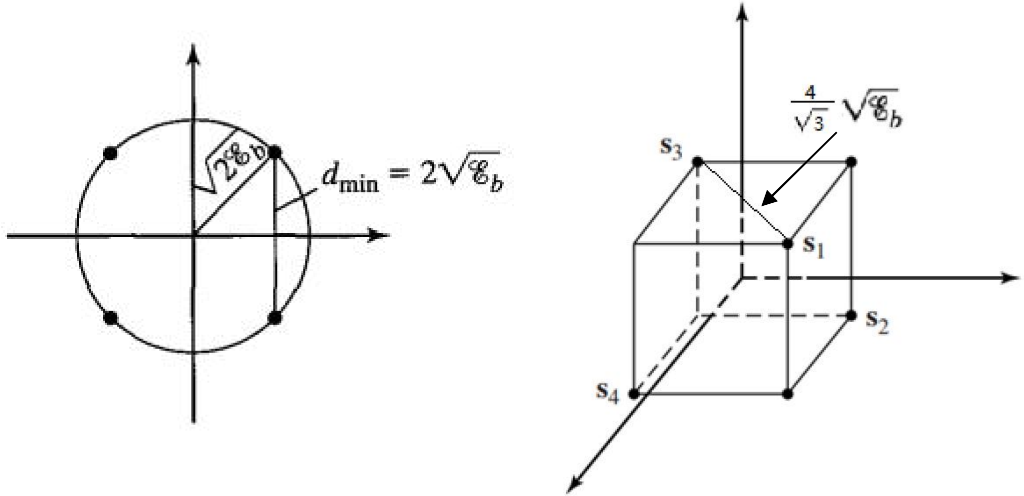
message is 1011

Performance Comparison



- ***Shannon Limit***: There exists a limiting value of *Eb/N0* below which there can be no error-free communication at any information rate.

QPSK and a Coding Scheme



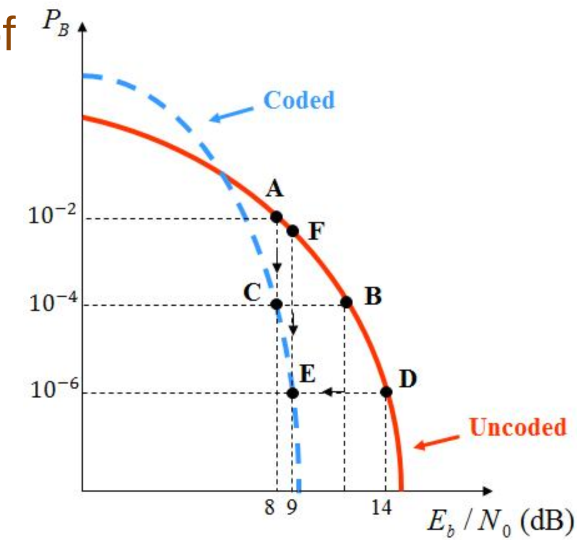
- Error probability is a decreasing function of the minimum ***Euclidean Distance***.
- Coding results in a lower error probability (which is equivalent to a ***higher effective SNR***) at the price of increasing the bandwidth and the complexity of the system.

Public

tieto

Why use error correction coding?

- *Trade-off 1 : Error performance vs. Bandwidth*
  - Consider a system does not use error-correction coding (point **A**, 10^{-2}), the customer complaints about the quality and suggests that the bit-error probability should be lower to 10^{-4}.
  - ***A→B*** : Suppose the SNR of 8dB is the most available in this system.
  - ***A→C*** : Improve error performance by error-correction coding. The price of the redundant bits is more transmission bandwidth.

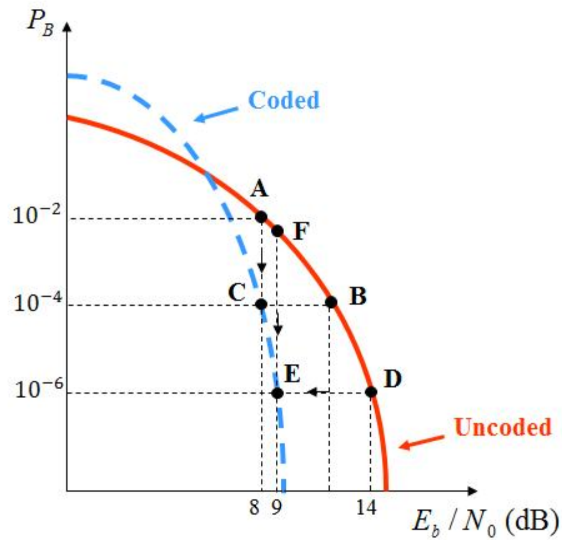


Public

Why use error correction coding?

- *Trade-off 2 : Power vs. Bandwidth*
  - Consider a system operating at point **D**, the customer complaints about the reliability, and suggests that the SNR or power could be reduced.
  - ***D→E*** :More bandwidth or delay.
  - ***Coding Gain***:

$$G(\mathrm{dB}) = \left(\frac{E_b}{N_0}\right)_{uncoded}(\mathrm{dB}) - \left(\frac{E_c}{N_0}\right)_{coded}(\mathrm{dB})$$



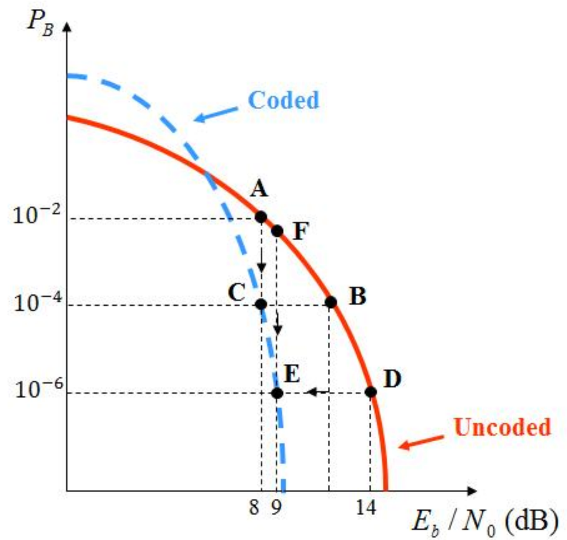
Public

Why use error correction coding?

- *Trade-off 3 : Data rate vs. Bandwidth*
  - Consider a system operating at point **D**, the customer's data rate requirement increases.
  - The relationship:

$$\frac{E_b}{N_0} = \frac{P}{N_0}\left(\frac{1}{R}\right)$$

  - ***D→F:*** If we do nothing except increase the R, the SNR would decrease.
  - ***D→E:*** The SNR is reduced, but the code facilitates getting the same error probability with a lower SNR.

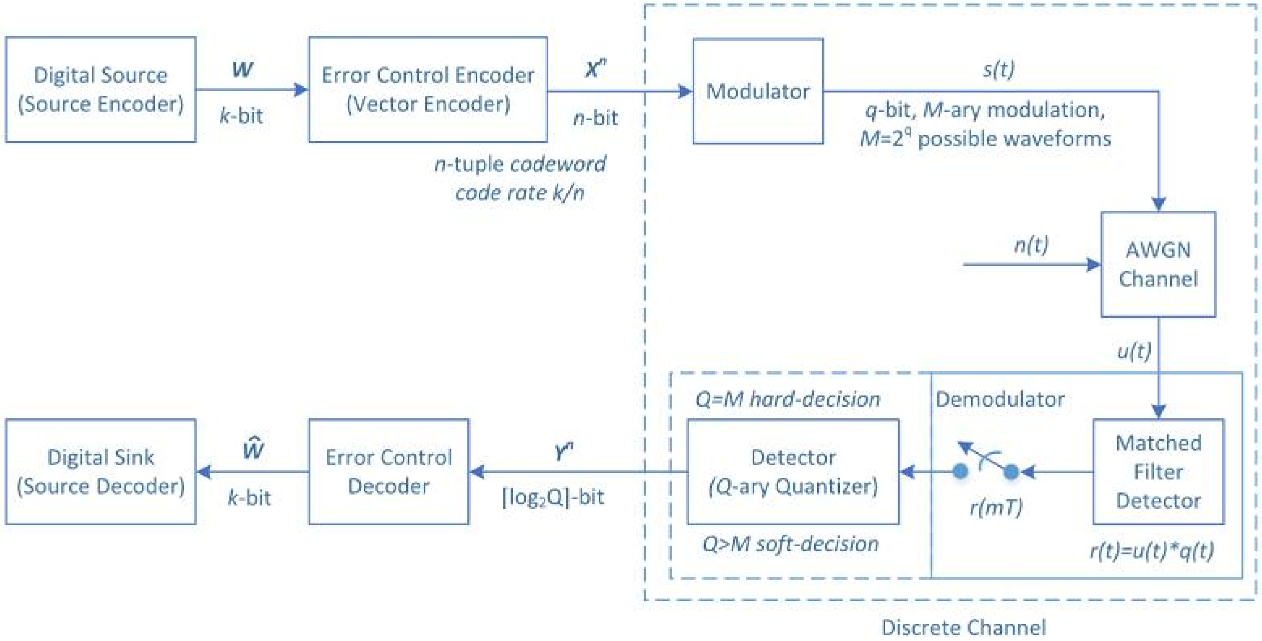


Why use error correction coding?

- *Trade-off 4 : Capacity vs. Bandwidth*
  - This is similar to trade-off 3. Consider a CDMA cellular, the capacity per cell is inversely proportional to E_b/E_0 .
- *What's more?*
  - Coded-Modulation: jointly optimize both channel coding and modulation, spectrally-efficient.
  - Adaptive Coded Modulation…
- *Conclusion: channel coding performance measure*
  - ✓ *Probability of Error, Q-function*
  - ✓ *Power, SNR, E_b/N_0*
  - ✓ *Shannon Limits, Capacity*
  - ✓ *Coding Gain*
  - ✓ *Required Bandwidth*
  - ✓ *Delay, Complexity*

Public

Discrete Memoryless Channels



We receive a sequence of bits $(B_1, B_2, \ldots, B_k)$ from the compressor that is uniformly distributed on $\{0,1\}^k$. We enumerate over all possible messages and assign each sequence a number W:

$$(B_1, B_2, \ldots, B_k) \to W, \quad \text{where } W \in \{1, \ldots, 2^k\}$$

Probability of error $p_e = P(\hat{W} \neq W)$, rate $R = k/n$.

Public

大知闲闲，小知间间；大言炎炎，小言詹詹。
——庄子 • 齐物论

Long Zhang

Hardware Engineer
Tieto Oyj, ZSR Product Development Services /
long.a.zhang@tieto.com