5.1 Wiener Filtering: Optimal Linear Estimation of Random Variables

Before we proceed, we would like to mention that estimation in the form of \( \hat{y} = a x + b \) is not in the strict sense "linear". (Convince yourself of this by showing that the conditions of linearity are not satisfied by the estimate). However, the above transformation is almost linear and is called an affine transformation. If the random variables had no deterministic component (like the mean), then one could easily argue that the constant term \( b \) would not be needed.

One can show that if the deterministic components are removed from the corresponding random quantities (i.e. means subtracted for the case of scalar random variables) linear estimation can applied identically in this modified framework as well. The take home message here is that statistical estimation cannot be used to estimate deterministic components and all determinism should be removed from the problem before using statistical estimation techniques as no operation (linear or non-linear) will transform a random quantity into a deterministic one.

- Let us rewrite the random variables in terms of a deterministic component and a random component as follows,

\[
x = \bar{x} + \mu_x
\]
\[
y = \bar{y} + \mu_y
\]

where \( \bar{x}, \bar{y} \) denote the random variables with their means subtracted (Thus, \( \langle \bar{x} \rangle = \langle \bar{y} \rangle = 0 \)). Define now the covariance of the random variables as \( c_{xy} = \langle \bar{x} \bar{y} \rangle \). The normalized covariance (covariance coefficient) can be expressed as \( \rho_{xy} = \frac{c_{xy}}{\sqrt{(\bar{x})^2}(\bar{y})^2} \). Remember, we had mentioned in class that the dot product of two zero-mean random variables can be defined as the expected value of their product. Thus, \( c_{xy} \) can be likened to be a dot product of \( \bar{x} \) and \( \bar{y} \).

We know that the dot product of two vectors \( \mathbf{a} \) and \( \mathbf{b} \) can be written as,

\[
\mathbf{a} \cdot \mathbf{b} = |\mathbf{a}| |\mathbf{b}| \cos \theta
\]

where \( |\mathbf{a}| = \sqrt{(\mathbf{a} \cdot \mathbf{a})} \) is like the length of the vector and \( \theta \) is the angle between the \( \mathbf{a} \) and \( \mathbf{b} \). Thus, we can write a similar equation for the random variables as

\[
c_{xy} = \sqrt{\langle \bar{x}^2 \rangle \langle \bar{y}^2 \rangle} \rho_{xy}
\]

Now, since \( \sigma_x = \sqrt{\langle \bar{x}^2 \rangle} \), it is clear that \( \rho_{xy} \) can be regarded as equivalent to some measure of the cosine of the angle between the abstract vectors representing \( x \) and \( y \). Note that the cosine of any angle lies between -1 and 1, and so \( \rho_{xy} \) lies in \([-1, 1]\).

The other way to think about this question is like this:

\[
C_{xy} = \langle (x - \mu_x)(y - \mu_y) \rangle
\]

\[
\Rightarrow C_{xy} = \langle xy - x \mu_y - y \mu_x + \mu_x \mu_y \rangle
\]

\[
\Rightarrow C_{xy} = \langle xy \rangle - \mu_x \langle y \rangle - \langle x \rangle \mu_y + \mu_x \mu_y
\]

\[
\Rightarrow C_{xy} = m_{xy} - \mu_x \mu_y
\]

\[^1\text{A more rigorous derivation of the above can be obtained by using the Schwartz inequality which states that for any two vectors } \mathbf{a} \text{ and } \mathbf{b}, |\mathbf{a} \cdot \mathbf{b}| \leq |\mathbf{a}| |\mathbf{b}|.\]
And from the Cauchy-Schwarz inequality, we know that:

\[ \langle xy \rangle^2 \leq \langle x^2 \rangle \langle y^2 \rangle \]

Which means:

\[ C_{xy}^2 = \langle (x - \mu_x)(y - \mu_y) \rangle^2 \leq \langle (x - \mu_x)^2 \rangle \langle (y - \mu_y)^2 \rangle \]

\[ \Rightarrow |C_{xy}| \leq \sigma_x \sigma_y \]

\[ \Rightarrow |\rho_{xy}| \leq 1 \]

• Now our estimate \( \hat{y} \) should be such that all the deterministic component of \( \hat{y} \) can be made equal to the deterministic component of \( y \). This is because if the deterministic components of \( y \) and \( \hat{y} \) were not the same there would be a larger error between them than if they were. Thus, in trying to minimize \( E \) it is appropriate to cancel out the deterministic components first so that the minimization is between statistical quantities. One can argue that this deterministic cancellation does not affect the minimization problem for the random quantities\(^2\). Thus, equating the mean values of \( y \) and \( \hat{y} \) we have,

\[ \mu_y = a \mu_x + b \]

Subtracting this from \( \hat{y} = ax + b \) we have,

\[ \hat{y} = a \hat{x}, \]

where \( \hat{y} = \hat{y} - \mu_y \).

Thus, we can choose the value of \( a \) which minimizes the error between \( \hat{y} \) and \( a \hat{x} \) and use the value of \( a \) to get \( b \). First, we can write the error \( E \) as,

\[ E = \langle (\hat{y} - a \hat{x})^2 \rangle = \langle \hat{y}^2 \rangle + a^2 \langle \hat{x}^2 \rangle - 2a \langle \hat{x} \hat{y} \rangle, \]

\[ E = \sigma_y^2 + a^2 \sigma_x^2 - 2a \rho_{xy} \sigma_x \sigma_y \]

• Differentiating the error \( E \) w.r.t. \( a \) and equating to zero gives the following equation that the optimal \( \hat{a} \) must satisfy,

\[ \frac{dE}{da} = 2a \sigma_x^2 - 2 \rho_{xy} \sigma_x \sigma_y = 0 \]

Also, the second derivative of \( E \) w.r.t. \( a \) is positive \( \left( \frac{d^2E}{da^2} = 2 \sigma_x^2 > 0 \right) \) and so the fixed point \( \hat{a} \) is indeed a minimum. Thus, since the optimal \( \hat{b} \) is related to \( \hat{a} \) we have,

\[ \hat{b} = \mu_y - \hat{a} \mu_x \]

Solving for \( \hat{a} \) and \( \hat{b} \) gives,

\[ \hat{a} = \frac{\rho_{xy} \sigma_y}{\sigma_x} = \frac{C_{xy}}{\sigma_x^2} \]

\[ \hat{b} = \mu_y - \frac{\rho_{xy} \sigma_y \mu_x}{\sigma_x} = \mu_y - \frac{C_{xy}}{\sigma_x^2} \mu_z \]

• The minimum possible error \( E^* \) is, after substituting for \( \hat{a} \),

\[ E^* = \sigma_y^2 (1 - \rho_{xy}^2) \]

\(^2\)There is an obvious attempt here to avoid using the brute force approach for there are some subtle nuances which are revealed using the approach we are pursuing.
Note that, \( E = 0 \) when \( \rho_{xy} = \pm 1 \) and \( E = \sigma_y^2 \) when \( \rho_{xy} = 0 \). This should be intuitively obvious since \( \rho_{xy} = \pm 1 \) mean that the random variables are perfectly correlated and so linearly related to each other. Thus, perfect correlation leads to perfect estimation. On the other hand \( \rho_{xy} = 0 \) means that the random variables have no correlation and so one cannot be estimated from the other, and the error in estimating \( y \) from \( x \) is the same as the error in estimating \( y \) itself as \( x \) conveys no information about \( y \). Since the best MS estimate of \( y \) in the absence of all information is its mean and the error in estimation is equal to the variance of \( y \), \( E^* = \sigma_y^2 \).

Thus, the covariance coefficient in effect captures the ability of one random variable to be estimated linearly from another. It is interesting to know that even when the correlation between two random variables is zero, one can be estimated from another using non-linear estimators.

5.2 Wiener Filtering: Optimal Linear Estimation of Random Processes

- To prove the orthogonality principle without using the fact that the derivative of \( E \) w.r.t. \( a \) is 0, let us assume that \( \hat{a} \) is any arbitrary value of \( a \). Also, let \( \hat{a} \) be a value of \( a \) which satisfies the orthogonality condition. We shall show that \( \hat{a} \) indeed minimizes \( E \). Decompose the error in estimating \( y \) using \( \hat{a} \), \( y - \hat{a}x \) as follows,

\[
y - \hat{a}x = y - \hat{\hat{a}}x + \hat{\hat{a}}x - \hat{\hat{a}}x
\]

The MS error \( \bar{E} \) in estimating using \( \hat{a} \) can be written as,

\[
\bar{E} = \langle (y - \hat{a}x)^2 \rangle = \langle (y - \hat{\hat{a}}x)^2 \rangle + \langle (\hat{\hat{a}} - \hat{a})((y - \hat{a}x)x) \rangle
\]

\[
\implies \langle (y - \hat{a}x)^2 \rangle = \langle (y - \hat{\hat{a}}x)^2 \rangle + \langle ((\hat{\hat{a}} - \hat{a})x)^2 \rangle
\]

\[
\implies \langle (y - \hat{\hat{a}}x)^2 \rangle \geq \langle (y - \hat{a}x)^2 \rangle
\]

Note that the expectation is taken over all possible \( x, y \) pairs. The last term in the above expression is zero as \( \hat{a} \) satisfies the orthogonality condition. Also, the second term is always positive unless \( \hat{a} = \hat{\hat{a}} \). Thus, \( \bar{E} \geq E \) for any arbitrary \( a \) and so \( \hat{a} \) is the optimal value.

- \( R_{zs}(\tau) = \langle x(t)s(t - \tau) \rangle = \langle x(t) \int_{t_1}^{\infty} h(t_1)x(t - \tau - t_1)dt_1 \rangle \)

\[
\implies R_{zs}(\tau) = \int_{-\infty}^{\infty} h(t_1)x(t - \tau - t_1)dt_1
\]

\[
\implies R_{zs}(\tau) = \int_{-\infty}^{\infty} h(t_1)R_{xz}(\tau + t_1)dt_1
\]

Let \( t_2 = -t_1 \)

\[
\implies R_{zs}(\tau) = \int_{-\infty}^{\infty} h(-t_2)R_{xz}(\tau - t_2)d(-t_2)
\]

\[
\implies R_{zs}(\tau) = h(-\tau) * R_{xz}(\tau)
\]

- \( R_{ss}(\tau) = \langle s(t)s(t - \tau) \rangle = \langle (\int_{-\infty}^{\infty} h(t_1)x(t - t_1)dt_1)s(t - \tau) \rangle \)

\[
\implies R_{ss}(\tau) = \int_{-\infty}^{\infty} h(t_1)x(t - t_1)s(t - \tau)dt_1
\]

\[
\implies R_{ss}(\tau) = \int_{-\infty}^{\infty} h(t_1)R_{sz}(\tau - t_1)dt_1
\]
\[ \Rightarrow R_{zz}(\tau) = h(\tau) * R_{zz}(\tau) \]

- After deriving the equations for the toy problem above, we shall now show how easily everything extends to the more complicated case of estimating one random process from another. The dot product in the case of scalar random variables was just the expected value of their product. However, a random process is an indexed set of random variables and the dot product of two random processes must be defined for all values of their indices. Thus, for two random processes \( x(t) \) and \( y(t) \) to be orthogonal we must have,

\[
\langle x(\tau) y(t) \rangle = 0 \quad \forall \, \tau, t
\]

where the expectation is taken over all possible combinations of signal waveform pairs \( x(t), y(t) \) for \(-\infty < t < \infty\). If the above equation were not true then some covariation between \( x(\tau) \) and \( y(t) \) could be used to estimate one from the other and so the optimal filter would have a corresponding coefficient to incorporate this. As was mentioned in class, in the case of random processes we need to derive the optimal filter \( h(t) \) which when convolved with the random process \( x(t) \) estimates best in a MS sense the process \( y(t) \). The optimal filter \( \hat{h}(t) \) must satisfy the orthogonality condition which in this case can be written as,

\[
\langle (y(t) - \hat{h}(t) * x(t)) x(t') \rangle = 0 \quad \forall \, t, t'
\]

\[
\Rightarrow \langle y(t) x(t') \rangle - \langle (\hat{h}(t) * x(t)) x(t') \rangle = 0
\]

The first term in the above equation is the cross-correlation of \( y(t) \) and \( x(t) \), since by definition, \( \langle y(t) x(t') \rangle = R_{y x}(t - t') \). The second term can be simplified as,

\[
\langle (\hat{h}(t) * x(t)) x(t') \rangle = \langle \int_{-\infty}^{\infty} \hat{h}(\tau) x(t - \tau) \, d\tau \rangle x(t')
\]

\[
= \int_{-\infty}^{\infty} \hat{h}(\tau) \langle x(t') x(t - \tau) \rangle \, d\tau
\]

(By interchanging the order of integration and \( \langle \cdot \rangle \))

We know that \( \langle x(t - \tau) x(t') \rangle = R_{xx}(t - t' - \tau) \). Substituting this in the equation above we have,

\[
\langle (\hat{h}(t) * x(t)) x(t') \rangle = R_{xx}(t - t') * \hat{h}(t - t')
\]

which gives us the equation that \( \hat{h} \) must satisfy to be optimal,

\[
R_{y x}(t) = \hat{h}(t) * R_{xx}(t)
\]

- Taking Fourier transforms of both sides we have,

\[
S_{y z}(\omega) = \hat{H}(\omega)S_{zz}(\omega)
\]

\[
\Rightarrow \hat{H}(\omega) = \frac{S_{y z}(\omega)}{S_{zz}(\omega)}
\]

where \( S_{y z}(\omega) \), \( S_{zz}(\omega) \) and \( \hat{H}(\omega) \) are the Fourier transforms of \( R_{y z}(t) \), \( R_{zz}(t) \) and \( \hat{h}(t) \) respectively. \( S_{zz}(\omega) \) is called the power spectrum of the process \( x(t) \) and it essentially measures how the power of \( x(t) \) is distributed across its various frequency components. \( S_{y z}(\omega) \) is called the cross-spectrum of the processes \( x(t) \) and \( y(t) \). A moment’s reflection will reveal the similarity between the optimal filter transfer function \( \hat{H}(\omega) \) and the optimal estimate \( \hat{a} \) for the scalar case.

- To mention an interesting aside here, the Fourier coefficients of a stationary random process are statistically independent. In other words, if we take the Fourier transform of a stationary process (which would also be a stationary random process, with frequency \( \omega \) as the indexing variable) then the random variables in the Fourier transform corresponding to two different values of \( \omega \) will be
statistically independent of each other. Thus, in estimating the Fourier transform of the process, knowledge at one particular frequency \( \omega_1 \) will not help/affect the estimation at some other frequency \( \omega_2 \). Thus, while designing the optimal filter if we choose to work in the frequency domain (which is simpler since convolution is replaced by multiplication) design of the filter coefficients at different frequency values is decoupled. This allows us to write down the values of \( H(\omega) \) at each frequency by just using the scalar result.

Since for the scalar case, \( E^* = \sigma_y^2(1 - \rho_{xy}^2) \), the error in estimating the Fourier transform of \( y(t) \) at \( \omega_1 \) from the Fourier transform of \( x(t) \) at \( \omega_1 \) will be

\[
E^*(\omega_1) = S_{yy}(\omega_1) \left(1 - \frac{|S_{yx}(\omega_1)|^2}{S_{xx}(\omega_1) S_{yy}(\omega_1)}\right)
\]

The parallels between two cases are very obvious to see. Now, since different frequency components are independent, the total error in estimation will be the integral over all values of \( \omega \). Thus,

\[
E^* = \int_{-\infty}^{\infty} S_{yy}(\omega)(1 - \frac{|S_{yx}(\omega)|^2}{S_{xx}(\omega) S_{yy}(\omega)}) d\omega
\]

5.3 Decoding Spike Trains by Linear Reconstruction

This problem will illustrate an approach which has been recently quite successfully used in addressing the complex problem of neural coding. Using linear reconstruction on a particular model (I & F) of a neuron we seek to reveal some properties of neural coding which have been shown to hold for some real neurons as well. We shall see that by only using a linear filter to estimate the stimulus from the spike train of a neuron we can estimate most of the variance of the input even though the transformation which relates the spike train to the input is highly non-linear.

5.3.1 The basics

In Fig.1, a comparison between the input and the reconstructed waveform is made. We have chosen an arbitrary value of the stimulus parameters (\( \sigma_i = 3 \), \( B_i = 50\text{Hz} \)) for the purposes of illustration. The input, reconstruction and the spike train have been scaled appropriately to fit on the same graph. In addition, the error (input - reconstruction) is offset by 1 for visual convenience.

- The stimulus waveform is shown in Fig.1.
- The generate membrane voltage and the output spike train is shown in Fig.2.
- From Fig.3, we can see that the reconstruction is fairly accurate for positive input values but it fails to track negative input swings. This can again be explained by the fact that for those times when the input goes negative there are very few spikes to give a good estimate of the signal. Verify this by observing the spike train record.
Figure 1: The stimulus waveform, with $B_s = 50Hz$ and $\sigma_s = 3$

Figure 2: The generate membrane voltage and the output spike train

- The optimal filter is here:

```matlab
J  = find(S_xx);
H_ = S_yx(J)/S_xx(J);
h  = fftshift(real(ifft(H_)));
```

These three lines do the following work:

$$H(\omega) = \frac{S_{yx}(\omega)}{S_{xx}(\omega)}$$
Figure 3: Comparison of Input, Reconstruction, Error in Reconstruction for Fixed $B_s$ and $\sigma_s$

$$h(t) = \text{IFFT}(H(\omega))$$