Massachusetts Institute of Technology
6.435 Theory of Learning and System Identification
(Spring 2007)

Prof. Dahleh, Prof. Mitter — Lecture 13
Scribed by Ilan Lobel — Th 3/22

In this lecture, we begin to study the problem of learning when the samples available are dependent. In particular, we focus on the case of linear regression.

1 Quasi-Stationary Signals

Definition 1. A discrete-time signal $\{X_t\}$ is quasi-stationary if $\mathbf{E}[X_t]$ is bounded for every $t\mathbb{N}$ and the following limit is well-defined and finite for every $\tau \in \mathbb{N}$:

$$R_X(\tau) = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathbf{E}[X_t X_{t+\tau}].$$

The set of quasi-stationary signals includes both stationary signals as well as several deterministic ones such as step functions and sinusoids.

Definition 2. A quasi-stationary signal $\{X_t\}$ has spectral density

$$\Phi_X(\omega) = \sum_{\tau=-\infty}^{\infty} R_X(\tau) e^{-i\omega\tau}.$$

To simplify notation we also define the following piece of notation $\overline{\mathbf{E}}$. For any signal $\{X_t\}$ where the limit is well-defined and finite,

$$\overline{\mathbf{E}}[X_t] = \lim_{N\to\infty} \frac{1}{N} \sum_{t=1}^{N} \mathbf{E}[X_t]. \tag{1}$$

2 Linear Filtering

A linear filter is characterized by a transfer function H, which is a function of the frequency ω of the input, as represented in Figure 1.

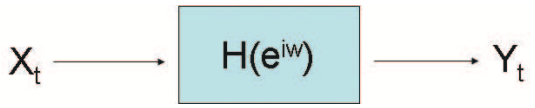

Figure 1: A Linear Filter

The spectral density of the output is given by

$$\Phi_Y(\omega) = |H(e^{i\omega})|^2 \Phi_X(\omega).$$

A white noise signal $\{X_t\}$ is characterized by

$$\Phi_X(\omega) = 1, \text{ for all frequencies } \omega.$$

It is a quasi-stationary signal with

$$R_X(\tau) = \delta(\tau), \text{ for every } \tau.$$

For any input $\{X_t\}$, the crosscorrelation is given by

$$\Phi_{XY}(\omega) = H(e^{i\omega})\Phi_X(\omega) \text{ and}$$

$$R_{XY}(\tau) = h * R_X(\tau),$$

where h represents the inverse Fourier Transform of the transfer function H and $*$ represents a convolution.

3 Problem Set-Up

Let Z^ℓ represent the observation:

$$Z^\ell = \{(x_1, y_1), ..., (x_\ell, y_\ell)\}.$$

Assume that the system is linear with parameters $n_a^\star$ and $n_b^\star$ such that

$$y_t = -\sum_{i=1}^{n_a^\star} a_i y_{t-i} + \sum_{i=0}^{n_b^\star} b_i x_{t-i} + e_t,$$

where $\{e_t\}$ is white gaussian noise with variance λ^2. Assume as well that X_t and Y_t are jointly quasi-stationary. Then, the following quantities are well-defined for every τ: $\overline{\mathbf{E}}[X_t X_{t+\tau}], \overline{\mathbf{E}}[Y_t Y_{t+\tau}]$ and $\overline{\mathbf{E}}[Y_t X_{t+\tau}]$, according to Eq. (1).

We also have a class of models to search over. Let n_a and n_b be a pair of numbers (not necessarily equal to $n_a^\star$ and $n_b^\star$). The class of models that we search over is parameterized by

$$\alpha = (a_1, a_2, ..., a_{n_a}, b_0, b_1, ..., b_{n_b})$$

and the models are of the form

$$y_t = -\sum_{i=1}^{n_a} a_i y_{t-i} + \sum_{i=0}^{n_b} b_i x_{t-i} + e_t,$$

where $\{e_t\}$ is zero-mean white noise with variance λ^2. We don't allow for the search of α over the entire space $\mathbb{R}^{n_a+n_b+1}$, but we require that $\alpha \in \Lambda$, where Λ is some compact subset of $\mathbb{R}^{n_a+n_b+1}$.

Finally, we also need to specify a cost function to determine how well a model fits the real underlying system. If we let for each $\alpha \in \Lambda$,

$$\hat{y}_t = \mathbf{E}[y_t | y_1, ..., y_{t-1}, x_1, ..., x_t, \alpha] = -\sum_{i=1}^{n_a} a_i y_{t-i} + \sum_{i=0}^{n_b} b_i x_{t-i},$$

then the latter represents the best prediction of y_t using all known information up to time t, given the model specified by α. The loss we associate with this prediction is

$$|y_t - \hat{y}_t|^2.$$

This choice places us in the Minimum Prediction Error (MPE) paradigm, where the empirical risk becomes

$$R_{emp}^\ell(\alpha) = \frac{1}{\ell}\sum_{t=1}^{\ell} |y_t - \hat{y}_t|^2.$$

The selected model will be

$$\alpha_\ell = \operatorname*{argmin}_{\alpha\in\Lambda} R_{emp}^\ell(\alpha).$$

Meanwhile, the true risk of a model parameterized by α is

$$R(\alpha) = \overline{\mathbf{E}}\left[|y_t - \hat{y}_t|^2\right].$$

4 Convergence Results

Claim. $R^{\ell}_{emp}(\alpha) \to R(\alpha)$ *uniformly on* Λ *almost surely.*

Proof. Pointwise convergence follows immediately. This is a proof that the convergence is uniform. Let $\hat{y}_t = \phi(t)'\alpha$, where

$$\phi(t) = (-y(t-1), ..., -y(t-n_a), x(t), ..., x(t-n_b)).$$

Then,

$$|y_t - \hat{y}_t|^2 = y_t^2 - 2y_t\phi(t)'\alpha + \alpha'\phi(t)\phi(t)'\alpha.$$

By quasi-stationarity,

$$\lim_{t\to\infty} |y_t - \hat{y}_t|^2 = \overline{\mathbf{E}}[y_t^2] - 2\overline{\mathbf{E}}[y_t\phi(t)']\alpha + \alpha'\overline{\mathbf{E}}[\phi(t)\phi(t)']\alpha.$$

Because all these terms are convergent, these functions are quadratic in α and α lies in a compact set, this family of functions is equicontinuous. An equicontinuous family of functions on a compact set that converges pointwise must also converge uniformly. □

Corollary. $\alpha_\ell = \text{argmin}_{\alpha\in\Lambda} R^{\ell}_{emp}(\alpha) \to \text{argmin}_{\alpha\in\Lambda} R(\alpha)$.

Claim. *The optimal* α_ℓ *satisfies*

$$\left[\frac{1}{\ell}\sum_{t=1}^{\ell}\phi(t)\phi(t)'\right]\alpha_\ell = \sum_{t=1}^{\ell}\phi(t)y_t.$$

This results follows immediately from the fact that determining α_ℓ is nothing more than solving a least squares problem. In particular, if the matrix is invertible,

$$\alpha_\ell = \left[\frac{1}{\ell}\sum_{t=1}^{\ell}\phi(t)\phi(t)'\right]^{-1}\sum_{t=1}^{\ell}\phi(t)y_t.$$

5 Persistence of Excitation

Definition 3. We say that a quasi-stationary $\{X_t\}$ is p.e. of order n if

$$\overline{R}_X = \begin{pmatrix} R_X(0) & ... & R_X(n-1) \\ ... & ... & ... \\ R_X(n-1) & ... & R_X(0) \end{pmatrix} \text{ is invertible.}$$

Note that a step input is p.e. of order 1 and a sinusoid is p.e. of order 2.

Claim. *If* $a_i = 0$ *for all* i*, then* $\overline{\mathbf{E}}[\phi(t)\phi(t)']$ *is invertible if* $\{X_t\}$ *is p.e. of order* $n_b + 1$.

In the case of $a_i = 0$ for all i, the matrices are essentially identical and the result follows trivially.

Claim. $\overline{\mathbf{E}}[\phi(t)\phi(t)']$ *is invertible if* $\{X_t\}$ *is p.e. of order* $n_a + n_b + 1$.