Parametric Signal Modeling and Linear Prediction Theory
1. Discrete-time Stochastic Processes

Electrical & Computer Engineering
University of Maryland, College Park

Acknowledgment: ENEE630 slides were based on class notes developed by Profs. K.J. Ray Liu and Min Wu. The LaTeX slides were made by Prof. Min Wu and Mr. Wei-Hong Chuang.

Contact: *minwu@umd.edu*. Updated: October 27, 2011.

Outline of Part-2

1. Discrete-time Stochastic Processes
2. Discrete Wiener Filtering
3. Linear Prediction

Outline of Section 1

- Basic Properties and Characterization
  - 1st and 2nd moment function; ergodicity
  - correlation matrix; power-spectrum density
- The Rational Transfer Function Model
  - ARMA, AR, MA processes
  - Wold Decomposition Theorem
  - ARMA, AR, and MA models and properties
  - asymptotic stationarity of AR process

Readings for §1.1: Haykin 4th Ed. 1.1-1.3, 1.12, 1.14;
see also Hayes 3.3, 3.4, and background reviews 2.2, 2.3, 3.2

Stochastic Processes

- To describe the time evolution of a statistical phenomenon according to probabilistic laws.

  Example random processes: speech signals, image, noise, temperature and other spatial/temporal measurements, etc.

- Discrete-time Stochastic Process $\{u[n]\}$
  - Focus on the stochastic process that is defined / observed at discrete and uniformly spaced instants of time
  - View it as an ordered sequence of random variables that are related in some statistical way:
    $\{\ldots u[n-M], \ldots, u[n], u[n+1], \ldots\}$
  - A random process is not just a single function of time; it may have an **infinite** number of different realizations

Parametric Signal Modeling

- A general way to completely characterize a random process is by joint probability density functions for all possible subsets of the r.v. in it:

$$\text{Probability of } \{u[n_1], u[n_2], \ldots, u[n_k]\}$$

- **Question:** How to use only a few parameters to describe a process?

  Determine a model and then the model parameters

  $\Rightarrow$ This part of the course studies the signal modeling (including models, applicable conditions, how to determine the parameters, etc)

(1) Partial Characterization by 1st and 2nd moments

It is often difficult to determine and efficiently describe the joint p.d.f. for a general random process.

As a compromise, we consider partial characterization of the process by specifying its 1st and 2nd moments.

Consider a stochastic time series $\{u[n]\}$, where $u[n], u[n-1], \ldots$ may be complex valued. We define the following functions:

- **mean-value function**: $m[n] = \mathbb{E}\left[u[n]\right]$, $n \in \mathbb{Z}$
- **autocorrelation function**: $r(n, n-k) = \mathbb{E}\left[u[n]u^*[n-k]\right]$
- **autocovariance function**:
  $c(n, n-k) = \mathbb{E}\left[(u[n] - m[n])(u[n-k] - m[n-k])^*\right]$

Without loss of generality, we often consider zero-men random process $\mathbb{E}\left[u[n]\right] = 0 \ \forall n$, since we can always subtract the mean in preprocessing.

Now the autocorrelation and autocovariance functions become identical.

Wide-Sense Stationary (w.s.s.)

Wide-Sense Stationarity

If $\forall n$, $m[n] = m$ and $r(n, n-k) = r(k)$ (or $c(n, n-k) = c(k)$), then the sequence $u[n]$ is said to be wide-sense stationary (w.s.s.), or also called stationary to the second order.

- The strict stationarity requires the entire statistical property (characterized by joint probability density or mass function) to be invariant to time shifts.
- The partial characterization using 1st and 2nd moments offers two important advantages:
  1. reflect practical measurements;
  2. well suited for linear operations of random processes

(2) Ensemble Average vs. Time Average

- Statistical expectation $\mathbb{E}(\cdot)$ as an ensemble average: take average across (different realizations of) the process
- Time-average: take average along the process.

  This is what we can rather easily measure from one realization of the random process.

Question: Are these two average the same?

Answer: No in general. (Examples/discussions from ENEE620.)

Consider two special cases of correlations between signal samples:

1. $u[n], u[n-1], \cdots$ i.i.d.
2. $u[n] = u[n-1] = \cdots$ (i.e. all samples are exact copies)

Mean Ergodicity

For a w.s.s. process, we may use the time average

$$\hat{m}(N) = \frac{1}{N} \sum_{n=0}^{N-1} u[n]$$

to estimate the mean m.

- $\hat{m}(N)$ is an unbiased estimator of the mean of the process.

$\because \mathbb{E}[\hat{m}(N)] = m \quad \forall N$.

- **Question:** How much does $\hat{m}(N)$ from one observation deviate from the true mean?

Mean Ergodic

A w.s.s. process $\{u[n]\}$ is mean ergodic in the mean square error sense if $\lim_{N\to\infty} \mathbb{E}\left[|m - \hat{m}(N)|^2\right] = 0$

Mean Ergodicity

A w.s.s. process $\{u[n]\}$ is mean ergodic in the mean square error sense if $\lim_{N\to\infty} \mathbb{E}\left[|m - \hat{m}(N)|^2\right] = 0$

Question: under what condition will this be satisfied?

(Details)

$\Rightarrow$ (nece.& suff.) $\lim_{N\to\infty} \frac{1}{N} \sum_{\ell=-N+1}^{N-1} (1 - \frac{|\ell|}{N}) c(\ell) = 0$

Mean ergodicity suggests that $c(\ell)$ is asymptotically decaying s.t. $\{u[n]\}$ is asymptotically uncorrelated.

Correlation Ergodicity

Similarly, let the autocorrelation estimator be

$$\hat{r}(k, N) = \frac{1}{N} \sum_{n=0}^{N-1} u[n]u^*[n-k]$$

The w.s.s. process $\{u[n]\}$ is said to be correlation ergodic in the MSE sense if the mean squared difference between $r(k)$ and $\hat{r}(k, N)$ approaches zero as $N \to \infty$.

(3) Correlation Matrix

Given an observation vector $\underline{u}[n]$ of a w.s.s. process, the correlation matrix $\mathbf{R}$ is defined as $\mathbf{R} \triangleq \mathbb{E}\left[\underline{u}[n]\underline{u}^H[n]\right]$

where H denotes Hermitian transposition (i.e., conjugate transpose).

$$\underline{u}[n] \triangleq \begin{bmatrix} u[n] \\ u[n-1] \\ \vdots \\ u[n-M+1] \end{bmatrix},$$

Each entry in $\mathbf{R}$ is

$[\mathbf{R}]_{i,j} = \mathbb{E}\left[u[n-i]u^*[n-j]\right] = r(j-i)$
$(0 \leq i, j \leq M-1)$

$$\text{Thus } \mathbf{R} = \begin{bmatrix} r(0) & r(1) & \cdots & \cdots & r(M-1) \\ r(-1) & r(0) & r(1) & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ r(-M+2) & \cdots & \cdots & r(0) & r(1) \\ r(-M+1) & \cdots & \cdots & \cdots & r(0) \end{bmatrix}$$

Properties of $\mathbf{R}$

1. $\mathbf{R}$ is Hermitian, i.e., $\mathbf{R}^H = \mathbf{R}$

   Proof (Details)

2. $\mathbf{R}$ is Toeplitz.

   A matrix is said to be Toeplitz if all elements in the main diagonal are identical, and the elements in any other diagonal parallel to the main diagonal are identical.

   $\mathbf{R}$ Toeplitz $\Leftrightarrow$ the w.s.s. property.

Properties of **R**

3. **R** is non-negative definite , i.e., $\underline{x}^H \mathbf{R} \underline{x} \geq 0, \ \forall \underline{x}$

Proof (Details)

- eigenvalues of a Hermitian matrix are real.

(similar relation in FT: real in one domain $\sim$ conjugate symmetric in the other)

- eigenvalues of a non-negative definite matrix are non-negative.

Proof (Details)

Properties of **R**

4. $\underline{u}^B[n] \triangleq \begin{bmatrix} u[n-M+1] \\ \vdots \\ u[n-1] \\ u[n] \end{bmatrix}$, i.e., reversely ordering $\underline{u}[n]$,

then the corresponding correlation matrix becomes

$$\mathbb{E}\left[\underline{u}^B[n](\underline{u}^B[n])^H\right] = \begin{bmatrix} r(0) & r(-1) & \cdots & r(-M+1) \\ r(1) & r(0) & & \vdots \\ \vdots & & \ddots & \vdots \\ r(M-1) & \cdots & \cdots & r(0) \end{bmatrix} = \mathbf{R}^T$$

Properties of **R**

5. Recursive relations: correlation matrix for $(M+1) \times 1$ $\underline{u}[n]$:
(Details)

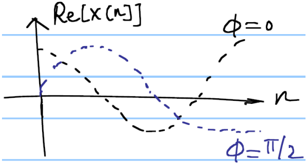
$$R_{M+1} = \begin{bmatrix} r(0) & r(1) & \cdots & r(M) \\ r^*(1) & r(0) & \cdots & r(M-1) \\ r^*(2) & r^*(1) & \ddots & \vdots \\ \vdots & \vdots & & \\ r^*(M) & r^*(M-1) & \cdots & r(0) \end{bmatrix} \quad \text{and} \quad \underline{u}_{M+1}[n] = \begin{bmatrix} \underline{u}_M[n] \\ \hline u(n-M) \end{bmatrix} = \begin{bmatrix} u[n] \\ \hline \underline{u}_M[n-1] \end{bmatrix}$$

$$= \begin{bmatrix} r(0) & \underline{r}^H \\ \underline{r} & R_M \end{bmatrix} = \begin{bmatrix} R_M & (\underline{r}^B)^* \\ (\underline{r}^B)^T & r(0) \end{bmatrix}$$

where $\underline{r} = \begin{bmatrix} r^*(1) \\ \vdots \\ r^*(M) \end{bmatrix}$, $\underline{r}^B = \begin{bmatrix} r^*(M) \\ \vdots \\ r^*(1) \end{bmatrix}$

(4) Example-1: Complex Sinusoidal Signal

$x[n] = A \exp\left[j(2\pi f_0 n + \phi)\right]$ where A and f_0 are real constant, $\phi \sim$ uniform distribution over $[0, 2\pi)$ (i.e., random phase)



$\mathbb{E}\left[x[n]\right] = ?$

$\mathbb{E}\left[x[n]x^*[n-k]\right] = ?$

Is $x[n]$ is w.s.s.?

Example-2: Complex Sinusoidal Signal with Noise

Let $y[n] = x[n] + w[n]$ where $w[n]$ is white Gaussian noise uncorrelated to $x[n]$, $w[n] \sim N(0, \sigma^2)$

Note: for white noise, $\mathbb{E}\left[w[n]w^*[n-k]\right] = \begin{cases} \sigma^2 & k = 0 \\ 0 & o.w. \end{cases}$

$r_y(k) = \mathbb{E}\left[y[n]y^*[n-k]\right] = ?$

$\mathbf{R}_y = ?$

Rank of Correlation Matrices $\mathbf{R}_x$, $\mathbf{R}_w$, $\mathbf{R}_y = ?$

(5) Power Spectral Density (a.k.a. Power Spectrum)

Power spectral density (p.s.d.) of a w.s.s. process $\{x[n]\}$

$$P_X(\omega) \triangleq \text{DTFT}[r_x(k)] = \sum_{k=-\infty}^{\infty} r_x(k) e^{-j\omega k}$$

$$r_x(k) \triangleq \text{DTFT}^{-1}[PX(\omega)] = \frac{1}{2\pi} \int_{-\pi}^{\pi} P_X(\omega) e^{j\omega k} d\omega$$

The p.s.d. provides frequency domain description of the 2nd-order moment of the process (may also be defined as a function of f: $\omega = 2\pi f$)

The power spectrum in terms of ZT:

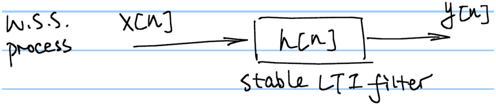
$$P_X(z) = \text{ZT}[r_x(k)] = \sum_{k=-\infty}^{\infty} r_x(k) z^{-k}$$

Physical meaning of p.s.d.: describes how the signal power of a random process is distributed as a function of frequency.

Properties of Power Spectral Density

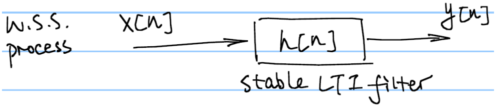
- $r_x(k)$ is conjugate symmetric: $r_x(k) = r_x^*(-k)$
  $\Leftrightarrow P_X(\omega)$ is real valued: $P_X(\omega) = P_X^*(\omega)$; $P_X(z) = P_X^*(1/z^*)$
- For real-valued random process: $r_x(k)$ is real-valued and even symmetric
  $\Rightarrow P_X(\omega)$ is real and even symmetric, i.e.,
  $P_X(\omega) = P_X(-\omega)$; $P_X(z) = P_X^*(z^*)$
- For w.s.s. process, $P_X(\omega) \geq 0$ (nonnegative)
- The power of a zero-mean w.s.s. random process is proportional to the area under the p.s.d. curve over one period 2π,
  i.e., $\mathbb{E}\left[|x[n]|^2\right] = r_x(0) = \frac{1}{2\pi}\int_0^{2\pi} P_X(\omega) d\omega$
  Proof: note $r_x(0) =$ IDTFT of $P_X(\omega)$ at $k = 0$

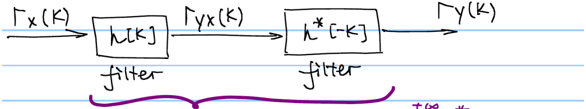
(6) Filtering a Random Process



(Details)

Filtering a Random Process

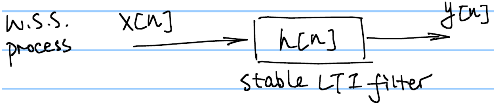




$$r_h[k] = h[k] * h^*[-k] = \sum_{l=-\infty}^{+\infty} h^*[l]\, h[k+l]$$

deterministic autocorrelation of filter's impulse response

Filtering a Random Process



In terms of ZT:

$$P_Y(z) = P_X(z)H(z)H^*(1/z^*)$$

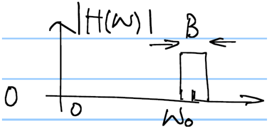
$$\Rightarrow P_Y(\omega) = P_X(\omega)H(\omega)H^*(\omega) = P_X(\omega)|H(\omega)|^2$$

When $h[n]$ is real, $H^*(z^*) = H(z)$

$$\Rightarrow P_Y(z) = P_X(z)H(z)H(1/z)$$

Interpretation of p.s.d.

If we choose $H(z)$ to be an ideal bandpass filter with very narrow bandwidth around any ω_0, and measure the output power:



$$\mathbb{E}\left[|y[n]|^2\right] = r_y(0) = \frac{1}{2\pi}\int_{-\pi}^{+\pi} P_Y(\omega)d\omega$$

$$= \frac{1}{2\pi}\int_{-\pi}^{+\pi} P_X(\omega)|H(\omega)|^2 d\omega = \frac{1}{2\pi}\int_{\omega_0-B/2}^{\omega_0+B/2} P_X(\omega)\cdot 1 \cdot d\omega$$

$$\doteq \frac{1}{2\pi} P_X(\omega_0)\cdot B \geq 0$$

$$\therefore P_X(\omega_0) \doteq \mathbb{E}\left[|y[n]|^2\right]\cdot \frac{2\pi}{B}, \text{ and } P_X(\omega) \geq 0 \ \ \forall \omega$$

i.e., p.s.d. is non-negative, and can be measured via power of $\{y[n]\}$.

⋇ $P_X(\omega)$ can be viewed as a density function describing how the power in $x[n]$ varies with frequency. The above BPF operation also provides a way to measure it by BPF.

Summary: Review of Discrete-Time Random Process

1. An "ensemble" of sequences, where each outcome of the sample space corresponds to a discrete-time sequence
2. A general and complete way to characterize a random process: through joint p.d.f.
3. w.s.s process: can be characterized by 1st and 2nd moments (mean, autocorrelation)
   - These moments are ensemble averages; $\mathbb{E}[x[n]]$, $r(k) = \mathbb{E}[x[n]x^*[n-k]]$
   - Time average is easier to estimate (from just 1 observed sequence)
   - **Mean ergodicity** and **autocorrelation ergodicity**: correlation function should be asymptotically decay, i.e., uncorrelated between samples that are far apart.
     $\Rightarrow$ the time average over large number of samples converges to the ensemble average in mean-square sense.

Characterization of w.s.s. Process through Correlation Matrix and p.s.d.

1. Define a vector on signal samples (note the indexing order): $\underline{u}[n] = [u(n), u(n-1), ..., u(n-M+1)]^T$
2. Take expectation on the outer product:
$$\mathbf{R} \triangleq \mathbb{E}\left[\underline{u}[n]\underline{u}^H[n]\right] = \begin{bmatrix} r(0) & r(1) & \cdots & \cdots & r(M-1) \\ r(-1) & r(0) & r(1) & \cdots & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ r(-M+1) & \cdots & \cdots & \cdots & r(0) \end{bmatrix}$$
3. Correlation function of w.s.s. process is a one-variable deterministic sequence $\Rightarrow$ take DTFT($r[k]$) to get p.s.d.
We can take DTFT on one sequence from the sample space of random process; different outcomes of the process will give different DTFT results; p.s.d. describes the statistical power distribution of the random process in spectrum domain.

Properties of Correlation Matrix and p.s.d.

4. Properties of correlation matrix:
   - Toeplitz (by w.s.s.)
   - Hermitian (by conjugate symmetry of $r[k]$);
   - non-negative definite

   Note: if we reversely order the sample vector, the corresponding correlation matrix will be transposed. This is the convention used in Hayes book (i.e. the sample is ordered from $n - M + 1$ to n), while Haykin's book uses ordering of n, $n - 1$, ... to $n - M + 1$.

5. Properties of p.s.d.:
   - real-valued (by conjugate symmetry of correlation function);
   - non-negative (by non-negative definiteness of **R** matrix)

Filtering a Random Process

1. Each specific realization of the random process is just a discrete-time signal that can be filtered in the way we've studied in undergrad DSP.
2. The ensemble of the filtering output is a random process. What can we say about the properties of this random process given the input process and the filter?
3. The results will help us further study such an important class of random processes that are generated by filtering a noise process by discrete-time linear filter with rational transfer function. Many discrete-time random processes encountered in practice can be well approximated by such a rational transfer function model: ARMA, AR, MA (see §II.1.2)