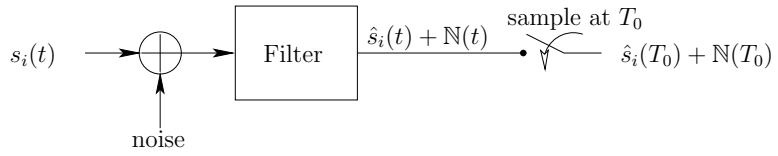
ECE 361: Lecture 3: Matched Filters – Part I

3.1 The Receiver Model

In Lecture 2, we studied the decision-making process in the digital communications receiver which was modeled as shown below. In our simple model, the signal $s_i(t)$, denoting one the two possible received signals $s_0(t)$ and $s_1(t)$ is processed through a filter and then sampled at time T_0. The received signal is corrupted by noise which also passes through the filter and corrupts the sample value which is thus $\mathbb{Y} = \hat{s}_i(T_0) + \mathbb{N}(T_0)$.



As we noted in Lecture 2, the pdf of $\mathbb{N}(T_0)$ does not depend on whether $s_0(t)$ or $s_1(t)$ is received, that is, the noise is assumed to be independent of the received signal. A good model for $\mathbb{N}(T_0)$ is a zero-mean Gaussian random variable with variance σ^2. The *minimax* decision rule is the same as the *maximum-likelihood* decision rule is the same as the *minimum-error-probability* decision rule (since we are assuming equally likely signals) and consists of deciding that a 0 or a 1 was transmitted according as $|\mathbb{Y} - \hat{s}_0(T_0)|$ is smaller than or larger than$|\mathbb{Y} - \hat{s}_1(T_0)|$. Equivalently, assuming that $\hat{s}_0(T_0) > \hat{s}_1(T_0)$, the receiver compares the sample value $\mathbb{Y}$ to a *threshold* $\theta = (\hat{s}_0(T_0) + \hat{s}_1(T_0))/2$ and decides that a 0 or a 1 was transmitted according as $\mathbb{Y}$ is larger than or smaller than θ. The error probability P_e achieved by this decision rule is $Q\left(\frac{\hat{s}_0(T_0) - \hat{s}_1(T_0)}{2\sigma}\right) = Q(\mathsf{SNR})$ where SNR (*signal-to-noise ratio*) is the value of the argument of the $Q(\cdot)$ function. Note that P_e decreases rapidly as SNR increases. It is also worth noting that there are many different definitions for SNR that are in common use, and thus, care must be taken in comparing systems from different designers since they be defining SNR differently. However, regardless of the exact definition of SNR, P_e decreases as SNR increases in a properly designed system, that is, everyone is in agreement that "increasing SNR is a good thing to do."

The noise in the system is almost always referred to as *channel noise* though most of it actually arises in the front end of the receiver. The random motion of electrons in the electrical conductors comprising the front end of the receiver creates *small* time-varying voltages – referred to as *thermal noise* voltages – that are on the order of a few microvolts or so. Based on experimental evidence, the thermal noise is modeled as a *stationary Gaussian random process*. The interested reader will find a short discussion of random processes in Appendix A of this Lecture, but for our purposes, it suffices to note that if we sample the noise at the receiver input at any time instant, then an reasonable model for this noise sample is a zero-mean Gaussian random variable whose variance is the same regardless of the choice of time instant. The reader may then wonder why it is necessary to use a filter as shown in the above figure of the receiver model. Why not just sample at the receiver input itself and make a decision based on that sample value? After all, the sample will be $s_i(T_0)$ (instead of $\hat{s}_i(T_0)$) plus a Gaussian noise variable which is exactly the situation we studied in Lecture 2! Unfortunately, in many instances, the noise voltages can be of the same order of magnitude as (or even considerably larger than) the voltages created by the received signals, especially when the transmitter is far away, or is restricted in transmitter power. Thus, the error probability can be unacceptably large if we were to make a decision based on a sample taken at the receiver input. But, can filtering ameliorate this situation? In almost all instances, it can. Informally speaking, the noise at the receiver input is *broadband* noise in comparison to which the received signals are *narrowband* signals.[1] Thus, even a simple bandpass filter which passes the signals $s_0(t)$ and $s_1(t)$ (and the *in-band* noise) unchanged while eliminating the *out-*

[1] Bandwidth is a precious commodity in a digital communication system. In wireline systems, the physical properties of the medium limit the usable bandwidth. For wireless systems, frequency bands are allocated by governments, and the signals used in a communication system are required to have negligible energy outside the frequency band allocated to that system.

of-band noise will reduce the noise variance considerably, and thereby reduce the error probability achieved. The use of a filter can be beneficial even when the received signals are strong enough that sampling at the receiver input gives acceptably small error probability. In such a case, the filter can be used not as a device for improving the error probability from acceptably small to fantastically small, but rather as a cost-effective way of achieving the same acceptably small error probability while increasing the spatial distance over which the communication system can operate, or reducing the transmitter power which in turn can have additional side benefits such as a reduction in the size and weight of the transmitter, an increase in battery life, etc. Finally, for those still unconvinced of the utility of filtering before sampling, consider that it is universal practice to *amplify* the receiver input before any sampling is done, and for reasons of power efficiency and ease of implementation, amplifiers also act as bandpass filters. The reason for amplification is that it is much easier to design a circuit that triggers when its input exceeds 0.25V (where the threshold may vary ± 0.01V (say) due to manufacturing process variations or circuit component tolerances) than it is to design a circuit that triggers at a threshold of $0.25 \pm 0.01\mu$V. Note that since P_e is determined by the *ratio* of signal level to noise level, and the amplifier increases the signal level and the noise level by the same factor, amplification by itself does not change the error probability – but amplification *does* make implementation of the sampler, A/D converter, threshold device etc. a lot easier. In summary, digital communication receivers amplify and filter the received signals (plus noise) before sampling and making a decision as to which bit was transmitted. Since the analysis of error probability is unaffected by the amplifier gain, we do not include amplifier gain explicitly in our analysis, though we do incorporate the filtering in the amplifier(s)[2] into the filter shown in the figure above. One other consequence of amplification is important, and simplifies our analysis. Remember that thermal noise is present in *all* the electrical conductors in the amplifier/filter combination and not just in those in the front end of the receiver. However, because of the *amplification* that the thermal noise present in the front end of the receiver undergoes as it passes through the amplifier/filter combination, this amplified noise from the front end is the *predominant* noise present in the output of the amplifier/filter combination, completely swamping out the few microvolts contribution to the noise from the electrical conductors in the output impedance of the amplifier/filter. In other words, for all practical purposes, the noise $\mathbb{N}(t)$ shown in the figure above can be taken to be the result of filtering the noise process in the front end of the receiver. Notice that such an assumption would not be valid in the absence of amplification since the thermal noise generated at the filter output would be comparable in magnitude to that passing through the filter and appearing at its output. Finally, following convention, we blame it all on the channel and say that the noise $\mathbb{N}(t)$ at the filter output is the result of filtering the Gaussan *channel noise process*. The channel itself is called a Gaussian noise channel.

Having justified the need for the filter shown in our model, let us consider what kind of filter we should use. For a Gaussian noise channel, the smallest error probability that can possibly be achieved with given received signals $s_0(t)$ and $s_1(t)$ is achievable by using a suitable linear filter, that is, a linear time-invariant system.[3] Call this smallest achievable error probability P_e^*. The *optimum linear filter* that achieves P_e^* is called a *matched filter* for signals $s_0(t)$ and $s_1(t)$. As the name implies, different signal sets have different matched filters (and achieve different minimum error probabilities). For given signals $s_0(t)$ and $s_1(t)$, the use of a linear filter *not matched* to them will result in a receiver with $P_e > P_e^*$. Notice also that the claim does not mean that a receiver with a nonlinear filter cannot achieve error probability P_e^*: it might well do so, but so can the linear matched filter receiver achieve error probability P_e^*. What the claim does mean is that *no* receiver, whether using a linear filter or a nonlinear filter, can achieve an error probability smaller than P_e^*. That is the least error probability that we can achieve with signals $s_0(t)$ and $s_1(t)$, and we can achieve it with a linear (matched filter) receiver. Looking at receivers with nonlinear filters in the hope of getting error probability smaller than P_e^* is futile.

Restricting ourselves to linear filters, what can be said about the noise process at the filter output? Since the channel noise process is a zero-mean stationary Gaussian random process that passes through the filter, the noise process at the filter output is also a zero-mean stationary Gaussian random process.[4] Thus, regardless of the choice of sampling instant T_0, $\mathbb{N}(T_0)$ is a zero-mean Gaussian random variable with fixed variance σ^2.

[2]Since amplifier gains totaling several tens of dB are usually required, what we have called the amplifier is generally a chain of amplifiers with small gains (on the order of 10 dB, say) instead of one humongous amplifier with a gain of (say) 50 dB.

[3]The proof of this assertion is beyond the scope of this course.

[4]See Appendix A for a *little* more explanation if you are really interested in knowing why.

It has been observed empirically that if the filter has impulse response $h(t)$ and transfer function $H(f)$, then σ^2 is *proportional* to the energy in the "signal" $h(t)$, that is,

$$\text{for all choices of } T_0,\ \mathsf{E}[(\mathbb{N}(T_0))^2] = \mathsf{var}(\mathbb{N}(T_0)) = \sigma^2 = \frac{N_0}{2}\int_{-\infty}^{\infty} |h(t)|^2\,dt = \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2\,df \tag{3.1}$$

where N_0 is a constant. Under these circumstances, the channel noise process is called a *white Gaussian noise process with two-sided power spectral density* $N_0/2$, and the channel is referred to as an *additive white Gaussian noise (AWGN) channel*. Of course, N_0 is the value of the *one-sided* power spectral density of the noise. Why is the constant of proportionality in (3.1) written as $\frac{N_0}{2}$ instead of just plain N_0? Well, N_0 is called the *noise power per unit bandwidth* measured in watts/Hz or volts^2/Hz in the sense that the noise power $\mathsf{E}[(\mathbb{N}(T_0))^2] = \sigma^2$ at the output of an ideal lowpass or bandpass filter of bandwidth B Hz and unit gain in the pass band is N_0B watts or volts^2. But since the pass band of the lowpass filter extends from $-B$ Hz to $+B$ Hz, (and the pass band of the bandpass filter of center frequency f_c extends from $-f_c - \frac{B}{2}$ Hz to $-f_c + \frac{B}{2}$ Hz, and from $f_c - \frac{B}{2}$ Hz to $f_c + \frac{B}{2}$ Hz) we need to use $\frac{N_0}{2}$ in (3.1) to get the pretty answer N_0B for the noise power.

3.2 The Communications Receiver with Linear Filter

To summarize the discussion in Section 3.1, our receiver model as shown in the figure in Section 3.1 is as follows. There are two possible equally likely received signals $s_0(t)$ and $s_1(t)$. The actual received signal, denoted by $s_i(t)$, $i = 0, 1$, is passed through a linear filter with impulse response $h(t)$ and transfer function $H(f)$. The output of the filter when its input is $s_i(t)$ is denoted by $\hat{s}_i(t)$. There must be at least one time instant T_0 at which $\hat{s}_0(T_0) \neq \hat{s}_1(T_0)$ (else why bother transmitting the signals?), and this in turn implies that $s_0(t) - s_1(t)$ is a signal with nonzero energy. In addition to one of the $s_i(t)$, there is (white Gaussian) noise with two-sided power spectral density $N_0/2$ at the filter input, and this noise passes through the filter to produce Gaussian noise $\mathbb{N}(t)$ at the output. Regardless of the choice of sampling time T_0, $\mathbb{N}(T_0)$ is a zero-mean Gaussian random variable with variance

$$\sigma^2 = \frac{N_0}{2}\int_{-\infty}^{\infty} |h(t)|^2\,dt = \frac{N_0}{2}\int_{-\infty}^{\infty} |H(f)|^2\,df.$$

Thus, the sample value $\mathbb{Y}$ is given by

$$\mathbb{Y} = \begin{cases} \hat{s}_0(T_0) + \mathbb{N}(T_0), & \text{if a 0 is transmitted,} \\ \hat{s}_1(T_0) + \mathbb{N}(T_0), & \text{if a 1 is transmitted,} \end{cases}$$

where $\mathbb{N}(T_0) \sim \mathcal{N}(0, \sigma^2)$. The receiver compares the sample value $\mathbb{Y}$ to the threshold $\theta = \dfrac{\hat{s}_0(T_0) + \hat{s}_1(T_0)}{2}$. Assume that $\hat{s}_0(T_0) > \hat{s}_1(T_0)$ which we can do without loss of generality.[5] Thus the receiver decides that a 0 was transmitted if $\mathbb{Y} > \theta$ and that a 1 was transmitted if $\mathbb{Y} < \theta$. As we saw in Lecture 2, the error probability achieved by the receiver is

$$P_e = Q\left(\frac{\hat{s}_0(T_0) - \hat{s}_1(T_0)}{2\sigma}\right) = Q(\mathsf{SNR}) \quad \text{where} \quad \mathsf{SNR} = \frac{\hat{s}_0(T_0) - \hat{s}_1(T_0)}{2\sigma}. \tag{3.2}$$

Since $Q(x)$ is a decreasing function x, we can reduce the value of P_e by maximizing SNR, and one obvious way of increasing SNR is to choose T_0 to be the time instant at which $\hat{s}_0(T_0) - \hat{s}_1(T_0)$ is as large as possible.[6] We shall assume henceforth that the sampling instant has indeed been chosen in accordance with this criterion.

[5] If $\hat{s}_0(t) < \hat{s}_1(t)$ for all t, then include an additional inverter in the filter, i.e. reverse the polarity of the filter output, or interchange the roles of s_0 and s_1, or reverse the inequalities in the next sentence in the text and replace the argument of Q in (3.2) by its negative.

[6] Remember that σ does not depend on the choice of sampling instant.

How small can the error probability P_e be? Equivalently, how large can SNR be? To study this question, we begin by noting that the numerator $\hat{s}_0(T_0) - \hat{s}_1(T_0)$ in the expression for SNR in (3.2) is the difference of two convolution integrals and thus can be expressed as

$$\hat{s}_0(T_0) - \hat{s}_1(T_0) = \int_{-\infty}^{\infty} h(t)[s_0(T_0 - t) - s_1(T_0 - t)]\, dt = \int_{-\infty}^{\infty} h(t)s(t)\, dt \tag{3.3}$$

where $s(t) = s_0(T_0 - t) - s_1(T_0 - t)$ and we know that the integral has positive value. As noted previously, the simple expedient of replacing $h(t)$ by $\lambda h(t)$ where $\lambda > 1$, that is, increasing the filter gain, does increase the value of the numerator in (3.2) by a factor of λ but, as can be seen from (3.1), the denominator also increases by the same factor λ. In other words, changing the filter gain does not change SNR or P_e. So, how large can SNR be? We can get an upper bound on SNR by invoking the Schwarz Inequality (sometimes called the Cauchy-Schwarz Inequality). Appendix B contains a detailed description of this result and a careful statement and proof. Here, we simply use the end result which for our application can be expressed as

$$|\hat{s}_0(T_0) - \hat{s}_1(T_0)|^2 = \left| \int_{-\infty}^{\infty} h(t)s(t)\, dt \right|^2 \leq \int_{-\infty}^{\infty} |h(t)|^2\, dt \int_{-\infty}^{\infty} |s(t)|^2\, dt \tag{3.4}$$

Since (3.1) gives that $4\sigma^2 = 2N_0 \displaystyle\int_{-\infty}^{\infty} |h(t)|^2\, dt$, we conclude that

$$\frac{|\hat{s}_0(T_0) - \hat{s}_1(T_0)|^2}{4\sigma^2} = \mathsf{SNR}^2 \leq \frac{1}{2N_0} \int_{-\infty}^{\infty} |s(t)|^2\, dt$$

Now, a simple change of variables $\tau = T_0 - t$ gives that

$$\begin{aligned} \int_{-\infty}^{\infty} |s(t)|^2\, dt &= \int_{-\infty}^{\infty} |s_0(T_0 - t) - s_1(T_0 - t)|^2\, dt = \int_{-\infty}^{\infty} |s_0(\tau) - s_1(\tau)|^2\, d\tau \\ &= \int_{-\infty}^{\infty} [s_0(\tau)]^2\, d\tau + \int_{-\infty}^{\infty} [s_1(\tau)]^2 - 2\int_{-\infty}^{\infty} s_0(\tau)s_1(\tau)\, d\tau \\ &= \mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1 \rangle \end{aligned}$$

where $\mathcal{E}_0$ and $\mathcal{E}_1$ respectively denote the *energy* in the given received signals $s_0(t)$ and $s_1(t)$ and $\langle s_0, s_1 \rangle$ denotes their *inner product*. Note that $\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1 \rangle$ is the energy in the *difference signal* $s_0(t) - s_1(t)$. It follows that

$$\mathsf{SNR} \leq \mathsf{SNR}^* = \sqrt{\frac{\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1 \rangle}{2N_0}}, \tag{3.5}$$

$$P_e \geq P_e^* \quad = Q\left(\sqrt{\frac{\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1 \rangle}{2N_0}} \right) \tag{3.6}$$

For given signals $s_0(t)$ and $s_1(t)$, regardless of what filter, what sampling time, and what threshold the receiver uses, SNR cannot exceed the upper bound SNR^* shown in (3.5) and the error probability cannot be smaller than the lower bound P_e^* shown in (3.6). Notice that we can compute the values of SNR^* and P_e^* just from knowledge of the signals $s_0(t)$ and $s_1(t)$ and the parameter N_0, that is, without knowing or taking into account the details of any particular receiver e.g. the filter impulse response, the filter outputs $\hat{s}_0(t)$ and $\hat{s}_1(t)$, the optimum sampling time, the minimax threshold, the noise variance, etc. All we need to do is to compute the signal energies $\mathcal{E}_i$, the inner product $\langle s_0, s_1 \rangle$, and substitute into right sides of (3.5) and (3.6) to obtain the values of SNR^* and P_e^*. *All* receivers will have SNR no larger than SNR^* and achieve error probability no smaller than P_e^*. Furthermore, since it *is* possible that equality holds in the Schwarz Inequality, it is in fact possible to design a receiver that achieves SNR^* and P_e^*. Such a receiver is called an *optimum receiver* and its receiver filter is called a *matched filter*. The concept of a matched filter is discussed in the next section. But note that we can find the error probability achieved by the optimum receiver *without* first finding the matched filter and figuring out its sampling time, threshold etc. The optimum receiver has

a signal-to-noise ratio of SNR^* and error probability P_e^* and the values of these quantities depend *only* on the signal energies, the inner product of the signals, and the parameter N_0.

It is worth emphasizing that SNR^* and P_e^* depend on the signals $s_0(t)$ and $s_1(t)$ only through the energy of the *difference signal* $s_0(t) - s_1(t)$. The exact *signal shapes and frequencies* of $s_0(t)$ and $s_1(t)$ do not matter except insofar as they affect the energy of the difference signal. For any given energy budget, it is possible to increase SNR by careful choice of $s_0(t)$ and $s_1(t)$ that maximizes the energy in the difference signal, and thus, with use of the appropriate matched filter receiver, further reduce the error probability.

3.3 The Matched Filter

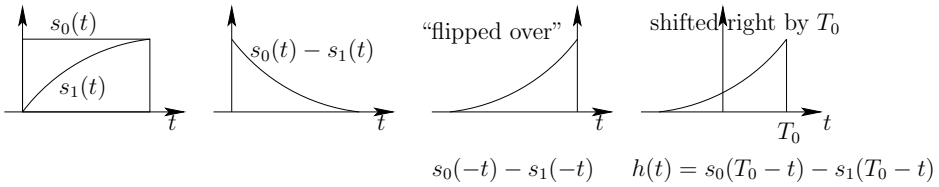
From (3.3) and (3.1), we see that for given signals $s_0(t)$ and $s_1(t)$, given receiver filter with impulse response $h(t)$ and given sampling time T_0, the signal-to-noise ratio SNR is given by

$$\mathsf{SNR} = \sqrt{\frac{\langle h, s\rangle}{2N_0\langle h, h\rangle}} \leq \sqrt{\frac{\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1\rangle}{2N_0}} = \mathsf{SNR}^*$$

where $s(t) = s_0(T_0 - t) - s_1(T_0 - t)$. According to a simplified version of the Schwarz Inequality as given in Appendix B, equality holds in the inequality above if there is a positive constant λ such that $h(t) = \lambda s(t)$. In other words, the maximum signal-to-noise ratio SNR^* and the minimum error probability P_e^* is obtained when the filter impulse response $h(t)$ is a positive multiple of $s(t) = s_0(T_0 - t) - s_1(T_0 - t)$. Thus, for given signals $s_0(t)$ and $s_1(t)$, if we pick our favorite sampling time T_0 and our favorite positive number λ and design the filter receiver to have impulse response $h(t) = \lambda s(t) = \lambda[s_0(T_0 - t) - s_1(T_0 - t)]$, then we shall have constructed an *optimum receiver* for these signals in additive white Gaussian noise. Here, *optimum receiver* means a receiver that achieves the minimum possible error probability. In fact, the receiver that we have described is optimum not just among all possible receivers that use linear filters, but among all possible receivers, whether with linear filters or nonlinear filters, or any other kind of processing of the sample value(s), i.e. with structures different from the one described in Section 3.1 above. As stated before, we shall not be proving this fact in this course, but it is nonetheless important to know and remember. The optimum receiver for signals in additive white Gaussian noise is a *linear* receiver: filtering, sampling, and comparing to a threshold is all that is required.

A filter with impulse response $\lambda[s_0(T_0 - t) - s_1(T_0 - t)]$, $\lambda > 0$, is said to be *matched* to the signals $s_0(t)$ and $s_1(t)$ at time T_0, or to be a *matched filter* for the signals $s_0(t)$ and $s_1(t)$ for sampling time T_0.

So, what does a matched filter impulse response look like? It is just the difference signal $s_0(t) - s_1(t)$ *reversed in time* or "flipped over" with respect to the vertical axis, *shifted to the right by* T_0, and *amplified (or attenuated)* by the positive factor λ. This is illustrated in the figure below for $\lambda = 1$.



As noted several times, the gain factor λ is irrelevant as far as the error probability analysis is concerned, and so we can simply set it to 1 for convenience. Note also that the specification of the optimum sampling time is built into the definition of matched filter above. *We* choose a convenient sampling time T_0 and the definition of the matched filter for sampling time T_0 guarantees maximum SNR, that is, maximum separation of the signals $\hat{s}_0$ and $\hat{s}_1$, at our chosen sampling time T_0. Indeed, the impulse responses of the matched filters for sampling time T_0 and T_1 differ only in how far to the right the "flipped over" signal $s_0(-t) - s_1(-t)$ is shifted. For this reason, some people apply the name matched filter *only* to the filter with impulse response $h(t) = s_0(-t) - s_1(-t)$, which corresponds to the special case $\lambda = 1$ and $T_0 = 0$ in the definition above. The claim is that a filter that maximizes SNR at time T_0 is just this canonical matched filter (which maximizes

SNR at 0) followed by a pure delay of T_0 to make this maximum occur at T_0 instead of 0. This is certainly very convenient for analysis purposes since we don't have to carry around an additional parameter T_0, but it should always be remembered that this canonical form of the matched filter is generally a *noncausal* filter. Indeed, if $s_0(t)$ and $s_1(t)$ have support $[0,T)$ (say), then any filter matched to $s_0(t)$ and $s_1(t)$ at any time $T_0 < T$ will be a noncausal filter. See the figure on the previous page for an illustration of this point.

For convenience, let us consider the canonical matched filter receiver where the filter impulse response is $h(t) = s_0(-t) - s_1(-t)$ and the sampling time is 0. What are the signal outputs $\hat{s}_0(0)$ and $\hat{s}_1(0)$ and the noise variance σ^2? Noting that $h(-t) = s_0(t) - s_1(t)$, we readily get

$$\hat{s}_0(0) = \int_{-\infty}^{\infty} h(0-t)s_0(t)\,dt = \int_{-\infty}^{\infty} [s_0(t) - s_1(t)]s_0(t)\,dt = \mathcal{E}_0 - \langle s_0, s_1\rangle \tag{3.7}$$

$$\hat{s}_1(0) = \int_{-\infty}^{\infty} h(0-t)s_1(t)\,dt = \int_{-\infty}^{\infty} [s_0(t) - s_1(t)]s_1(t)\,dt = \langle s_0, s_1\rangle - \mathcal{E}_1 \tag{3.8}$$

$$\sigma^2 = \frac{N_0}{2}\int_{-\infty}^{\infty} |h(t)|^2\,dt \quad = \frac{N_0}{2}\int_{-\infty}^{\infty} [s_0(t) - s_1(t)]^2\,dt = \frac{N_0}{2}\Big[\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1\rangle\Big] \tag{3.9}$$

Consequently,

$$\hat{s}_0(0) - \hat{s}_1(0) = \mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1\rangle, \quad 2\sigma = \sqrt{2N_0[\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1\rangle]},$$

and

$$\mathsf{SNR} = \frac{\hat{s}_0(0) - \hat{s}_1(0)}{2\sigma} = \frac{\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1\rangle}{\sqrt{2N_0[\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1\rangle]}} = \sqrt{\frac{\mathcal{E}_0 + \mathcal{E}_1 - 2\langle s_0, s_1\rangle}{2N_0}}.$$

Note also that the minimax threshold is

$$\theta = \frac{\hat{s}_0(0) + \hat{s}_1(0)}{2} = \frac{\mathcal{E}_0 - \mathcal{E}_1}{2}.$$

Thus, for *equal-energy* signals, including as a special case, *antipodal* signals $s_1(t) = -s_0(t)$, the threshold is 0. From the practical viewpoint, a threshold of 0 is very important, especlally in communication systems in which the signal strength (and hence received energy) can change as the spatial distance between transmitter and receiver varies, or the signals are subject to *fading*. To set a nonzero threshold in such systems would need additional circuitry to estimate the signal energies $\mathcal{E}_0$ and $\mathcal{E}_1$ and an adjustable threshold that would require resetting on a continuing basis. For this reason (as well as other reasons that we will explore in the next Lecture) equal energy or antipodal signals are preferred in digital communication systems.

Appendix A: All That You Never Wanted to Know About Random Processes – And Were Desperately Hoping That You Would Never Need to Learn

The material in this section is not required reading for ECE 361.

Fundamental notions

The mathematical theory of *random processes*, also called *stochastic processes*, is the basis of the study of noise and its effects on communication systems. What is observed physically is a noise waveform $x(t)$. The mathematical model for this is a *collection* of random variables, one for each time instant t, and this collection of random variables is called a random process. It is common in the engineering literature to denote the collection as $\{\mathbb{X}(t)\}$ where the { and } indicate that it is a *set* that is under consideration, and even more common to simply abuse notation and write "the noise process $\mathbb{X}(t)$". Mathematicians usually write $\{\mathbb{X}_t\}$. What is meant in either case is that we have a collection of random variables whose generic name (or family name) is $\mathbb{X}$, and $\mathbb{X}(t)$ (or $\mathbb{X}_t$) is the random variable for time instant t, that is, t is the first name that uniquely identifies the random variable under discussion among the whole family bearing the name $\mathbb{X}$. So, what is the relationship between $x(t)$ and $\mathbb{X}(t)$ or $\{\mathbb{X}(t)\}$? In probability theory, when the experiment is performed and outcome ω is observed, each random variable defined on the experiment is observed to take on a numerical value: the random variable $\mathbb{Y}$ takes on value $\mathbb{Y}(\omega)$ which is just some real number. If the experiment were repeated again with different outcome $\hat{\omega}$, $\mathbb{Y}$ would take on value $\mathbb{Y}(\hat{\omega})$. Thus, $x(t_1)$ is the value taken on by the random variable $\mathbb{X}(t_1)$ *on that particular trial of the experiment*, $x(t_2)$ is the value taken on by the random variable $\mathbb{X}(t_2)$ *on that particular trial of the experiment*, and so on. The entire waveform $x(t)$ that we have observed is thus comprised of the values taken on by all the random variables in the random process *on that particular trial of the experiment*. The waveform $x(t)$ is called the *realization* of the random process on that trial of the experiment. If the experiment were to be repeated, we would observe a different waveform $\hat{x}(t)$(that is, we would observe a different realization) corresponding to the values taken on by the random variables in the set $\{\mathbb{X}(t)\}$ on the new trial of the experiment.

Probability Distribution Functions

The model of a noise process as a random process (or collection of random variables) naturally leads to the questions: What are the pdfs of all these random variables? What are the two-dimensional joint pdfs? the three-dimensional pdfs? and so on. For the noise processes that plague digital communication systems, experimental observation shows that it is reasonable to model the random variables as Gaussian random variables. We will always assume that all the random variables in a noise process are zero-mean (jointly) Gaussian random variables[7] with common variance $\sigma^2_{\mathbb{X}}$. Such processes are called *Gaussian random processes* (What else would you call them?). Another property observed experimentally for the noise processes that occur in communication systems is that $\mathsf{cov}(\mathbb{X}(t), \mathbb{X}(t+\tau))$, the *covariance* of the two random variables $\mathbb{X}(t)$ and $\mathbb{X}(t+\tau)$ depends only on τ, the separation in time of the two random variables, and not at all on the value of t. For example, taking $\tau = 2$, we have that $\mathsf{cov}(\mathbb{X}(1), \mathbb{X}(3)) = \mathsf{cov}(\mathbb{X}(4), \mathbb{X}(6))$. Now, since $\mathbb{X}(t)$ and $\mathbb{X}(t+\tau)$ are jointly Gaussian random variables, their joint pdf depends only on their means (0 for both), their variances ($\sigma^2_{\mathbb{X}}$ for both), and their correlation coefficient which is just $\mathsf{cov}(\mathbb{X}(t), \mathbb{X}(t+\tau))/\sigma^2_{\mathbb{X}}$. So, if we know the function $R_{\mathbb{X}}(\tau) = \mathsf{E}[\mathbb{X}(t)\mathbb{X}(t+\tau)] = \mathsf{cov}(\mathbb{X}(t), \mathbb{X}(t+\tau))$ for all τ, we can write down the joint pdf of any two random variables in the random process. Notice, for example, that the joint pdf of $\mathbb{X}(1)$ and $\mathbb{X}(3)$ is the same as the joint pdf of $\mathbb{X}(4)$ and $\mathbb{X}(6)$. More generally, the joint pdf of n zero-mean Gaussian random variables $\mathbb{X}(t_1)$, $\mathbb{X}(t_1)$, ..., $\mathbb{X}(t_n)$ of common variance $\sigma^2_{\mathbb{X}}$ can be written down if we know the $\binom{n}{2}$ covariances, and all these covariances are known if $R_{\mathbb{X}}(\tau)$ is known for all τ. Note that the covariances depend only on the time differences $t_i - t_j$. Thus, the joint pdf of $\mathbb{X}(t_1)$, $\mathbb{X}(t_1)$, ..., $\mathbb{X}(t_n)$ is the same as the joint pdf of $\mathbb{X}(t_1 + T)$, $\mathbb{X}(t_1 + T)$, ..., $\mathbb{X}(t_n + T)$ because the increase of T in *all* the time instants leaves the time differences unchanged. Random processes (whether Gaussian or not) with the property that the joint pdf remains unchanged if all the time instants are increased by the same amount, for example $(1, 2, 8.5) \to (3, 4, 10.5)$, are called *stationary* random processes.

[7] The Gaussian assumption can also be justified by invoking the central limit theorem of probability theory. Thermal noise, the dominant source of noise in electronic circuits, is caused by the random motion of a very large number of electrons within conductors, and the observed voltage at time t is due to the net electric field created by the random positions at time t of the electrical charges carried by the electrons. The central limit theorem says that the limiting form of the distribution of the voltage is a Gaussian distribution, and the number of electrons is large enough that the actual distribution is very close to the limit.

Wide-sense Stationary Random Processes
A random process $\{\mathbb{X}(t)\}$ is said to be a wide-sense stationary (WSS) random process if

- All the random variables have the same mean: $\mathsf{E}[\mathbb{X}(t)] = \mu_{\mathbb{X}}$ for all t, and
- $\mathsf{E}[\mathbb{X}(t)\mathbb{X}(t+\tau)]$ is a function of τ but does not depend at all on the value of t.

Note that the noise processes typically encountered in communication systems are WSS random processes with the additional properties that the mean $\mu_{\mathbb{X}}$ is zero, and all the random variables are jointly Gaussian random variables. It is usual to denote $\mathsf{E}[\mathbb{X}(t)\mathbb{X}(t+\tau)]$ by $R_{\mathbb{X}}(\tau)$, and this function is called the *autocorrelation function* of the WSS random process. $R_{\mathbb{X}}(\tau)$ is an even function of τ with a maximum at $\tau = 0$. In fact, $|R_{\mathbb{X}}(\tau)|$ is strictly smaller than $R_{\mathbb{X}}(0)$ for all $\tau \neq 0$, unless $R_{\mathbb{X}}(\tau)$ is a periodic function of τ in which case the peak repeats at periodic intervals and the strict inequality holds between the peaks. Note also that $R_{\mathbb{X}}(0) = \mathsf{E}[(\mathbb{X}(t))^2] = \mathsf{var}(\mathbb{X}(t)) + \mu_{\mathbb{X}}^2$ and thus all the random variables in the process have the same variance $\sigma_{\mathbb{X}}^2 = R_{\mathbb{X}}(0) - \mu_{\mathbb{X}}^2$. For any random process, $\mathsf{E}[(\mathbb{X}(t))^2]$ is called the *total instantaneous power in the process at time* t, and since $\mathsf{E}[(\mathbb{X}(t))^2] = R_{\mathbb{X}}(0)$ has the same value for all t, a WSS random process delivers the same total (instantaneous) power at all time instants t.

The Fourier transform $S_{\mathbb{X}}(f)$ of the autocorrelation function $R_{\mathbb{X}}(\tau)$ of a WSS random process is called the (two-sided) *power spectral density* (PSD) of the process. $S_{\mathbb{X}}(f)$ is a *real-valued*, *even*, and *nonnegative* function of f, and it describes how the power in the process is distributed as a function of frequency. The word "density" is important here: there is no power *at* a frequency f_0 unless $S_{\mathbb{X}}(f)$ has impulses at $\pm f_0$.[8] Assuming that there are no impulses at $\pm W_1$ or $\pm W_2$, the power in the process between W_1 Hz and W_2 Hz is the *area* under the PSD in that band,[9] namely,

$$\int_{-W_2}^{-W_1} S_{\mathbb{X}}(f)\, df + \int_{W_1}^{W_2} S_{\mathbb{X}}(f)\, df = 2\int_{W_1}^{W_2} S_{\mathbb{X}}(f)\, df$$

since $S_{\mathbb{X}}(f)$ is an even function of f. If the PSD does contains impulses at $\pm W_1$ or $\pm W_2$, then we have to be careful about whether to include or exclude their contribution to the integrals depending on what exactly we mean by the phrase "between W_1 Hz and W_2 Hz".[10] As a special case, the *total power in the process* is

$$R_{\mathbb{X}}(0) = \int_{-\infty}^{\infty} S_{\mathbb{X}}(f)\, df$$

which is the total area under the PSD.

Note that nothing has been said about the pdfs of the random variables in a WSS random process: $\mathbb{X}(t_1)$ and $\mathbb{X}(t_2)$ might well have different pdfs though of course the WSS property requires that they have the same mean $\mu_{\mathbb{X}}$ and the same variance $\sigma_{\mathbb{X}}^2$.) Thus, a random process can be wide-sense stationary without necessarily being stationary in the sense discussed earlier of the pdfs not changing when the time instants are increased by the same amount. On the other hand, a stationary process is always wide-sense stationary. This is because for a stationary process, the pdf of $\mathbb{X}(t)$ is the same for all choices of t, and so $\mathsf{E}[\mathbb{X}(t)]$ is the same for all t. Also, the joint pdf of $\mathbb{X}(t_1)$ and $\mathbb{X}(t_1+\tau)$ is the same as the joint pdf of $\mathbb{X}(t_2)$ and $\mathbb{X}(t_2+\tau)$ for all choices of t_1 and t_2, and so $\mathsf{E}[\mathbb{X}(t_1)\mathbb{X}(t_1+\tau) = \mathsf{E}[\mathbb{X}(t_2)\mathbb{X}(t_2+\tau)$, that is, $\mathsf{E}[\mathbb{X}(t)\mathbb{X}(t+\tau)]$ does not depend on t, only on τ. For example, the Gaussian noise processes discussed above are stationary random processes, and thus they are also WSS random processes. Conversely, if a random process is known to be a WSS random process and we make the further assumption that it is a Gaussian random process – meaning that all the random variables in the process are jointly Gaussian random variables – then the process is also a stationary random process.

[8] As a special case, note that if $\mu_{\mathbb{X}} \neq 0$, then $S_{\mathbb{X}}(f)$ contains an impulse $\mu_{\mathbb{X}}^2\delta(f)$ at the origin.

[9] This is analogous to the concept of the probability *density* function (pdf) of a random variable. Recall that the value of a pdf at a point α is *not* the probability that the random variable has value α – in fact, a pdf can have value greater than 1 – and the probability that the random variable takes on value in $[\alpha, \beta]$ is the *area* under the pdf curve in the interval $[\alpha, \beta]$.

[10] Returning to the analogy with random variables, recall that if $\mathbb{Z}$ is a discrete random variable or a *mixed* random variable that takes on discrete values with nonzero probabilities and exhibits continuous behavior elsewhere, we cannot cavalierly ignore the difference between $<$ and $\leq$ signs since, for example, $P\{\alpha < \mathbb{Z} < \beta\}$ need not be the same as $P\{\alpha \leq \mathbb{Z} \leq \beta\}$: the two probabilities would always be equal if $\mathbb{Z}$ were a continuous random variable.

Random Processes in Linear Systems

Suppose that a linear time-invariant system has impulse response $h(t)$ and transfer function $H(f)$. What is the output of this system if its input is a realization $x(t)$ of a random process $\{\mathbb{X}(t)\}$? Well, $x(t)$ is just some function of time (input signal), and so the output of the linear system is just some other time function $y(t)$ where of course $y = x \star h$. Now, subject to some mild restrictions that always hold in actual systems, the response $y(t)$ to the input $x(t)$ is a realization of another random process $\{\mathbb{Y}(t)\}$. It is as if the random process $\{\mathbb{X}(t)\}$ is passing through the linear system and producing the random process $\{\mathbb{Y}(t)\}$ at the output. In fact, the random variable $\mathbb{Y}(t)$ can be expressed as

$$\mathbb{Y}(t) = \int_{-\infty}^{\infty} \mathbb{X}(\tau)h(t-\tau)\,d\tau = \int_{-\infty}^{\infty} \mathbb{X}(t-\tau)h(\tau)\,d\tau \tag{3.A.1}$$

which looks remarkably like a convolution integral. But, what does it mean? Well, an ordinary integral $\int_a^b g(t)\,dt$ is the *limit of a Riemann sum* of the form $\sum_{i=0}^{n-1} g(t_i)\Delta t_i$. The Riemann sums whose limits give the integrals on the right side of (3.A.1) are *weighted linear combinations of a finite number of random variables* in the input process $\{\mathbb{X}(t)\}$. Thus, $\mathbb{Y}(t)$ is the limit of a weighted sum $\sum_{i=0}^{n-1} \mathbb{X}(t-(t_i))h(t_i)\Delta t_i$ of random variables from the process $\{\mathbb{X}(t)\}$ with weights given by the values of $h(t)$ and the Δt_i's. As a special case, suppose that $\{\mathbb{X}(t)\}$ is a Gaussian random process. Then, since linear combinations of Gaussian random variables are Gaussian random variables (and this holds true in the limit as well), $\{\mathbb{Y}(t)\}$ is also a Gaussian random process. If this sounds very complicated and difficult to understand, do not be alarmed. Fortunately for us, an in-depth understanding of the meaning of (3.A.1) is *not required* for carrying out many of the computations needed in the analysis of digital communication systems.

Wide-Sense Stationary Random Processes in Linear Systems

Now suppose that $\{\mathbb{X}(t)\}$ is a WSS random process. Then, the output random process $\{\mathbb{Y}(t)\}$ is also a WSS process whose mean $\mu_{\mathbb{Y}}$, autocorrelation function $R_{\mathbb{Y}}(t)$, and power spectral density $S_{\mathbb{Y}}(f)$ are all related to $\mu_{\mathbb{X}}$, $R_{\mathbb{X}}(t)$, and $S_{\mathbb{X}}(f)$ via $h(t)$ and $H(f)$. In particular, we have

$$\mu_{\mathbb{Y}} = \mu_{\mathbb{X}} \star h \qquad\qquad = \mu_{\mathbb{X}} \int_{-\infty}^{\infty} h(t)\,dt, \tag{3.A.2}$$

$$R_{\mathbb{Y}}(\tau) = R_{\mathbb{X}} \star (h \star \tilde{h}) = R_{\mathbb{X}} \star R_h = \int_{-\infty}^{\infty} R_{\mathbb{X}}(\tau - t)R_h(t)\,dt, \tag{3.A.3}$$

$$S_{\mathbb{Y}}(f) = S_{\mathbb{X}}(f)|H(f)|^2. \tag{3.A.4}$$

Here, $\tilde{h}(t) = h(-t)$ is the impulse response reversed in time whose Fourier transform is $\tilde{H}(f) = H^*(f)$, and $R_h(\tau) = (h \star \tilde{h})$ is the *autocorrelation function* $\int_{-\infty}^{\infty} h(t+\tau)h(t)\,dt$ of the impulse response $h(t)$. There are several consequences to these results that are worth noting and remembering.

- If $\mu_{\mathbb{X}} = 0$, then $\mu_{\mathbb{Y}} = 0$, i.e., if $\{\mathbb{X}(t)\}$ is a zero-mean process, then $\{\mathbb{Y}(t)\}$ also is a zero-mean process. If $H(0) = \displaystyle\int_{-\infty}^{\infty} h(t)\,dt = 0$, then $\mu_{\mathbb{Y}} = 0$, i.e., if the filter does not pass DC, then the output process has zero mean even if the mean of the input process is a nonzero constant.
- Since $R_{\mathbb{Y}}(\tau)$ is an even function of τ, the integral in (3.A.3) can be written as $\displaystyle\int_{-\infty}^{\infty} R_{\mathbb{X}}(t+\tau)R_h(t)\,dt$ showing that $R_{\mathbb{Y}}(\tau)$ also can be thought of as the *cross-correlation function* of $R_{\mathbb{X}}$ and R_h.
- The output process $\{\mathbb{Y}(t)\}$ has no power at frequencies f where $H(f) = 0$, and little power in frequency bands where $|H(f)|$ is small, that is, the linear system is filtering the input process exactly as we expect.
- The total power in the output process is $R_{\mathbb{Y}}(0) = \displaystyle\int_{-\infty}^{\infty} R_{\mathbb{X}}(t)R_h(t)\,dt = \int_{-\infty}^{\infty} S_{\mathbb{X}}(f)|H(f)|^2\,df$.
  If $\mu_{\mathbb{Y}} = 0$ (as is true for all noise processes considered in this course), then we also have

$$\sigma_{\mathbb{Y}}^2 = R_{\mathbb{Y}}(0) = \int_{-\infty}^{\infty} R_{\mathbb{X}}(t)R_h(t)\,dt = \int_{-\infty}^{\infty} S_{\mathbb{X}}(f)|H(f)|^2\,df. \tag{3.A.5}$$

- If $\{\mathbb{X}(t)\}$ is a *zero-mean stationary Gaussian process*, then $\{\mathbb{Y}(t)\}$ is also a zero-mean stationary Gaussian process, and each random variable $\mathbb{Y}(t)$ in the output process has variance $\sigma^2_{\mathbb{Y}}$ given by (3.A.5).

The White Noise Myth

The result of thermal noise at the input of an amplifier (which we model as a linear filter with gain) is a randomly varying output that can be treated as a random process in the sense described above. Experimental observation reveals that it is reasonable to model this output process as a zero-mean WSS process $\{\mathbb{Y}(t)\}$ with autocorrelation function $R_{\mathbb{Y}}(\tau)$ that is proportional to $R_h(\tau)$ where $h(t)$ is the impulse response of the filter.[11] The constant of proportionality is usually written as $N_0/2$ and so we have

$$R_{\mathbb{Y}}(\tau) = \frac{N_0}{2}R_h(\tau); \quad S_{\mathbb{Y}}(f) = \frac{N_0}{2}|H(f)|^2; \quad \text{and} \quad \sigma^2_{\mathbb{Y}} = \frac{N_0}{2}\int_{-\infty}^{\infty}|h(t)|^2\,dt = \frac{N_0}{2}\int_{-\infty}^{\infty}|H(f)|^2\,df. \tag{3.A.6}$$

Why is the constant of proportionality written as $\frac{N_0}{2}$ instead of plain simple N_0? Well, N_0 is the *noise power per unit bandwidth* measured in watts/Hz or volts^2/Hz in the sense that the noise power at the output of an ideal lowpass or bandpass filter of bandwidth B Hz and unit gain in the pass band is N_0B watts or volts^2. But since the pass band of the lowpass filter extends from $-B$ Hz to $+B$ Hz, (and the pass band of the bandpass filter of center frequency f_c extends from $-f_c - \frac{B}{2}$ Hz to $-f_c + \frac{B}{2}$ Hz, and from $f_c - \frac{B}{2}$ Hz to $f_c + \frac{B}{2}$ Hz) we need to use $\frac{N_0}{2}$ in (3.A.6) to get the pretty answer N_0B for the noise power.

Now, if we wish to fit these experimental results into the Procrustean bed of the theory of WSS random processes in linear systems as expounded in (3.A.3)-(3.A.4), then we have to assume that the input process $\{\mathbb{X}(t)\}$ has PSD given by $S_{\mathbb{X}}(f) = \frac{N_0}{2}, -\infty < f < \infty$. According to (3.A.4), with this as the PSD of the input process, the PSD of the output process is $S_{\mathbb{Y}}(f) = S_{\mathbb{X}}(f)|H(f)|^2 = \frac{N_0}{2}|H(f)|^2$ exactly as stated in (3.A.6). Since $R_{\mathbb{X}}(\tau)$ is the inverse Fourier transform of $S_{\mathbb{X}}(f)$ and $\mathcal{F}^{-1}[\frac{N_0}{2}] = \frac{N_0}{2}\delta(\tau)$ where $\delta(\tau)$ is the unit impulse, we conclude that the input process has autocorrelation function $R_{\mathbb{X}}(\tau) = \frac{N_0}{2}\delta(\tau)$. Naturally, (3.A.3) becomes

$$R_{\mathbb{Y}}(\tau) = \int_{-\infty}^{\infty}\frac{N_0}{2}\delta(\tau - t)R_h(t)\,dt = \frac{N_0}{2}R_h(t)$$

as stated in (3.A.6). Similarly, from (3.A.5) we have that

$$\sigma^2_{\mathbb{Y}} = \int_{-\infty}^{\infty}R_{\mathbb{X}}(t)R_h(t)\,dt = \int_{-\infty}^{\infty}\frac{N_0}{2}\delta(t)R_h(t)\,dt = \frac{N_0}{2}R_h(0) = \frac{N_0}{2}\int_{-\infty}^{\infty}|h(t)|^2\,dt$$

as stated in (3.A.6). On the other hand, since the total area under the PSD $S_{\mathbb{X}}(f)$ is infinite, this assumed input process $\{\mathbb{X}(t)\}$ has infinite power!

It should be clear that the input process $\{\mathbb{X}(t)\}$ that we have hypothesized is a myth: it has infinite power! Thus, the input process is a mathematical fiction, but a very useful fiction too – indeed, one that has been adopted by engineers with the greatest delight – that allows us to use (3.A.3)-(3.A.4) consistently to model the experimentally obtained results at the outputs of filters. This mythical input process is termed *white noise*. White noise *does not exist physically*. It cannot be observed (or sampled) by sticking a probe into the filter input because all measuring devices implicitly or explicitly filter whatever they are measuring. All that we can do is observe the result of passing the mythical white noise process through a filter, and this physically observable process is a zero-mean WSS process with PSD proportional to $|H(f)|^2$.

Definition: White noise is a fictitious zero-mean WSS random process that is assumed to exist in the input of a linear time-invariant system (impulse response $h(t)$ and transfer function $H(f)$) whenever the system output includes a zero-mean WSS random process with power spectral density proportional to $|H(f)|^2$. The constant of proportionality is written as $\frac{N_0}{2}$ and is called *the two-sided power spectral density* of the white noise process since $\frac{N_0}{2}$ is the value of the power spectral density of the white noise process for all frequencies f, $-\infty < f < \infty$. The autocorrelation function of the white noise process is $\frac{N_0}{2}\delta(\tau)$ where $\delta(\tau)$ denotes the unit impulse.

[11] In practice, it is often the PSD of the noise that is measured with a spectrum analyzer and found to be proportional to $|H(f)|^2$. That is, $S_{\mathbb{Y}}(f) \propto |H(f)|^2$ which of course implies that $R_{\mathbb{Y}}(\tau) \propto R_h(\tau)$.

Why does all this work? Well, in many cases, the noise present at the filter input has power spectral density that can be approximated as $S_{\mathbb{X}}(f) = \frac{N_0}{2}\text{rect}(f/2f_0)$ where f_0 is on the order of 10^{13} Hz or so. Now, for all practically implementable filters, $|H(f)|^2$ is *very small* for large $|f|$, and thus there is a negligibly small difference between the "exact" value $\frac{N_0}{2}\int_{-f_0}^{+f_0}|H(f)|^2\,df$ of the noise variance at the filter output and the approximate value $\frac{N_0}{2}\int_{-\infty}^{\infty}|H(f)|^2\,df$ given by the white noise assumption. On the other hand, the mathematical manipulations are a lot easier with the white noise assumption, e.g. integrals whose integrands include impulses are easy to evaluate. Thus, with the caveat that we never try to observe white noise physically, we never look very closely at the random variables in the white noise process, and we never try to tap the infinite power of a white noise process to solve the world's energy problems, this fictitious random process can help simplifying the analysis of many problems in digital communication systems.

In most cases, the thermal noise that is omnipresent in electrical circuits tends to produce stationary Gaussian noise processes at the outputs of linear filters, and this can be explained by the assumption that the input process is a stationary Gaussian process. Furthermore, if the output PSD is proportional to $|H(f)|^2$, then the fictitious white noise input process that is assumed to have passed through the filter and resulted in the output process is called a *white Gaussian noise* (WGN) process. White Gaussian noise has all the properties of white noise stated in the boxed text above *and* the additional property that the output process is a stationary Gaussian process. One should not, however, infer that the random variables in the WGN process are themselves Gaussian random variables. Each variable in a white noise process technically has infinite variance which makes it hard to fit into the Gaussian mold of $(1/\sigma\sqrt{2\pi})\exp(-u^2/2\sigma^2)$.

Appendix B: The Schwarz Inequality

The material in this section is not required reading for ECE 361.

The Schwarz Inequality, also known as the Cauchy-Schwarz Inequality, is a bound on the inner product of two vectors. It is often taught in linear algebra courses and usually proved for vectors in the finite-dimensional vector spaces $\mathbb{R}^n$ or $\mathbb{C}^n$, the vector spaces of n-tuples of real numbers or complex numbers respectively. The Schwarz Inequality also applies to finite-energy signals, but the proof in this case needs a little extra care.

Let $x(t)$ and $y(t)$ denote finite-energy signals, that is, $\displaystyle\int_{-\infty}^{\infty}|x(t)|^2\,dt < \infty$ and $\displaystyle\int_{-\infty}^{\infty}|y(t)|^2\,dt < \infty$. The *inner product* of $x(t)$ and $y(t)$ is defined as $\langle x, y\rangle = \displaystyle\int_{-\infty}^{\infty}x(t)[y(t)]^*\,dt$ where * denotes complex conjugation. Note that the inner product is just the *crosscorrelation function* of x and y evaluated at 0. According to the Schwarz Inequality, the inner product of two finite-energy signals has finite magnitude. More specifically,

$$\left|\int_{-\infty}^{\infty}x(t)[y(t)]^*\,dt\right|^2 \le \int_{-\infty}^{\infty}|x(t)|^2\,dt\int_{-\infty}^{\infty}|y(t)|^2\,dt$$

The inner product has the obvious properties (i) $\langle y, x\rangle = [\langle x, y\rangle]^*$, and (ii) $\langle x, \mathbf{0}\rangle = \langle \mathbf{0}, x\rangle = 0$ where $\mathbf{0}(t)$ denotes the signal that has value 0 for all t. The inner product of $x(t)$ with itself is just the energy of the signal $x(t)$, and it is usual to write $\langle x, x\rangle = \| x \|^2$ where $\| x \|$, the positive square root of $\langle x, x\rangle$, is called the *norm* of x. Note that $\| x \| \ge 0$ for all signals x, but that in sharp contrast to the vector spaces $\mathbb{R}^n$ or $\mathbb{C}^n$, $\| x \| = 0$ *does not* imply that $x(t) = \mathbf{0}(t)$. For example, if $x(0) = 1$ and $x(t) = 0$ for all $t \ne 0$, then $\| x \| = 0$, but $x(t) \ne \mathbf{0}(t)$. However, if $\| x \| = 0$, then $x(t)$ *behaves* like $\mathbf{0}(t)$ in inner products in the sense that $\langle x, y\rangle = \langle y, x\rangle = 0$ for all finite-energy signals $y(t)$. We say that $x(t)$ is *equivalent* to $\mathbf{0}(t)$. Similarly, if $\| x - y \| = 0$, then we say that $x(t)$ and $y(t)$ are *equivalent signals* in the sense that one can be substituted for the other in inner products: $\langle x, z\rangle = \langle y, z\rangle$ for all finite-energy signals $z(t)$.

If α and β are constants, then we have the following very useful result whose proof is left as an exercise:

$$\| \alpha x + \beta y \|^2 = |\alpha|^2 \| x \|^2 + 2\Re[\alpha\beta^*\langle x, y\rangle] + |\beta|^2 \| y \|^2 \tag{3.B.1}$$

Now suppose that $\| x \| = 0$. Setting $\alpha = -c[\langle x, y\rangle]^*$ where c is a real number and $\beta = 1$ in (3.B.1), we get

$$\| \alpha x + \beta y \|^2 = 2\Re\big[-c[\langle x, y\rangle]^*\langle x, y\rangle\big] + \| y \|^2 = -c|\langle x, y\rangle|^2 + \| y \|^2 \ge 0.$$

But, $-c|\langle x,y\rangle|^2 + \| y \|^2$ cannot be nonnegative for *every* choice of real number c unless $\langle x,y\rangle = 0$. In summary, $\boxed{\text{If } \| x \| = 0 \text{ then } \langle x,y\rangle = \langle y,x\rangle = 0 \text{ for all finite-energy signals } y(t)}$ as was claimed above. Next, note that $\langle x,z\rangle - \langle y,z\rangle = \langle x-y,z\rangle$, and so if $\| x-y \| = 0$, then $\langle x-y,z\rangle = 0$, that is, $\langle x,z\rangle = \langle y,z\rangle$ for all finite-energy signals $z(t)$: equivalent signals can be substituted one for the other in inner products.

The Schwarz Inequality: If $\| x \| < \infty$ and $\| y \| < \infty$, then

$$|\langle x,y\rangle| \le \| x \| \| y \| . \tag{3.B.2}$$

Furthermore, if $\| y \| > 0$, then equality holds in (3.B.2) if and only if there is a number λ such that $\| x - \lambda y \| = 0$.

Proof: If $\| y \| = 0$, then $\langle x,y\rangle = 0$, and thus both sides of (3.B.2) are 0, that is, not only is (3.B.2) valid, but equality holds in the expression. Next, note that from (3.B.1) we have that for *any* β,

$$\| x - \beta y \|^2 = \| x \|^2 - 2\Re[\beta^* \langle x,y\rangle] + |\beta|^2 \| y \|^2 \ge 0. \tag{3.B.3}$$

If $\| y \| > 0$, we can choose $\beta = \dfrac{\langle x,y\rangle}{\| y \|^2}$ in (3.B.3) and use the fact that $|\beta| = \dfrac{|\langle x,y\rangle|}{\| y \|^2}$ to write (3.B.3) as

$$\| x - \beta y \|^2 = \| x \|^2 - 2\frac{|\langle x,y\rangle|^2}{\| y \|^2} + \frac{|\langle x,y\rangle|^2}{\| y \|^4} \| y \|^2 = \| x \|^2 - \frac{|\langle x,y\rangle|^2}{\| y \|^2} \ge 0. \tag{3.B.4}$$

Re-arranging the inequality on the right and taking square roots, we get the Schwarz Inequality.

Continuing to suppose that $\| y \| > 0$, let us consider the case when x is such that $\| x - \lambda y \| > 0$ for *every* choice of λ. Then, in (3.B.4) above, we have that $\| x - \lambda y \|^2 > 0$ and therefore it must be that $|\langle x,y\rangle| < \| x \| \| y \|$. Thus, strict inequality holds in (3.B.2). Alternatively, if there *is* some number λ such that $\| x - \lambda y \| = 0$, that is, $x - \lambda y$ is *equivalent* to the zero signal $\mathbf{0}(t)$, then

$$\langle x - \lambda y, x\rangle = 0 \Rightarrow \| x \|^2 = \lambda \langle y,x\rangle \quad \text{and} \quad \langle x - \lambda y, y\rangle = 0 \Rightarrow \lambda \| y \|^2 = \langle x,y\rangle$$

and hence

$$\| x \|^2 \lambda \| y \|^2 = \lambda \langle y,x\rangle \langle x,y\rangle = \lambda [\langle x,y\rangle]^* \langle x,y\rangle = \lambda |\langle x,y\rangle|^2 \Rightarrow |\langle x,y\rangle|^2 = \| x \|^2 \| y \|^2 .$$

that is, equality holds in (3.B.2). Thus, if $\| y \| > 0$, then equality holds in the Schwarz Inequality (3.B.2) if and only if there is a number λ such that $\| x - \lambda y \| = 0$. □

Exercise: We canceled λ from $\lambda \| x \|^2 \| y \|^2 = \lambda |\langle x,y\rangle|^2$ (and took square roots) to arrive at the conclusion that $|\langle x,y\rangle| = \| x \| \| y \|$. Such a cancellation is not permissible when $\lambda = 0$. Show that the conclusion that $|\langle x,y\rangle| = \| x \| \| y \|$ nevertheless holds when $\lambda = 0$.

The condition for equality in the Schwarz Inequality is often stated as $x = \lambda y$ but, as we saw above, the slightly weaker condition that there exist a number λ such $x(t)$ and $\lambda y(t)$ are *equivalent* signals suffices: we do not insist that $x(t) = \lambda y(t)$ for all t, $-\infty < t < \infty$.

For real-valued signals $x(t)$ and $y(t)$, the inner product $\langle x,y\rangle$ is also real-valued, and the Schwarz Inequality can be re-written as

$$-\| x \| \| y \| \le \langle x,y\rangle \le \| x \| \| y \| \tag{3.B.5}$$

Now suppose that $\| y \| > 0$ and there is a real number λ such that $\| x - \lambda y \| = 0$. But under these conditions, we saw above that $\lambda \| y \|^2 = \langle x,y\rangle$. and so $\langle x,y\rangle$ is positive or negative or zero according as λ is positive or negative or zero. In other words, $\mathrm{sgn}(\langle x,y\rangle) = \mathrm{sgn}(\lambda)$. Equality must also hold in (3.B.2) under these conditions, and thus we conclude that for real-valued signals, $\langle x,y\rangle = \mathrm{sgn}(\lambda) \| x \| \| y \|$ whenever equality holds in the Schwarz Inequality, that is, $\langle x,y\rangle = \| x \| \| y \|$ when $\lambda > 0$ and $\langle x,y\rangle = -\| x \| \| y \|$ when $\lambda < 0$.