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 $\langle X, Y \rangle = \mathbb{E}[XY^*]$. 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* $\boldsymbol{Z_1}, \boldsymbol{Z_2}$

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

The ij-th entry of the matrix is simply the scalar inner product $\mathbb{E}[\boldsymbol{X_i}\boldsymbol{Y_j}^*]$ where $\boldsymbol{X_i}$ and $\boldsymbol{Y_j}$ are the ith and jth entries of $\boldsymbol{X}$ and $\boldsymbol{Y}$ respectively. This means the matrix is equivalent to the cross correlation $\boldsymbol{R}_{XY}$ between the two vectors. We can also specify the auto-correlation $\boldsymbol{R}_X = \langle \boldsymbol{X}, \boldsymbol{X} \rangle$ and auto-covariance $\boldsymbol{\Sigma}_X = \langle \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}], \boldsymbol{X} - \mathbb{E}[\boldsymbol{X}] \rangle$. 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* $\|\boldsymbol{Z}\|^2 = \langle \boldsymbol{Z}, \boldsymbol{Z} \rangle$.

Since we are thinking of inner products as matrices instead of scalars, we can 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 llinear function of $\boldsymbol{Y}$. If $\boldsymbol{Y}$ is finite dimensional, then we can say $\hat{\boldsymbol{X}}(\boldsymbol{Y}) = \boldsymbol{W}\boldsymbol{Y}$ where $\boldsymbol{W}$ is some matrix. Using theorem 1 and the orthogonality principle, we know that

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

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

$$\begin{aligned} \|\boldsymbol{X}\|^2 = \|\hat{\boldsymbol{X}}\|^2 + \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 \implies \\ \|\boldsymbol{X} - \hat{\boldsymbol{X}}\|^2 = \|\boldsymbol{X}\|^2 - \|\hat{\boldsymbol{X}}\|^2 = \boldsymbol{R}_X - \boldsymbol{R}_{XY}\boldsymbol{R}_Y^{-1}\boldsymbol{R}_{YX} \end{aligned}$$

2.1 Affine Estimation

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

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

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

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

$\boldsymbol{\Sigma}_Y$ and $\boldsymbol{\Sigma}_{XY}$ are the auto-covariance and cross-covariance respectively. Recall that if

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

then

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

The mean is the best affine estimator of $\boldsymbol{X}$, and the covariance is the estimation error. This has two interpretations.

1. Under the Gaussian assumption, the best nonlinear estimator $\mathbb{E}[\boldsymbol{X}|\boldsymbol{Y}]$ 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 $\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{x}+\boldsymbol{Z}$ and Z is zero-mean with $\langle \boldsymbol{Z}, \boldsymbol{Z} \rangle = \boldsymbol{I}$, $\boldsymbol{H}$ is full-column rank, then $\hat{\boldsymbol{x}}_b = (\boldsymbol{H}^*\boldsymbol{H})^{-1}\boldsymbol{H}^*\boldsymbol{Y}$ 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 $\mu_n = \mathbb{E}[X_n]$ and an autocorrelation function $R_X(n_1, n_2) = \mathbb{E}\left[X_{n_1} X_{n_2}^*\right]$

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.

3.1 Spectral Analysis

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 5 *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 3 *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 6 *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 4 *If $x[n]$ is a l^2 signal, then the DTFT $X(ej\omega)$ 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 7 *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 8 *The Energy Spectral Density $x[n]$ with auto-correlation $a[n]$ is given by*

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

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

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

Since summing over each frequency gives us the energy spectral density, we can think of $A(\omega)$ 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 9 *The Power Spectral Density of a Wide-Sense Stationary random process is given by*

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

Note that when considering stochastic signals, the metric changes from energy to power. This is because stochastic processes are not generally l_2, and 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

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

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 $\frac{X(\omega)d\omega}{2\pi}$. This representation has an analog for stochastic signals as well.

Theorem 5 (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 $\frac{S_X(\omega)}{2\pi}$.

Besides giving us a decomposition of a WSS random process, theorem 5 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. $\mathbb{E}\left[|dF(\omega)|^2\right] = \frac{S_X(\omega)}{2\pi}$

3.2 LTI Filtering

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 6 *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 $x[n]$.

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

Definition 10 *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

$$S_{XY}(z) = S_{YZ}^*(z^{-*}).$$

Theorem 7 *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 $S_{ZY}(z) = S_{ZX}(z)H^*(z^{-*})$.*

3.2.1 Wiener Filtering

Suppose we have a stochastic WSS process Y_n that is jointly WSS with X_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. Starting with the orthogonality principle,

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

$$\therefore R_{XY}(k) = \sum_{m\in\mathbb{Z}} h(m) R_Y(k-m) \implies S_{XY}(e^{j\omega}) = H(e^{j\omega})S_Y(e^{j\omega})$$

$$\therefore H(e^{j\omega}) = \frac{S_{XY}(e^{j\omega})}{S_Y(e^{j\omega})}$$

Definition 11 *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.*

$$H(e^{j\omega}) = \frac{S_{XY}(e^{j\omega})}{S_Y(e^{j\omega})}$$

For a specific ω, we can understand $H(e^{j\omega})$ an optimal linear estimator for $F_X(\omega)$ where $F_X(\omega)$ is the the stochastic process given by the Cramer-Khinchin decomposition (theorem 5). 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.