EECS225A Course Notes

Anmol Parande

Fall 2021 - Professor Jiantao Jiao

Disclaimer: These notes reflect EECS225A when I took the course (Fall 2021). They may not accurately reflect current course content, so use at your own risk. If you find any typos, errors, etc, please raise an issue on the GitHub repository.

Contents

1 Hilbert Space Theory

Complex random vectors form a Hilbert space with inner product $\langle X, Y\rangle = \mathbb{E}\left[XY^*\right]$. If we have a random complex vector, then we can use Hilbert Theory in a more efficient manner by looking at the matrix of inner products. For simplicity, we will call this the "inner product" of two complex vectors.

Definition 1 *Let the inner product between two random, complex vectors $\boldsymbol{Z_1}, \boldsymbol{Z_2}$ be defined as*

$$\langle \boldsymbol{Z_1}, \boldsymbol{Z_2}\rangle = \mathbb{E}\left[\boldsymbol{Z_1}\boldsymbol{Z_2}^*\right]$$

The ij-th entry of the matrix is simply the scalar inner product $\mathbb{E}\left[X_iY_j^*\right]$ where X_i and Y_j are the ith and jth entries of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. This means the matrix is equivalent to the cross correlation R_{XY} between the two vectors. We can also specify the auto-correlation $R_X = \langle \boldsymbol{X}, \boldsymbol{X}\rangle$ and auto-covariance $\Sigma_X = \langle \boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right], \boldsymbol{X} - \mathbb{E}\left[\boldsymbol{X}\right]\rangle$. One reason why we can think of this matrix as the inner product is because it also satisfies the properties of inner products. In particular, it is

1. Linear: $\langle \alpha_1\boldsymbol{V_1} + \alpha_2\boldsymbol{V_2}, \boldsymbol{u}\rangle = \alpha_1\langle \boldsymbol{V_1}, u\rangle + \alpha_2\langle \boldsymbol{V_2}, u\rangle$.
2. Reflexive: $\langle \boldsymbol{U}, \boldsymbol{V}\rangle = \langle \boldsymbol{V}, \boldsymbol{U}\rangle^*$.
3. Non-degeneracy: $\langle \boldsymbol{V}, \boldsymbol{V}\rangle = \boldsymbol{0} \Leftrightarrow \boldsymbol{V} = \boldsymbol{0}$.

Since we are thinking of the matrix as an inner product, we can also think of the norm as a matrix.

Definition 2 *The norm of a complex random vector is given by $\|\boldsymbol{Z}\|^2 = \langle \boldsymbol{Z}, \boldsymbol{Z}\rangle$.*

When thinking of inner products as matrices instead of scalars, we must rewrite the Hilbert Projection Theorem to use matrices instead.

Theorem 1 (Hilbert Projection Theorem) *The minimization problem $\min_{\hat{\boldsymbol{X}}(\boldsymbol{Y})} \|\hat{\boldsymbol{X}}(\boldsymbol{Y}) - \boldsymbol{X}\|^2$ has a unique solution which is a linear function of $\boldsymbol{Y}$. The error is orthogonal to the linear subspace of $\boldsymbol{Y}$ (i.e $\langle \boldsymbol{X} - \hat{\boldsymbol{X}}, \boldsymbol{Y}\rangle = \boldsymbol{0}$)*

When we do a minimization over a matrix, we are minimizing it in a PSD sense, so for any other linear function $\boldsymbol{X}'$,

$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 \preceq \|\boldsymbol{X} - \boldsymbol{X}'\|^2.$$

2 Linear Estimation

In Linear Estimation, we are trying to estimate a random variable $\boldsymbol{X}$ using an observation $\boldsymbol{Y}$ with a linear function of $\boldsymbol{Y}$. If $\boldsymbol{Y}$ is finite dimensional, then we can say $\hat{\boldsymbol{X}}(\boldsymbol{Y}) = W\boldsymbol{Y}$ where W is some matrix. Using theorem 1 and the orthogonality principle, we know that

$$\langle \boldsymbol{X} - W\boldsymbol{Y}, \boldsymbol{Y} \rangle = \mathbf{0} \Leftrightarrow R_{XY} = W\boldsymbol{R}_Y$$

This is known as the **Normal Equation**. If R_Y is invertible, then we can apply the inverse to find W. Otherwise, we can apply the pseudoinverse $R_Y^\dagger$ to find W, which may not be unique. If we want to measure the quality of the estimation, since $\boldsymbol{X} = \boldsymbol{X} + (\boldsymbol{X} - \hat{\boldsymbol{X}})$,

$$\|\boldsymbol{X}\|^2 = \|\hat{\boldsymbol{X}}\|^2 + \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 \implies$$
$$\|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 = \|\boldsymbol{X}\|^2 - \|\hat{\boldsymbol{X}}\|^2 = R_X - R_{XY}R_Y^{-1}R_{YX}$$

2.1 Affine Estimation

If we allow ourselves to consider an affine function for estimation $\hat{\boldsymbol{X}}(\boldsymbol{Y}) = W\boldsymbol{Y} + b$, then this is equivalent to instead finding an estimator

$$\hat{\boldsymbol{X}}(\boldsymbol{Y}') = W\boldsymbol{Y}' \qquad \text{where } \boldsymbol{Y}' = \begin{bmatrix} \boldsymbol{Y} \\ 1 \end{bmatrix}$$

This is equivalent to the following orthogonality conditions:

1. $\langle \boldsymbol{X} - \hat{\boldsymbol{X}}, \boldsymbol{Y} \rangle$
2. $\langle \boldsymbol{X} - \hat{\boldsymbol{X}}, 1 \rangle$

Solving gives us

$$\hat{\boldsymbol{X}}(\boldsymbol{Y}) = W(\boldsymbol{Y} - \boldsymbol{\mu}_Y) + \mu_x \qquad \text{where } W\Sigma_Y = \Sigma_{XY}.$$

Σ_Y and Σ_{XY} are the auto-covariance and cross-covariance respectively. Recall that if

$$\begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \boldsymbol{\mu}_X \\ \boldsymbol{\mu}_Y \end{bmatrix}, \begin{bmatrix} \Sigma_X & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_Y \end{bmatrix} \right)$$

then

$$\boldsymbol{X}|\boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{\mu}_X + \Sigma_{XY}\Sigma_Y^{-1}(\boldsymbol{Y} - \boldsymbol{\mu}_Y), \Sigma_X - \Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} \right)$$

Thus in the Joint Gaussian case, the mean of the conditional distribution is the best affine estimator of $\boldsymbol{X}$ using $\boldsymbol{Y}$, and the covariance is the estimation error. This has two interpretations.

1. Under the Gaussian assumption, the best nonlinear estimator $\mathbb{E}\left[\boldsymbol{X}|\boldsymbol{Y}\right]$ is affine
2. Gaussian random variables are the hardest predict because nonlinearity should improve our error, but it does not in the Gaussian case. This means if affine estimation works well, we shouldn't try and find better non-linear estimators.

2.2 Least Squares

The theory of linear estimation is very closely connected with the theory behind least squares in linear algebra. In least squares, we have a deterministic $\boldsymbol{x}$ and assume nothing else about it, meaning we are looking for an unbiased estimator. theorem 2 tells us how to find the best linear unbiased estimator in a linear setting.

Theorem 2 (Gauss Markov Theorem) *Suppose that $\boldsymbol{Y} = H\boldsymbol{x} + \boldsymbol{Z}$ and Z is zero-mean with $\langle \boldsymbol{Z}, \boldsymbol{Z} \rangle = \boldsymbol{I}$, H is full-column rank, then $\hat{\boldsymbol{x}}_b = (H^*H)^{-1}H^*\boldsymbol{Y}$ is the best linear unbiased estimator.*

3 Discrete Time Random Processes

Definition 3 *A Discrete-Time Random Process is a countably infinite collection of random variables on the same probability space $\{X_n : n \in \mathbb{Z}\}$.*

Discrete Time Random Processes have a mean function $\mu_n = \mathbb{E}\left[X_n\right]$ and an autocorrelation function $R_X(n_1, n_2) = \mathbb{E}\left[X_{n_1}X_{n_2}^*\right]$

3.1 Types of Random Processes

Definition 4 *A Wide-Sense Stationary Random Process is a disrete-time random process with constant mean, finite variance, and an autocorrelation function that can be re-written to only depend on $n_1 - n_2$.*

We call this wide-sense stationary because the mean and covariance do not change as the process evolves. In a strict-sense stationary process, the distribution of each random variable in the process would not change.

Definition 5 *A WSS process $Z \sim \mathcal{WN}(0, \sigma^2)$ is a white noise process with variance σ^2 if and only if $\mathbb{E}\left[Z_n\right] = 0$ and $\mathbb{E}\left[Z_nZ_m^*\right] = \sigma^2\delta[n, m]$.*

3.1.1 Hidden Markov Processes

Definition 6 *We say that random variables X, Y, Z form a Markov Triple X—Y—Z if and only if X and Z are conditionall independent on Y*

Mathematically, Markov triplets satisfy three properties.

1. $p(x, z|y) = p(x|y)p(z|y)$
2. $p(z|x, y) = p(z|y)$
3. $p(x|y, z) = p(x|y)$

Because of these rules, the joint distribution can be written as $p(x, y, z) = p(x)p(y|x)p(z|y)$.

Theorem 3 *Random variables X, Y, Z form a Markov triplet if and only if there exist ϕ_1, ϕ_2 such that $p(x, y, z) = \phi_1(x, y)\phi_2(y, z)$.*

To simplify notation, we can define $X_m^n = (X_m, X_{m+1}, \cdots, X_n)$ and $X^n = X_1^n$.

Definition 7 *A Markov Process is a Discrete Time Random Process $\{X_n\}_{n\geq 1}$ where X_n—X_{n-1}—X^{n-2} for all $n \geq 2$*

Because of the conditional independence property, we can write the joint distribution of all states in the Markov process as

$$p(x^n) = \prod_{t=1}^{n} p(x_t|x^{t-1}) = \prod_{t=1}^{n} p(x_t|x_{t-1}).$$

Definition 8 *If $\{X_n\}_{n\geq 1}$ is a Markov Process, then $\{Y_n\}_{n\geq 1}$ is a Hidden Markov Process if we can factorize the conditional probability density*

$$p(y^n, x^n) = \prod_{i=1}^{n} p(y_i|x_i)$$

We can think of Y as a noisy observation of an underlying Markov Process. The joint distribution of $\{X_n\}_{n\geq 1}$ and $\{Y_n\}_{n\geq 1}$ can be written as

$$p(x^n, y^n) = p(x^n)p(y^n|x^n) = \prod_{t=1}^{n} p(x_t|x_{t-1}) \prod_{i=1}^{n} p(y_i|x_i).$$

Hidden Markov Models can be represented by undirected graphical models. To create an undirected graphical model,

1. Create a node for each random variable.
2. Draw an edge between two nodes if a factor of the joint distribution contains both nodes.

Undirected graphical models of Hidden Markov Processes are useful because they let us derive additional Markov dependedepencies between groups of variables.

Theorem 4 *For 3 disjoint sets S_1, S_2, S_3 of notes in a graphical model, if any path from S_1 to S_3 passes through a node in S_2, then S_1—S_2—S_3.*

3.2 Spectral Density

Recall that the Discrete Time Fourier Transform is given by

$$X(e^{j\omega}) = \sum_{n=-\infty}^{\infty} x[n]e^{-j\omega n}.$$

The Inverse Discrete Time Fourier Transform is given by

$$x[n] = \frac{1}{2\pi}\int_{-\pi}^{\pi} X(e^{j\omega})e^{j\omega n}d\omega.$$

Since the DTFT is an infinite summation, it may or may not converge.

Definition 9 *A signal $x[n]$ belongs to the l^1 class of signals if the series converges absolutely. In other words,*

$$\sum_{k=-\infty}^{\infty} |x[k]| < \infty.$$

This class covers most real-world signals.

Theorem 5 *If $x[n]$ is a l^1 signal, then the DTFT $X(e^{j\omega})$ converges uniformly and is well-defined for every ω. $X(e^{j\omega})$ is also a continuous function.*

Definition 10 *A signal $x[n]$ belongs to the l^2 class of signals if it is square summable. In other words,*

$$\sum_{k=-\infty}^{\infty} |x[k]|^2 < \infty.$$

The l^2 class contains important functions such as sinc.

Theorem 6 *If $x[n]$ is a l^2 signal, then the DTFT $X(e^{j\omega})$ is defined almost everywhere and only converges in the mean-squared sense:*

$$\lim_{N\to\infty}\int_{-\pi}^{\pi}\left|\left(\sum_{k=-N}^{N}x[k]e^{-j\omega n}\right)-X(\omega)\right|^2 d\omega = 0$$

Tempered distributions like the Dirac Delta function are other functions which are important for computing the DTFT, and they arise from the theory of generalized functions.

Suppose we want to characterize the signal using its DTFT.

Definition 11 *The energy of a deterministic, discrete-time signal $x[n]$ is given by*

$$\sum_{n\in\mathbb{Z}}|x[n]|^2.$$

The autocorrelation of $x[n]$, given by $a[n] = x[n] * x^*[-n]$, is closely related to the energy of the signal since $a[0] = \sum_{n\in\mathbb{Z}}|x(n)|^2$.

Definition 12 *The Energy Spectral Density $x[n]$ with auto-correlation $a[n]$ is given by*

$$A(e^{j\omega}) = \sum_{n\in\mathbb{Z}}a[n]e^{-j\omega n}$$

We call the DTFT of the autocorrelation the energy spectral density because, by the Inverse DTFT,

$$a[0] = \frac{1}{2\pi}\int_{-\pi}^{\pi}A(e^{j\omega})d\omega.$$

Since summing over each frequency gives us the energy, we can think of $A(e^{j\omega})$ as storing the energy density of each spectral component of the signal. We can apply this same idea to wide-sense stationary stochastic processes.

Definition 13 *The Power Spectral Density of a Wide-Sense Stationary random process is given by*

$$S_X(e^{j\omega}) = \sum_{k\in\mathbb{Z}}R_X(k)e^{-j\omega k}.$$

Note that when considering stochastic signals, the metric changes from energy to power. This is because if X_n is Wide-Sense Stationary, then

$$\mathbb{E}\left[\sum_{n\in\mathbb{Z}}|X_n|^2\right] = \infty,$$

so energy doesn't even make sense. To build our notion of power, let $A_T(\omega)$ be a truncated DTFT of the auto-correlation of a wide-sense stationary process, then

$$\begin{aligned}
\lim_{T\to\infty}\frac{\mathbb{E}\left[A_T(e^{j\omega})\right]}{2T+1} &= \lim_{T\to\infty}\frac{1}{2T+1}\left(\sum_{n=-T}^{T}x[n]e^{-j\omega n}\right)\left(\sum_{m=-T}^{T}x^*[m]e^{j\omega m}\right)\\
&= \lim_{T\to\infty}\frac{1}{2T+1}\sum_{n,m\in[-T,T]}\mathbb{E}\left[x[n]x^*[m]\right]e^{-j\omega(n-m)}\\
&= \lim_{T\to\infty}\frac{1}{2T+1}\sum_{n,m\in[-T,T]}R_x(n-m)e^{-j\omega(n-m)}\\
&= \lim_{T\to\infty}\sum_{k=-2T}^{2T}R_X(k)e^{-j\omega k}\left(1-\frac{|k|}{2T+1}\right)\\
&= \sum_{k=-\infty}^{\infty}R_X(k)e^{-j\omega k}
\end{aligned}$$

The DTFT of the auto-correlation function naturally arises out of taking the energy spectral density and normalizing it by time (the truncated sequence is made of $2T+1$ points). In practice, this means to measure the PSD, we need to either use the distribution of the signal to compute R_X, or estimate the PSD by averaging multiple realizations of the signal.

The inverse DTFT formula tells us that we can represent a deterministic, discrete-time signal $x[n]$ as a sum of complex exponentials weighted by $\frac{X(e^{j\omega})d\omega}{2\pi}$. This representation has an analog for stochastic signals as well.

Theorem 7 (Cramer-Khinchin) *For a complex-valued WSS stochastic process X_n with power spectral density $S_X(\omega)$, there exists a unique right-continuous stochastic process $F(\omega), \omega \in (-\pi, \pi]$ with square-integrable, orthogonal increments such that*

$$X_n = \int_{-\pi}^{\pi} e^{j\omega n}dF(\omega)$$

where for any interval $[\omega_1, \omega_2], [\omega_3, \omega_4] \subset [-\pi, \pi]$,

$$\mathbb{E}\left[(F(\omega_2)-F(\omega_1))(F(\omega_4)-F(\omega_3))^*\right] = f((\omega_1, \omega_2] \cap (\omega_3, \omega_4])$$

where f is the structural measure of the stochastic process and has Radon-Nikodym derivative $\frac{S_X(e^{j\omega})}{2\pi}$.

Besides giving us a decomposition of a WSS random process, theorem 7 tells a few important facts.

1. $\omega_1 \neq \omega_2 \implies \langle dF(\omega_1), dF(\omega_2) \rangle = 0$ (i.e different frequencies are uncorrelated).
2. $\mathbb{E}\left[|dF(\omega)|^2\right] = \frac{S_X(e^{j\omega})d\omega}{2\pi}$

3.3 Z-Spectrum

Recall that the Z-transform converts a discrete-time signal into a complex representation. It is given by

$$X(z) = \sum_{n=-\infty}^{\infty} x[n]z^{-n}.$$

It is a special type of series called a **Laurent Series**.

Theorem 8 *A Laurent Series will converge absolutely on an open annulus*

$$A = \{z | r < |z| < R\}$$

for some r and R.

We can compute r and R using the signal $x[n]$.

$$r = \limsup_{n\to\infty} |x[n]|^{\frac{1}{n}}, \qquad \frac{1}{R} = \limsup_{n\to\infty} |x[-n]|^{\frac{1}{n}}.$$

In some cases, it can be useful to only compute the Z-transform of the right side of the signal.

Definition 14 *The unilateral Z-transform of a sequence $x[n]$ is given by*

$$\left[X(z)\right]_+ = \sum_{n=0}^{\infty} x[n]z^{-n}$$

If the Z-transform of the sequence is a rational function, then we can quickly compute what the unilateral Z-transform will be by leveraging its partial fraction decomposition.

Theorem 9 *Any arbitrary rational function $H(z)$ with region of convergence including the unit circle corresponds with the unilateral Z-transform*

$$\left[H(z)\right]_+ = r_0 + \sum_{i=1}^{m} \sum_{k=1}^{l_i} \frac{r_{ik}}{(z+\alpha_i)^k} + \sum_{i=m+1}^{n} \sum_{k=1}^{l_i} \frac{r_{ik}}{\beta_i^k}$$

where $|\alpha_i| < 1 < |\beta_i|$.

Definition 15 *For two jointly WSS processes X_n, Y_n, the z-cross spectrum is the Z-Transform of the correlation function $R_{YX}(k) = \mathbb{E}\left[Y_n X^*_{n-k}\right]$.*

$$S_{YX}(z) = \sum_{k\in\mathbb{Z}} R_{YX}(k) z^{-k}$$

Using this definition, we can see that

$$S_{XY}(z) = S^*_{YX}(z^{-*}).$$

We can also look at the Z-transform of the auto-correlation function of a WSS process X to obtain $S_X(z)$.

Definition 16 *For a rational function $S_X(z)$ with finite power $\left(\int_{-\pi}^{\pi} S_X(e^{j\omega})d\omega < \infty\right)$ and is strictly positive on the unit circle, the canonical spectral factorization decomposes $S_X(z)$ into a product of a $r_e > 0$ and the transfer function of a minimum phase system $L(z)$ with $L(\infty) = 1$*

$$S_X(z) = L(z) r_e L^*(z^{-*})$$

Because $L(z)$ is minimum phase and $L(\infty) = 1$, it must take the form

$$L(z) = 1 + \sum_{i=1}^{\infty} l[i] z^{-i}$$

since minimum phase systems are causal. Using definition 16, we can express $S_X(z)$ as the product of a right-sided and left-sided process.

$$S_X(z) = (\sqrt{r_e}L(z))(\sqrt{r_e}L^*(z^{-*})) = S^+_X(z)S^-_X(z)$$

Note that $S^-_X(e^{j\omega}) = \left(S^+_X(e^{j\omega})\right)^*$. Using the assumptions built into $definition$ 16, we can find a general form for $L(z)$ since we know $S_Y(z)$ takes the following form

$$S_Y(z) = r_e \frac{\prod_{i=1}^m (z-\alpha_i)(z^{-1}-\alpha^*_i)}{\prod_{i=1}^n (z-\beta_i)(z^{-1}-\beta^*_i)} \quad |\alpha_i| < 1, |\beta_i| < 1, r_e > 0.$$

If we let the $z - \alpha_i$ and $z - \beta_i$ terms be part of $L(z)$, then

$$L(z) = z^{n-m} \frac{\prod_{i=1}^m (z-\alpha_i)}{\prod_{i=1}^n (z-\beta_i)}.$$

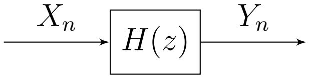

Figure 1: Filtering a Disrete Time Random Process with an LTI system with transfer function $H(z)$

4 Filtering

If we think of our signal as a discrete time random process, then like a normal deterministic signal, we can try filtering our random process. Filtering can either be accomplished with an LTI system or some other non-linear system just like with deterministic signals. If we use an LTI filter on a WSS process, then we can easily compute how the filter impacts the spectrum of the signal.

Theorem 10 *When $Y(n)$ is formed by passing a WSS process X_n through a stable LTI system with impulse response $h[n]$ and transfer function $H(z)$, then $S_Y(z) = H(z)S_X(z)H^*(z^{-*})$ and $S_{YX}(z) = H(z)S_X(z)$. If we have a third process Z_n that is jointly WSS with (Y_n, X_n), then $S_{ZY}(z) = S_{ZX}(z)H^*(z^{-*})$.*

This gives us an interesting interpretation of the spectral factorization $(definition\ 16)$ since it essentially passing a WSS process with auto-correlation $R_W(k) = r_e\delta[n]$ through a minimum-phase filter with transfer function $L(z)$.

4.1 Wiener Filter

Suppose we have a stochastic WSS process Y_n that is jointly WSS with X_n and that we want to find the best linear estimator of X_n using Y_n. The best linear estimator of X_n given the observations Y_n can be written as

$$\hat{X}_n = \sum_{m\in\mathbb{Z}} h(m)Y_{n-m} = h[n] * Y_n.$$

This is identical to passing Y_n through an LTI filter. If we restrict ourselves to using $\{Y_i\}_{i=-\infty}^{n}$ to estimate X_n, then the best linear estimator can be written as

$$\hat{X}_n = \sum_{m=0}^{\infty} h(m)Y_{n-m} = h[n] * Y_n.$$

It is identical to passing Y_n through a causal LTI filter. Since we are trying to find a best linear estimator, it would be nice if each of the random variables we are using for estimating were uncorrelated with each other. In other words, instead of using Y directly, we want to transform Y into a new process W where $R_W(k) = \delta[k]$.

This transformation is known as whitening. From the spectral factorization of Y, we know if we use the filter $G(z) = \frac{1}{S_Y^+(z)}$ then

$$S_W(z) = \frac{S_Y(z)}{S_Y^+(z)S_Y^{+*}(z^{-*})} = \frac{S_Y(z)}{S_Y^+(z)S_Y^-(z)} = 1.$$

Now we want to find the best linear estimator of X using our new process W by designing an LTI filter $Q(z)$.

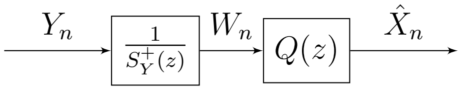

Figure 2: Finding the best linear estimator of X using W with a two-stage filter that first whitens the input.

4.1.1 Non-Causal Case

Starting with noncausal case, we can apply the orthogonality principle,

$$\mathbb{E}\left[(X_n - \hat{X}_n)W_{n-k}^*\right] = 0 \implies \mathbb{E}\left[X_n W_{n-k}^*\right] = \sum_{m\in\mathbb{Z}} q(m)\mathbb{E}\left[W_{n-m}W_{n-k}^*\right]$$

$$\therefore R_{XW}(k) = \sum_{m\in\mathbb{Z}} q(m)R_W(k-m) \implies S_{XW}(z) = Q(z)S_W(z)$$

$$\therefore Q(z) = \frac{S_{XW}(z)}{S_W(z)} = S_{XW}(z) = S_{XY}(z)(S_Y^+(z^{-*}))^{-*} = \frac{S_{XY}(z)}{S_Y^-(z)}$$

When we cascade these filters,

$$H(z) = Q(z)G(z) = \frac{S_{XY}(z)}{S_Y^-(z)}\frac{1}{S_Y^+(z)} = \frac{S_{XY}(z)}{S_Y(z)}.$$

Definition 17 *The best linear estimator of X_n using Y_n where (X_n, Y_n) is jointly WSS is given by the non-causal Wiener filter.*

$$H(z) = \frac{S_{XY}(z)}{S_Y(z)}$$

If we interpret definition 17 in the frequency domain, for a specific ω, we can understand $H(e^{j\omega})$ as an optimal linear estimator for $F_X(\omega)$ where $F_X(\omega)$ is the

the stochastic process given by the Cramer-Khinchin decomposition (theorem 7). More specifically, we can use the Cramer-Khinchin decomposition of Y_n.

$$\begin{aligned}\hat{X}_n &= \sum_{i\in\mathbb{Z}} h[i] \int_{-\pi}^{\pi} e^{j\omega(n-i)} dF_Y(\omega) \\ &= \int_{-\pi}^{\pi} \left(\sum_{i\in\mathbb{Z}} h[i] e^{-j\omega i} \right) e^{j\omega n} dF_Y(\omega) \\ &= \int_{-\pi}^{\pi} H(e^{j\omega}) e^{j\omega n} dF_Y(\omega)\end{aligned}$$

Since F_X and F_Y have jointly orthogonal increments, this tells us that $H(e^{j\omega})$ is just the optimal linear estimator of $dF_X(\omega)$ using $dF_Y(\omega)$. $dF_X(\omega)$ and $dF_Y(\omega)$ exist on a Hilbert space, meaning we are essentially projecting each frequency component of X_n onto the corresponding frequency component of Y_n.

4.1.2 Causal Case

First, note that in the causal case, whitening doesn't break causality because $\frac{1}{S_Y^+(z)}$ is causal. When we apply the orthogonality principle,

$$\mathbb{E}\left[(X_n - \hat{X}_n) W_{n-k}^* \right] = 0 \implies \mathbb{E}\left[X_n W_{n-k}^* \right] = \sum_{m=0}^{\infty} q(m) \mathbb{E}\left[W_{n-m} W_{n-k}^* \right]$$

$$\therefore R_{XW}(k) = \sum_{m=0}^{\infty} q[m] R_W(k-m) \qquad k \geq 0$$

We can't take the Z-transform of both sides because the equation is not necessarily true for $k < 0$. Instead, we can look at the function

$$f(k) = R_{XW}(k) - \sum_{m=0}^{\infty} R_W(k-m) q[m] = \begin{cases} 0 & k \geq 0, \\ ? & \text{else.} \end{cases}$$

Taking the unilateral Z-transform of both sides,

$$\begin{aligned}[F(z)]_+ &= [S_{XW}(z) - S_W(z)Q(z)]_+ = [S_{XW}(z)]_+ - Q(z) = 0 \\ Q(z) &= [S_{XW}(z)]_+ = \left[\frac{S_{XY}(z)}{S_Y^-(z)} \right]_+\end{aligned}$$

Thus the filter H which gives the causal best linear estimator of X using Y is

$$H(z) = Q(z)G(z) = \left[\frac{S_{XY}(z)}{S_Y^-(z)} \right]_+ \frac{1}{S_Y^+(z)}.$$

Definition 18 *The best linear estimator of X_n using $\{Y_i\}_{i=-\infty}^{n}$ is given by the causal Wiener filter.*

$$H(z) = Q(z)G(z) = \left[\frac{S_{XY}(z)}{S_Y^-(z)}\right]_+ \frac{1}{S_Y^+(z)}.$$

Intuitively, this should make sense because we are using the same W process as in the non-causal case, but only the ones which we are allowed to use, hence use the unilateral Z-transform of the non-causal Wiener filter, which amounts to truncated the noncausal filter to make it causal.

Theorem 11 *If $\hat{X}_{NC}(n)$ is the non-causal Wiener filter of X, then the causal wiener filter of X given Y is the same as the causal wiener filter of $\hat{X}_{NC}$ given Y, and if Y is white noise, then*

$$\hat{X}_C(n) = \sum_{i=0}^{\infty} h[i] Y_{n-i}$$

4.1.3 Vector Case

Suppose that instead of a Wide-Sense Stationary process, we an N length signal $\boldsymbol{X}$ which we want to estimate with another N length signal $\boldsymbol{Y}$. We can represent both $\boldsymbol{X}$ and $\boldsymbol{Y}$ as vectors in $\mathbb{C}^N$. If we are allowed to use all entries of $\boldsymbol{Y}$ to estimate $\boldsymbol{X}$, this is identical to linear estimation.

Definition 19 *The non-causal Wiener filter of a finite length N signal $\boldsymbol{Y}$ is given by*

$$K_s = R_{\boldsymbol{XY}} R_{\boldsymbol{Y}}^{-1}.$$

Note that this requires $R_{\boldsymbol{Y}} \succ 0$. Suppose that we wanted to design a causal filter for the vector case, so $\hat{X}_i$ only depends on $\{Y_j\}_{j=1}^{i}$. By the orthogonality principle,

$$\forall 1 \leq l \leq i,\ \mathbb{E}\left[X_i - \sum_{j=1}^{i} K_{f,ij} Y_j Y_l^*\right] = 0 \implies R_{\boldsymbol{XY}}(i,l) = \sum_{j=1}^{i} K_{f,ij} R_{\boldsymbol{Y}}(j,l)$$

In matrix form, this means

$$R_{\boldsymbol{XY}} - K_f R_{\boldsymbol{Y}} = U^+$$

where U^+ is strictly upper triangular.

Theorem 12 *If matrix $H \succ 0$, then there exists a unique lower-diagonal upper triangular factorization of $H = LDL^*$ where L is lower diagonal and invertible with unit diagonal entries and D is diagonal with positive entries.*

Applying the LDL decomposition, we see that

$$\begin{aligned} R_{\boldsymbol{XY}} - K_f LDL^* = U^+ \implies R_{\boldsymbol{XY}} L^{-*} D^{-1} - K_f L &= U^+ L^{-*} D^{-1} \\ \therefore [R_{\boldsymbol{XY}} L^{-*} D^{-1}]_L - K_f L &= 0 \end{aligned}$$

where $[\cdot]_L$ represent the lower triangular part of a matrix.

Definition 20 *The causal Wiener filter of a finite length N signal $\boldsymbol{Y}$ is given by*

$$K_f = [R_{\boldsymbol{XY}} L^{-*} D^{-1}]_L L^{-1}$$

4.2 Hidden Markov Model State Estimation

Suppose we have a Hidden Markov Process $\{Y_n\}_{n\geq 1}$. We can think of determining the state $\{X_n\}_{n\geq 1}$ as filtering $\{Y_n\}_{n\geq 1}$.

4.2.1 Causal Distribution Estimation

Suppose we want to know the distribution of X_t after we have observed Y^t.

$$\begin{aligned} p(x_t|y^t) &= \frac{p(x_t, y^t)}{p(y^t)} = \frac{p(x_t)p(y_t, y^{t-1}|x_t)}{\sum_x p(y_t, y^{t-1}|x_t = x)p(x_t = x)} \\ &= \frac{p(x_t)p(y_t|x_t)p(y^{t-1}|x_t)}{\sum_x p(y_t|x_t = x)p(y^{t-1}|x_t = x)p(x_t = x)} = \frac{p(y_t|x_t)p(y^{t-1})p(x_t|y^{t-1})}{\sum_x p(y_t|x_t = x)p(y^{t-1})p(x_t = x|y^{t-1})} \\ &= \frac{p(y_t|x_t)p(x_t|y^{t-1})}{\sum_x p(y_t|x_t = x)p(x_t = x|y^{t-1})} \end{aligned}$$

Now if we know $p(x_t|y^{t-1})$, then we are set.

$$\begin{aligned} p(x_t|y^{t-1}) &= \sum_x p(x_t, x_{t-1} = x|y^{t-1}) = \sum_x p(x_{t-1} = x|y^{t-1})p(x_t|x_{t-1} = x, y^{t-1}) \\ &= \sum_x p(x_{t-1} = x|y^t)p(x_t|x_{t-1} = x) \end{aligned}$$

Now we have a recursive algorithm for computing the distribution of x_t.

Algorithm 1: Forward Recursion

```
$\beta_1(x_1) = p(x_1)$;
for $t \geq 1$ do
    $\alpha_t(x_t) = p(x_t|y^t) = \frac{\beta_t(x_t)p(y_t|x_t)}{\sum_x \beta_t(x)p(y_t|x_t)}$ (Measurement Update);
    $\beta_{t+1}(x_{t+1}) = p(x_t|y^{t-1}) = \sum_x \alpha_t(x)p(x_{t+1}|x_t = x)$ (Time Update);
end
```

4.2.2 Non-Causal Distribution Estimation

Suppose we are allowed to non-causally filter our signal and we care about the distribution of X_t after we have observed Y^n. In other words, for $t \geq n$, we want to find $\gamma_t(x_t) = p(x_t|y^n)$. When $t = n$, $\gamma_n(x_n) = \alpha_n(x_n)$. If we continue expanding backwards, then

$$\begin{aligned} p(x_t|y^n) &= \sum_x p(x_t, x_{t+1} = x|y^n) = \sum_x p(x_{t+1} = x|y^n)p(x_t|x_{t+1} = x, y^t, y_{t+1}^n) \\ &= \sum_x p(x_{t+1} = x|y^n)p(x_t|x_{t+1}, y^t) = \sum_x p(x_{t+1} = x|y^n)\frac{p(x_t|y^t)p(x_{t+1} = x|x_t, y^t)}{p(x_{t+1} = x|y^t)} \\ &= \sum_x \gamma_{t+1}(x)\frac{\alpha_t(x_t)p(x_{t+1} = x|x_t)}{\beta_{t+1}(x)} \end{aligned}$$

This gives us a clear algorithm for non-causally computing the distribution of x_t.

Algorithm 2: Backward Recursion

```
Run Forward Recursion;
$\gamma_n(x_n) = \alpha_n(x_n)$;
for $t = n - 1$ to 1 do
    $\gamma_t(x_t) = \sum_x \gamma_{t+1}(x)\frac{\alpha_t(x_t)p(x_{t+1}=x|x_t)}{\beta_{t+1}(x)}$;
end
```

4.2.3 State Sequence Estimation

Suppose we want to find the most likely sequence of states given our observations. This means we should compute

$$\hat{X}^n = \underset{X^n}{\operatorname{argmax}}\, p(x^n|y^n)$$

$$\begin{aligned} p(x^t, y^t) &= p(x^{t-1}, y^{t-1})p(x_t, y_t|x^{t-1}, y^{t-1}) \\ &= p(x^{t-1}, y^{t-1})p(x_t|x^{t-1}, y^{t-1})p(y_t|x_t, x^{t-1}, y^{t-1}) \\ &= p(x^{t-1}, y^{t-1})p(x_t|x_{t-1})p(y_t|x_t) \end{aligned}$$

We see that there is a recursion in the joint distribution, so if we let $V_t(x_t) = \max_{x^{t-1}} p(x^t, y^t)$, then

$$\begin{aligned} V_t(x_t) &= \max_{x^{t-1}} p(x^t, y^t) = p(y_t|x_t) \max_{x^{t-1}} p(x^{t-1}, y^{t-1}) p(x_t|x_{t-1}) \\ &= p(y_t|x_t) \max_{x^{t-1}} \left[p(x_t|x_{t-1}) \max_{x^{t-2}} p(x^{t-1}, y^{t-1}) \right] \\ &= p(y_t|x_t) \max_{x^{t-1}} p(x_t|x_{t-1}) V_{t-1}(x_{t-1}) \end{aligned}$$

The base case is that $V_1(x_1) = p(x_1)p(y_1|x_1)$. V_t is useful because $\hat{x}_n = \operatorname{argmax}_{x_n} V_n(x_n)$. This is because we can first maximize over $\hat{X}^{n-1}$ and Y^n, so the only thing left to maximize is $\hat{x}_n$. Once we have $\hat{x}_t$, then we can comptue $\hat{x}_{t-1}$ by

$$\hat{x}_{t-1} = \operatorname*{argmax}_{x_{t-1}} p(\hat{x}_t|x_{t-1}) V_{t-1}(x_{t-1}).$$

Putting these equations gives us the Viterbi algorithm.

Algorithm 3: Viterbi Algorithm

```
$V_1(x_1) = p(x_1)p(y_1|x_1)$;
for $t = 2$ to $n$ do
    $V_t(x_t) = p(y_t|x_t) \max_{x_{t-1}} p(x_t|x_{t-1}) V_{t-1}(x_{t-1})$;
end
$\hat{x}_n = \operatorname{argmax}_{x_n} V_n(x_n)$;
for $t = n$ to $2$ do
    $\hat{x}^{t-1} = \operatorname{argmax}_{x_{t-1}} p(\hat{x}_t|x_{t-1}) V_{t-1}(x_{t-1})$;
end
```