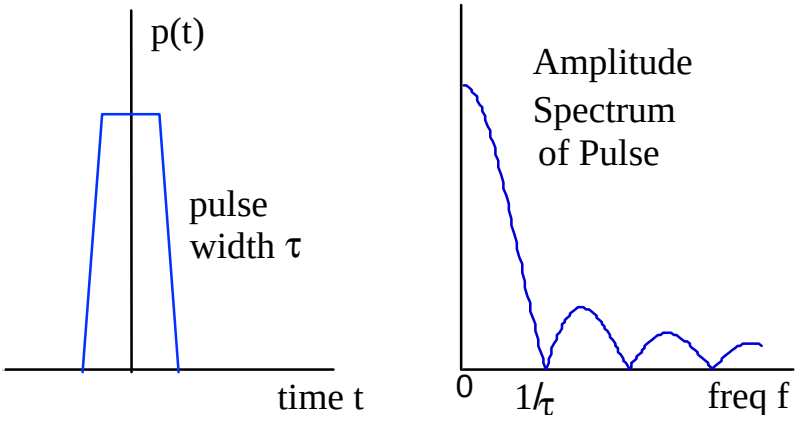
TCOM 370 NOTES 99-5

MODULATION AND DEMODULATION
Modems for Data Transmission over Analog Links

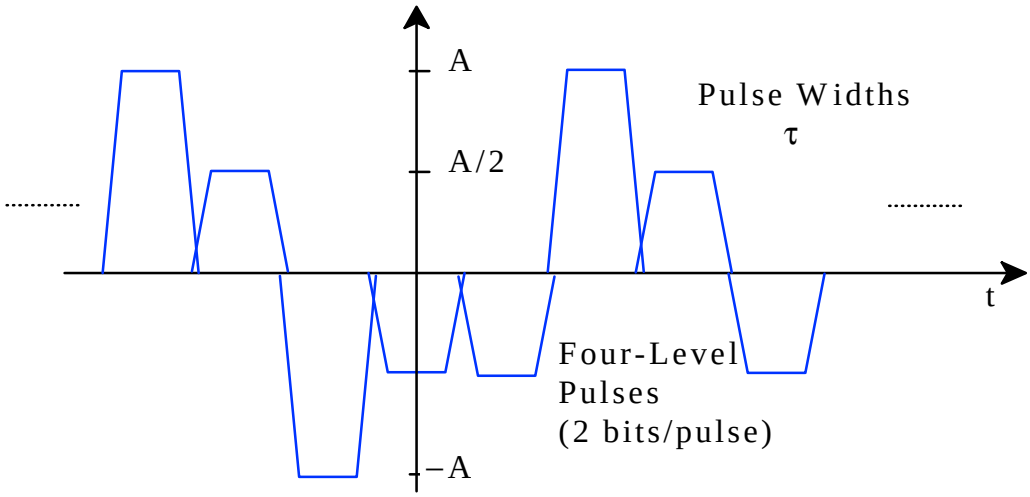
1. REVIEW

PULSE TRAIN BANDWIDTH

- **Single pulse of duration τ** sec. needs **frequency band of 0 to approx. 1/τ** for transmission. The larger fraction of the pulse energy is near 0 frequency. The bandwidth required is approximately **W=1/τ** Hz



- For **train of such pulses** with individual amplitudes corresponding to some ***arbitrary*** data stream, to receive each pulse individually at the receiver, **bandwidth required** is also **W**. With this bandwidth we can transmit **W pulses/second (each pulse of width τ=1/W).**



[Detail:

"Arbitrary data stream" implies that the pulse amplitudes in the sequence of pulses appear statistically independent of each other. For example, for bipolar (two-amplitude, +A and -A) pulses the two amplitudes are assumed to occur with equal probability for each pulse, independently of the amplitudes of other pulses. Such a sequence of pulse amplitudes *does not show any regular or deterministic pattern.* This is typical of a pulse train representing a real data stream. It is for such typical situations that the bandwidth required for a pulse train is the same as the bandwidth for a single pulse. Otherwise, consider as an example rectangular width-τ pulses each with the *same* amplitude A and generated at a rate of $1/\tau$ pulses per sec. (This might represent the unusual data pattern of all "1" in the bit stream.) This pulse train produces just a constant value of A, it is a dc waveform, with no frequency components other then the 0 frequency (only a_0 is non-zero in the Fourier series.) The other extreme of this is the text-book's ***worst-case binary sequence*** 101010101..... so that if we use a pulse p(t) with amplitude A for bit "1" and pulse amplitude 0 (or –A) for bit "0" then we have a sequence of alternating pulses with period $T=2\tau$ secs. Thus the fundamental frequency for a Fourier series representation for this train of pulses is $f_0= 1/(2\tau)$. If we only use the fundamental frequency (not a very good approx.) we need a bandwidth of W/2 if we define $W=1/\tau$.. On the other hand, the next harmonic is the third harmonic, because the second harmonic coefficient (a_2) in this case is 0. For this better approx. the bandwidth is therefore 3W/2. *Our value of W=1/t for the bandwidth is a better and easier to remember value*, and applies to the typical case of arbitrary (not special) data streams.]

PULSE TRAIN TRANSMISSION

While it is possible to transmit pulses directly as electrical waveforms over wireline links (e.g. twisted pair, or coaxial cable, as used for example in ethernet LANs), there are situations where we have to ***modulate* a higher *carrier frequency* by the pulse train** in order to produce a signal that can be carried on the link. A common example of this is in the use of the public switched telephone network to transmit data signals using *modems (MOdulators/DEModulators).* This is of course not the only place where modulation techniques are used; they are absolutely necessary in wireless links, for example. We will use the PSTN for data transmission as a commonly encountered application where modulation is used.

[The rectangular-type pulses that we have sketched in the previous figures are **"baseband"** pulses, because the frequencies they are comprised of go all the way to 0 Hz (the base of the frequency spectrum from 0 to ∞)]

2. ANALOG PSTN

The Public Switched Telephone Network was designed to carry **voice signals** as analog signals (continuous-time, continuously-varying amplitude) representing the sound pressure waveform directly.

- The band of frequencies that each end-user's signal occupies is approximately **between 300 and 3400 Hz** or a **bandwidth of approximately 3.1 KHz.**

The telephone line from the end-user (subscriber) premises terminates at the *central office* serving the local area in the telephone system. (The distance is of the order of a few thousand feet.)

The *lower limit on the frequency band* comes from limitations of the circuits coupling the electrical signal in the telephone set to the telephone line (twisted pair of wires)[1]. This attenuates very low frequencies. The very low frequency end is also used for dial tone and ringer signals.

The *upper frequency limit*, and the total bandwidth limitation of approximately 3 KHz, is not a limitation of the medium (twisted pair), which can carry much higher frequencies. It is the design of the system and equipment that allocates only this band of frequencies to each end-user; the telephone system equipment in the local office attenuates frequencies beyond 3400 Hz very significantly, and absolutely no frequencies beyond 4 KHz can be passed. Each user voice signal is allocated only this bandwidth, so that the large total available bandwidth within the system may be shared amongst a large number of subscribers.

MULTIPLEXING:

Within the long-distance telephone network, high bandwidth transmission links (such as terrestrial microwave or fiber optic) are used to carry large numbers of voice channels, by means of multiplexing techniques. **Multiplexing** means the sharing of a single high bandwidth channel amongst a number of individual users. This may be accomplished by shifting (through *modulation*) individual user signals to different 4-KHz-

[1] The coupling is AC coupling, via transformers.

wide frequency channels in a **frequency division multiplexing or FDM** scheme. In a **time-division multiplexing or TDM** scheme, each user signal is *sampled in time* at a uniform rate (of 8 K samples/sec) and different user channel samples are "interleaved" into a much higher rate stream. Multiplexing is required to simultaneously carry multiple low-bandwidth signals over the same high-bandwidth transmission facility.

THE NEED FOR MODULATION:

In the PSTN, the available band of frequencies to the end-users (300-3400 Hz) is fixed, is not very wide, and does not support frequencies close to zero (d.c. and very low frequency signals are blocked); the PSTN is designed basically to carry analog voice signals.

The strict **bandwidth limit** means that we have to use multi-level pulses carrying several bits/pulse in order to get reasonable data rates. Also, consider the **limitation of not being able to transmit low frequencies** (<300 Hz approximately). A train of rectangular-type pulses on the other hand has its most significant frequencies at the low frequency end.

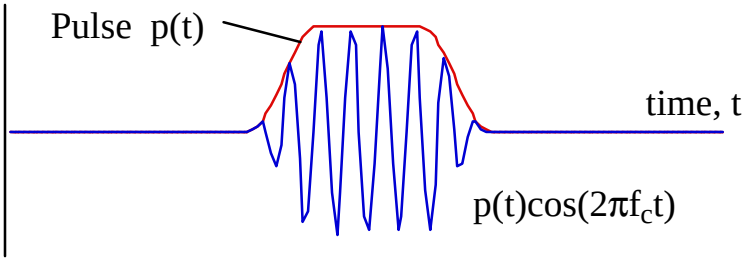
- Thus to send data (a sequence of binary digits) over the existing telephone system we have to use special techniques for multilevel pulse transmission.

The techniques used involve most importantly the operations of MOdulation and DEModulation, performed by modems.

(Other refinements to make high rate data transmission possible over such analog systems involve the use of hybrids to allow full-duplex operation over two-wire lines, and equalization.)

3. FREQUENCY SHIFTING BY AMPLITUDE MODULATION

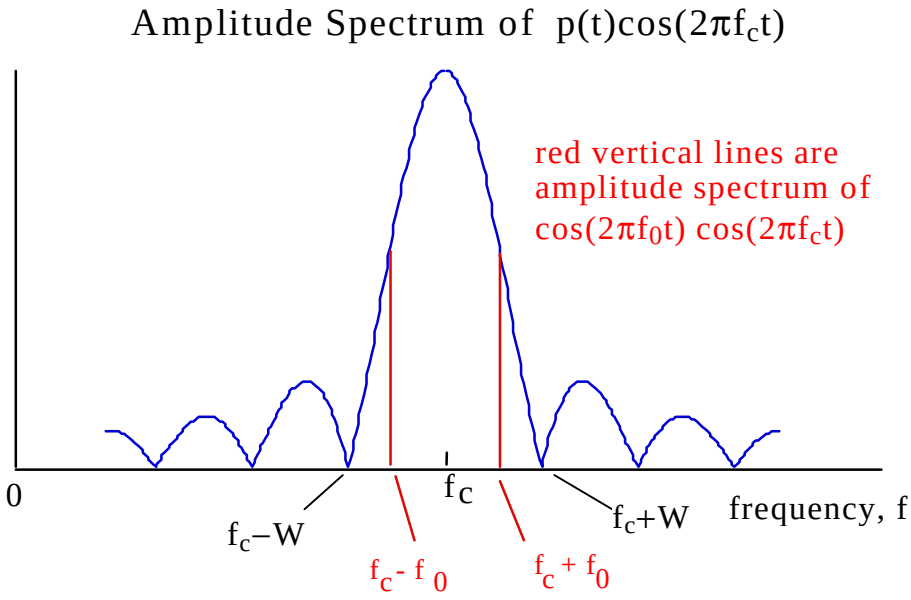
Let p(t) be a baseband pulse (e.g. the ideal rectangular pulse) and suppose we **multiply it by a "carrier" frequency signal** $\cos(2\pi f_c t)$, where f_c is some fixed frequency. (Multiplication of two signals can be implemented easily with electronic circuits).



Now p(t) is made up of an infinite number of frequencies between 0 and W; consider what happens to **any one** of these frequencies say $\cos(2\pi f_0 t)$ when it is multiplied by the carrier frequency.

$$\cos(2\pi f_0 t)\cos(2\pi f_c t) = \frac{1}{2}\cos(2\pi[f_c+f_0]t) + \frac{1}{2}\cos(2\pi[f_c-f_0]t)$$

- Each frequency f of the pulse gives rise to *two* frequencies, one on *either side of* f_c (with half the amplitude each.)
- In addition, the frequency 0 simply becomes f_c.
- Thus the effect of multiplication of a pulse by $\cos(2\pi f_c t)$ is to produce a signal with an **amplitude spectrum** that is the original one from 0 to W reproduced symmetrically on both sides of the frequency f_c.
- This is called **AMPLITUDE MODULATION (AM)**, i.e. the pulse modulates (becomes) the amplitude of the carrier.



- Amplitude modulation of the carrier by the pulse results in an amplitude spectrum approximately contained between frequencies f_c-W to f_c+W. The **total bandwidth is now 2W, centered at f_c**.

A typical f_c for low-rate telephone line modems is 1800 Hz (approximate center of 300 -- 3400 band), and for 1200 baud signaling (1200 pulses per sec., requiring W=1200 Hz) the range of frequencies required for transmission is therefore 600 -- 3000 Hz, within the capability of a telephone voice channel.

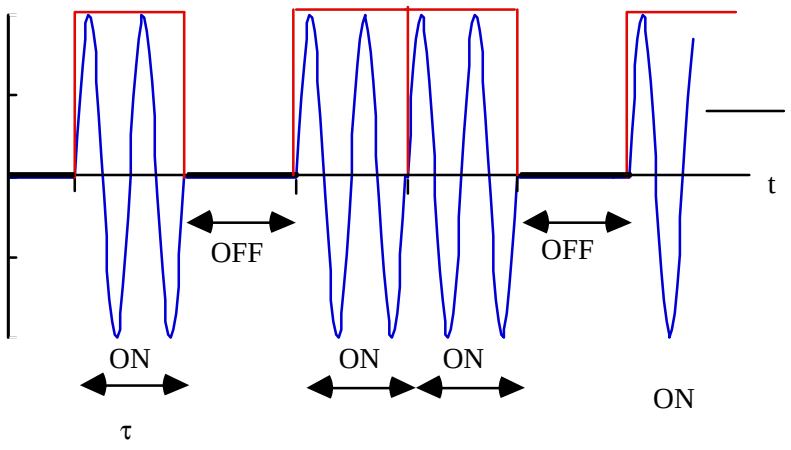
[Actually, it is possible to suppress one side of the frequencies centered on f_c, because we have in principle all the information needed to reconstruct the pulse p(t) from a modulated carrier with only frequencies on one side of f_c. This is called a **single-sideband** (SSB) signal, and requires a bandwidth of W for pulses of width τ=1/W. SSB systems are more complex to implement.]

DEMODULATION:

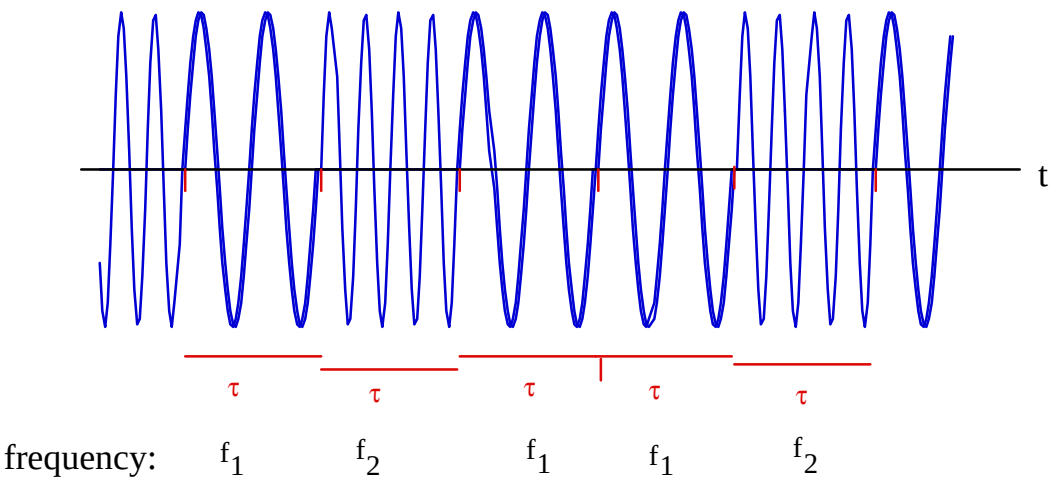
If we receive $p(t)\cos(2\pi f_c t)$ we can simply **multiply it with a copy of $\cos(2\pi f_c t)$** to get $p(t)\cos^2(2\pi f_c t) = \frac{1}{2}p(t)[1+\cos(2\pi 2 f_c t)]$. As long as $W < f_c$, as is the case here, the *lowest* frequency $2f_c - W$ in $\frac{1}{2}p(t)\cos(2\pi 2 f_c t)$ is *higher than* the *highest* frequency W in $\frac{1}{2}p(t)$, and we can filter out $\frac{1}{2}p(t)\cos(2\pi 2 f_c t)$ with a simple circuit that suppresses frequencies higher than W **(a low-pass filter)** to obtain a replica of p(t). This is called **"coherent" demodulation,** since we require an exact (phase-matched) copy of the frequency f_c at the receiver.

AMPLITUDE SHIFT KEYING

If we take a sequence of ON-OFF rectangular pulses (pulse amplitudes A or 0, a unipolar pulse sequence) representing binary digits, and use it to modulate a carrier frequency as above, we call the resulting signal an **AMPLITUDE SHIFT KEYING (ASK)** signal; it can be thought of as a carrier frequency f_c that is either turned ON or OFF to represent binary digits 1 or 0, respectively.



4. FREQUENCY SHIFT KEYING (FSK)



In this type of modulation scheme for *two-level pulses* we *always* put out a pulse p(t) that is amplitude modulating a carrier frequency, whether the data digit is a 1 or a 0, but we distinguish between the two digits by using **two slightly different carrier frequencies** f_1 and f_2.

For example, the very low-speed Bell System 103 Modem standard for 300 baud (and 300 bps) uses frequencies of 2025 and 2225 Hz for sending data in one direction. Clearly the total bandwidth in this case is roughly from 2025-300=1725 to 2225+300=2525. This is obtained by considering the band of frequencies for pulses with carrier frequency 2025 Hz and separately the band of frequencies for those with carrier frequency 2225 Hz. The lowest frequency needed is the lower edge of the band for carrier at 2025 Hz, and the highest is the higher edge of the band at 2225 Hz carrier. In the figure above, the pattern of carrier frequencies f_1 and f_2 transmits the same data sequence as the ON-OFF amplitude pattern of pulses in the previous figure.

For data in the opposite direction, carrier frequencies of 1070 and 1270 Hz are used in the Bell 103 modem, or the frequency band 770 -- 1570Hz. Note that two directions of data flow can be accommodated **simultaneously**, the receive and transmit circuits using distinct frequency sub-bands within the overall 3 KHz bandwidth.

The frequency occupancies given above for the Bell 103 Modem example are approximate, because the spectrum of an FSK signal is mathematically much more complicated to derive exactly. (This is because the "1" positions and "0" positions in a bit stream are disjoint and hence the train of pulses of frequency f_1 and that of the pulses with frequencies f_2 are deterministically related (they occupy disjoint time slots), so that the two types of pulses in any one direction of transmission cannot really be considered separately). Nonetheless, the approach used above to obtain frequency band occupancy of the FSK signal gives useful practical approximations.

(Frequency shift keying (FSK) described above is a special case of the general technique of frequency modulation (FM), in which a message signal s(t) is made to modulate the instantaneous frequency of a carrier; the frequency f in a carrier $\cos(2\pi f_c t)$ is varied about f_c as a function of time in accordance with the amplitude variations of s(t) about 0).

5. FULL DUPLEX OPERATION

- Any scheme that allows data transmission to take place simultaneously in two directions is called **full duplex** operation. In a **half-duplex** scheme data transmission can take place in only one direction at a time.

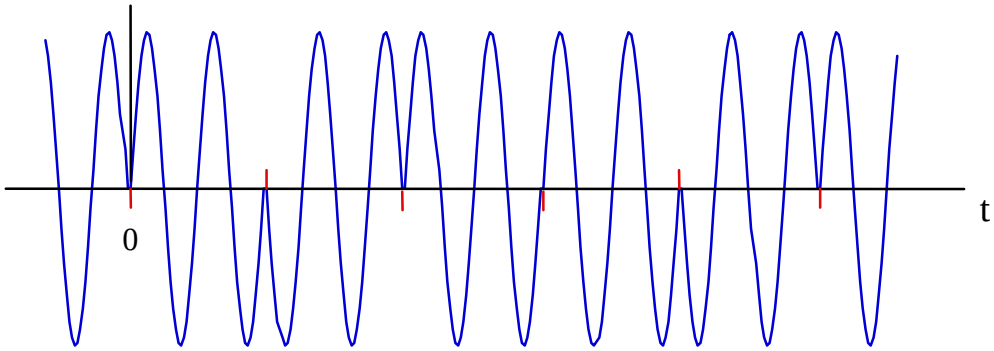
- One way to provide for full-duplex operation is through the use of a separate pair of wires for transmission of signals in each of the two directions. (This is called *four-wire* operation). Note that the telephone link between a subscriber set and the central office is a *two-wire* connection.

- It is possible to use *time-division duplexing* for full-duplex operation, in addition to *frequency division duplexing.* (Alternating use of time-slots on a line, as opposed to using two different frequency sub-bands). TDD is not used in telephone voiceband modems.

- It is possible to use two-wire lines to obtain full-duplex operation and obtain the full resources of the line in both directions simultaneously, through the use of *hybrids* with special techniques for *echo-cancellation*.

6. PHASE SHIFT KEYING (PSK)

Consider changing not the amplitude of a carrier as in ASK or the frequency of the carrier as in FSK to signal a "1" or a "0", but its PHASE. Thus we might send out a carrier frequency $\sin(2\pi f_c t)$ for a duration τ with phase either 0^o or 180^o (π).



Phase of $\sin(2\pi f_c t)$: 0 π 0 0 π

[We can equivalently think of the above as showing a pulse train with a carrier $\cos(2\pi f_c t)$ in which the carrier phase is either $-\pi/2$ or $+\pi/2$; note that $\cos(2\pi f_c t - \pi/2) = \sin(2\pi f_c t)$ and $\cos(2\pi f_c t + \pi/2) = -\sin(2\pi f_c t) = \sin(2\pi f_c t + \pi)$. In both cases the phase difference between the two alternatives is π.]

Notice that since a phase shift of π means an inversion of sign of the carrier waveform, this is **equivalent to ASK** not with an ON-OFF pulse train but **with a POS-NEG (bipolar)** pulse train. At the receiver we can demodulate with a coherent demodulator (multiplying the received signal with $\cos(2\pi f_c t - \pi/2)$ or $\sin(2\pi f_c t)$, and low-pass filtering) to get either a positive amplitude or a negative amplitude and decide between bit 1 and bit 0.

Even though the two-level phases of 0 and π result in a scheme that is basically an AM scheme, more generally with multiple phase levels per pulse we get a distinct modulation format with special advantages.

The **bandwidth** is the same as in the case of amplitude modulation of a carrier with a pulse train; it is **2W**

MULTI-LEVEL PHASE

More generally, we can use a set of more than 2 possible phases in each pulse or signaling interval of width τ; for example, **4-PSK, 8-PSK** and **16-PSK.** In **4-PSK**, for example, we use phases 0, $\pi/2$, π, and $3\pi/2$, and are able **to convey 2 bits of data** per pulse (M=4). Of course for M-level pulses we convey $\log_2 M$ bits of data per pulse.

The **bandwidth** required for this remains unchanged from the binary case for the case of arbitrary (random) data streams. Of course, the **SNR needed** for reliable demodulation now **increases**, because we have to make finer distinctions between closer phases at the receiver.

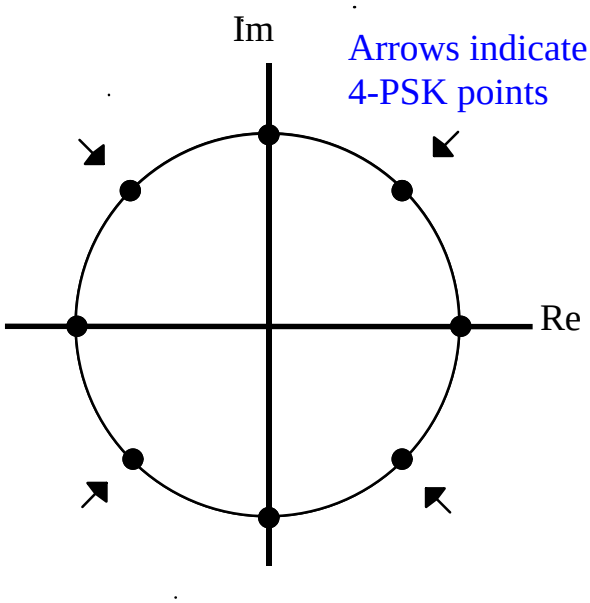
[A practical note: since it is non-trivial to obtain absolute phase information at the receiver, a more clever practical approach uses "differential" coding in the following way. Instead of using the chosen phase as the absolute phase of a carrier pulse, the pulse is given the phase of the preceding pulse shifted by the chosen phase. At the demodulator, each pulse is compared to the preceding pulse to determine the signaled phase value. This is called **Differential PSK (DPSK)**.]

7. QUADRATURE AMPLITUDE MODULATION (QAM)

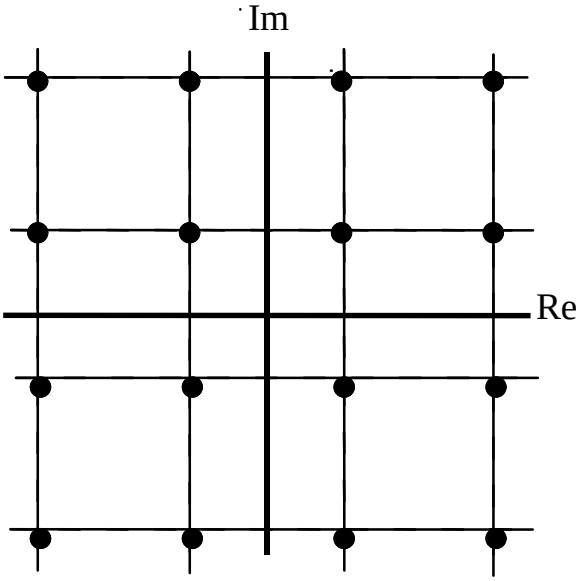
This is a more sophisticated and very useful modulation scheme. It is used for all ***high-speed voiceband*** telephone line modems *(e.g. the 14.4 Kbps V.32bis and 28.8 Kbps V.34 standard modems).* The idea is to **combine amplitude and phase modulation** producing multi-level signaling "constellations" with a large number M of allowed amplitude/phase combinations on each pulse or symbol interval of length τ. This allows us to obtain large M values in practice and therefore **increase the bit rate for a fixed baud rate.**

For the V.32bis modem, the baud rate is 2400 per sec. and the constellation size (number of different amplitude and phase combinations for each pulse) is 128, but the effective M is 64 and allows $\log_2(64)=6$ bits/pulse. [$\log_2(128)=7$, but 1 bit out of 7 is actually used to provide error control].

The constellation in a QAM scheme is the set of amplitude/phase combinations, depicted as points on the complex plane of real and imaginary axes. Amplitudes with **0 phase and phase π** are positive and negative values, respectively, on the **real axis**. A phase of π/2 takes a point from the real axis to the imaginary axis (90^o rotation). A combination of amplitude A and phase θ is the tip of a vector of length A at an angle θ with respect to the positive real axis. Generally, a carrier at fixed frequency f_c with amplitude A and phase angle θ is written as $A\cos(2\pi f_c t+\theta)$ = Real Part of $\{Ae^{j\theta} e^{j2\pi f_c t}\}$. Thus the different allowed $Ae^{j\theta}$ points are of interest, once we have fixed the carrier frequency.



8-PSK Points



16-QAM Points

Note that 4-QAM is the same as 4-PSK.

- QAM allows us to obtain ***better separation*** between constellation points with a given amount of average signal power than a pure AM or PSK scheme for the same number of points M, and thus leads to better noise immunity.

This is the key point about QAM. Of course **pure AM** uses only different amplitudes, and its constellation is a **one-dimensional set of points** on the real axis. By going to two-dimensional constellations, we can achieve better separation between individual constellation points for the same amount of power. Consider 16-PSK vs. 16-QAM. The 16 PSK points would be spaced uniformly around the circle above, and each would get quite "close" to its neighbors. Making the circle larger means increasing the amplitude which translates into transmitting higher power. Of course QAM is also subject to power considerations. For QAM we cannot expand the constellation arbitrarily, because this also entails transmitting higher amplitude pulses and therefore more power. The fact is that for a given transmitted power of 1, say, the amplitude of the PSK circle is also 1 (amp^2 =power). In QAM since the amplitudes are not all equal, some can have amplitudes larger than 1 if others are smaller than 1. Assuming each constellation point will occur with equal probability, we can work out exactly the spacing between nearest neighbors in the QAM case also. It turns out that the QAM constellation is more efficient in distributing (spacing) points for a fixed amount of power than is PSK. This translates into better performance in the presence of noise, which will tend to confuse the receiver especially between closely spaced alternatives. As M increases, the advantage of QAM over PSK grows. However, PSK does have the constant-amplitude property which makes it less prone to channel nonlinearities (as in satellite transmission) and makes demodulation and recovery of data bits easier to implement. One-dimensional AM constellations are more prone to errors in the presence of noise, but have the advantage of not requiring phase handling circuits.