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

Prof. Dahleh, Prof. Mitter
Scribed by Sleiman Itani

Lecture 14
T 4/3

1 Review

In most of this lecture (and the previous one), we operate under the assumption that we have a set of data points (x_i, y_i),$i \in \{1, 2, ...\ell\}$, that were produced by the following system:

$$y_t = \sum_{i=1}^{n_a^\star} -a_i^\star y_{t-i} + \sum_{j=0}^{n_b^\star} b_j^\star x_{t-j} + w_t \tag{1}$$

Where $y_t = 0$ and $x_t = 0$ if $t \leq 0$, and w_t is i.i.d with $\text{var}(w_t) = \lambda^2$. We hypothesize a model that has the form:

$$y_t = \sum_{i=1}^{n_a} -a_i y_{t-i} + \sum_{j=0}^{n_b} b_j x_{t-j}$$

with specific n_a and n_b, and try to minimize the prediction error (empirical risk), defined as:

$$R_{emp}^{\ell}(a_1, a_2, ..., a_{n_a}, b_0, b_1, ..., b_{n_b}) = \frac{1}{\ell} \sum_{t=1}^{\ell} |y_t - \hat{y}_t|^2$$

A useful alternative way of writing the model is as follows:

$$y_t = \phi_t' \alpha$$

$$\phi_t = [-y_{t-1}, -y_{t-2}, ..., -y_{t-n_a}, x_t, x_{t-1}, ..., x_{t-n_b}]'$$

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

$$R_{emp}^{\ell}(\alpha) = \frac{1}{\ell} \sum_{t=1}^{\ell} |y_t - \phi_t' \alpha|^2$$

We assume that the input signal is quasi-stationary, that is, $\{x_t\}$ satisfies:

$$\overline{\mathbf{E}}[x_t x_{t-\tau}] = \lim_{N \to \infty} \frac{1}{N} \sum_{k=1}^{N} \mathbf{E}[x_{t+k} x_{t+k-\tau}]$$

$$= R_x(\tau)$$

In the previous lecture, we used the assumption that $\alpha \in \Lambda$, which is compact, to prove that:

$$\lim_{\ell \to \infty} R_{emp}^{\ell}(\alpha) \xrightarrow{uniformly,\ w.p.1} \overline{\mathbf{E}}[x_t x_{t-\tau}]$$

$$= R(\alpha) \quad \text{(The "true" risk)}$$

Under this setting, we have a result that is similar to the case where the data was produced as i.i.d. samples, but with some caveats, such as no results similar to the VC dimension. Also, we are working within a very restricted setting, that of linear regression. Some extensions exist for nonlinear settings, such as the case of pre-processing or post-processing by Lipschitz continuous functions.

1.1 The Estimator

Last time, we also wrote the expression for the estimate:

$$\hat{\alpha}_l = (\frac{1}{\ell}\sum_{t=1}^{\ell}\phi_t\phi_t')^{-1}(\frac{1}{\ell}\sum_{t=1}^{\ell}\phi_t y_t)$$

This expression assumes that the inverse exists, and its existence is guaranteed by the persistence of excitation of the input x_t. If we define

$$\xi_t' = (x_t, x_{t-1}, ..., x_{t-n}),$$

we say that x_t is persistently exciting (p.e.) of order n if

$$\overline{\mathbf{E}}\xi_t'\xi_t > 0$$

Intuitively, if we are trying to generate an input to test/learn a linear system with a certain number of degrees of freedom, then the input needs to have enough degrees of freedom.

2 P.E. Sufficient condition

Claim.

$$\frac{1}{\ell}\sum_{t=1}^{\ell}\phi_t\phi_t'$$

is invertible for large enough ℓ if x_t is persistently exciting of order $n_a + n_b + 1$.

Notice that this is a sufficient condition. We can get away with less, considering how the input is generated, taking into account w's excitation, the problem of pole-zero cancellation (where dimensionality is lost) etc... (ref. Ljung's book). Basically, there is a notion of identifiability of the system.

Corollary. *(Condition for p.e.): x_t is p.e. of order n if $\Phi_X(w) \neq 0$ for at least n frequencies.*

Examples:

White noise (flat spectrum)/ Approximation: pseudo-binary input, deteministic, periodic signals with similar spectral properties./n Sinusoids.

Because we allowed mismatch between $n_{a,b}$ and $n_{a,b}^{\star}$, $R(\alpha)$ is not zero. We do converge nonetheless. Now we will do a quantitative comparison when there is no mismatch.

3 Asymptotic Properties of $\hat{x}_l$

Here we assume that $n_a = n_a^{\star}$ and $n_b = n_b^{\star}$. Notice that $y_t = \phi_t'\alpha$ now. The estimator can be written as:

$$\hat{\alpha}_l = (\frac{1}{\ell}\sum_{t=1}^{\ell}\phi_t\phi_t')^{-1}(\frac{1}{\ell}\sum_{t=1}^{\ell}\phi_t[\phi_t'\alpha_0 + w_t])$$

$$= \alpha_0 + \underbrace{(\frac{1}{\ell}\sum_{t=1}^{\ell}\phi_t\phi_t')^{-1}}_{I}\underbrace{(\frac{1}{\sqrt{\ell}}\sum_{t=1}^{\ell}\phi_t w_t)}_{II}\frac{1}{\sqrt{\ell}}$$

I : From quasi-stationarity, we know that:

$$\frac{1}{\ell}\sum_{t=1}^{\ell}\phi_t\phi_t' \to \overline{\mathbf{E}}\phi_t\phi_t' = P$$

II : This term converges in distribution to a zero-mean normal random variable with covariance P. In the general case, this requires an elaborate proof (c.f. Ljung, chapter 9). However, for the case where $na = 0$, on can see that:

- Each element in the sum has zero mean.
- Due to finite regression, $\phi_t w_t$ values distant enough are independent. We can split the sum into $n_b + 1$ sums.
- Now, $\frac{1}{\sqrt{\ell}} \sum v_i$, with independent v_i's, converges to a normal distribution, by the central limit theorem. These can be combined to yield a special case of the general result.

Thus $(\hat{\alpha}_l - \alpha_0)\sqrt{\ell} \sim P^{-1} Z$, where $Z \sim \mathcal{N}(0, Q)$. By computation,

$$\lim_{\ell \to \infty} (\frac{1}{\sqrt{\ell}} \sum_{t=1}^{\ell} \phi_t w_t)^2 \to \lambda^2 P$$

And finally,

$$(\hat{\alpha}_l - \alpha_0)\sqrt{\ell} \sim \mathcal{N}(0, \lambda^2 P^{-1})$$

Which gives us a convergence rate of $\frac{1}{\ell}$ in variance, and gives us design guidelines. For example, to make P^{-1} as small as possible, we should make the smallest singular value of P as large as possible.[1]

4 State Space Models

The next step is to look for other types of dependencies. A general setting is that of state space representations, as used in linear system theory. hilosophically, a state is a canonical representation of memory. As we weill see, this notion forms a natural extension from the class of dependencies that we have seen to that of Hidden Markov Models (HMM's).

A state space description of a system is as follows:

$$x_{t+1} = f(x_t, u_t, w_t)$$

$$y_t = g(x_t, u_t, w_t)$$

$$x_t \in \mathbb{X},\ \mathbb{X} = \mathbf{R}^n \text{ or } \mathbb{X} = \{1, 2, ..., n\}$$
$$y_t \in \mathbb{Y},\ \mathbb{Y} = \mathbf{R}^m \text{ or } \mathbb{Y} = \{1, 2, ..., m\}$$
$$u_t \in \mathbb{U},\ \mathbb{U} = \mathbf{R}^p \text{ or } \mathbb{U} = \{1, 2, ..., p\}$$

Assume w_t and v_t are i.i.d., and $v_t \perp w_t$. Note the following:

- This representation can either be perceived as a direct dynamic description, or as a probabilistic description that gives Markovianity.
- Auto Regressive with External input (ARX) models (models like in equation (1) can be written as s.s. models.)
- HMM's are probabilistic state space systems. There, the system is not driven, i.e. $u_t = 0$. We will elaborate on this next time.

[1] This completes a scan over Ljung's book. We can make models more complicated, but these are the basic ideas. Reference: OCW website of old 6.435 for examples/details.