Chapter 6

Prediction Error Methods

> "How far can we go by optimizing the predictive performance of an estimated model?"

This chapter studies the parameter estimation technique called the Prediction Error Method (PEM). The idea is that rather than a plain least squares approach, or a statistical maximum likelihood approach there is a third important principle in use for estimating the parameters of a dynamic model based on recorded observations. This technique considers the accuracy of the predictions computed for the observations, rather than the model mismatch are the likelihood of the corresponding statistical model. This technique is perhaps the most tightly connected to systems theory as it explicitly exploits the dynamical structure of the studied system. Those three design principles are represented schematically in Fig. (6.1). In a number of cases the three design decision leads to the same estimators as will be discussed in some detail.

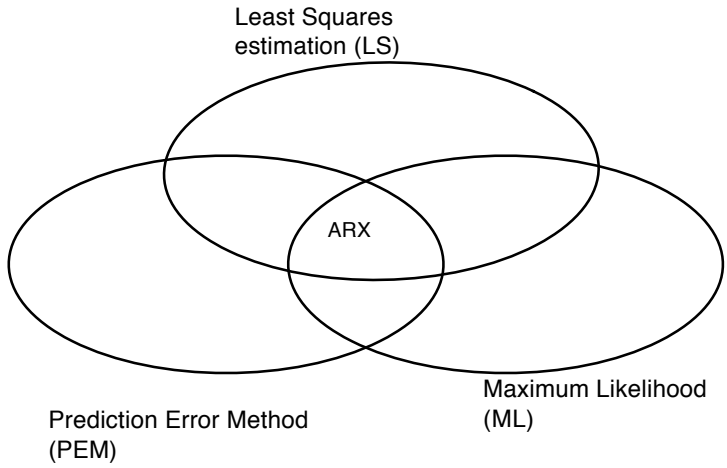

Figure 6.1: Schematic Illustration of the different approaches which one could take for estimation of parameters.

6.1 Identification of an ARX model

Let's start the discussion with a study of a convenient model. The result is not too difficult to obtain, but the consecutive steps in the analysis will come back over and over in later sections. Consider a system relating two signals $\{u_t\}_{t=-\infty}^{\infty}$ and $\{y_t\}_{t=-\infty}^{\infty}$ which is modeled as

$$A(q^{-1})y_t = B(q^{-1})u_t + e_t,\ \forall t = \ldots, 0, 1, , 2, \ldots, \tag{6.1}$$

where for given $n_a, n_b > 0$ one has $A(q^{-1}) = 1 + a_1 q^{-1} + \cdots + a_{n_a} q^{-n_a}$ and $B(q^{-1}) = b_1 q^{-1} + \cdots + b_{n_b} q^{-n_b}$, with fixed but unknown coefficients $\{a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b}\}$. Here the residuals $\{e_t\}_t$ are small in some sense, but unknown otherwise. This system can be written equivalently as

$$y_t = \varphi_t^T \theta + e_t,\ \forall t = \ldots, 0, 1, , 2, \ldots, \tag{6.2}$$

where

$$\begin{cases} \varphi_t^T = (-y_{t-1}, \ldots, -y_{t-n_a}, u_{t-1}, \ldots, u_{t-n_b})^T \in \mathbb{R}^{n_a+n_b},\ \forall t \\ \theta = (a_1, \ldots, a_{n_a}, b_1, \ldots, b_{n_b})^T \in \mathbb{R}^{n_a+n_b}. \end{cases} \tag{6.3}$$

The model is linear in the parameters, hence it is already known how to estimate the parameter vector θ from given samples $\{(\varphi_t, y_t)\}_{t=1}^n$ induced by the signals $\{u_t\}_t$ and $\{y_t\}_t$. Note that if the signals are only recorded at time instances $t = 1, 2, \ldots, n$, one can only construct the samples $\{(\varphi_t, y_t)\}_{t=1+\max(n_a,n_b)}^n$. - for notational convenience we shall assume further that the signals are observed fully such that $\{(\varphi_t, y_t)\}_{t=1}^n$ can constructed. The Least Squares (LS) estimation problem is

$$\min_{\hat{\theta}=(a_1,\ldots,a_{n_a},b_1,\ldots,b_{n_b})} \sum_{t=1}^{n} (y_t + a_1 y_{t-1} + \cdots + a_{n_a} y_{t-n_a} - b_1 u_{t-1} - \cdots - b_{n_b} u_{t-n_b})^2 = \sum_{t=1}^{n} (\varphi_t \hat{\theta} - y_t)^2, \tag{6.4}$$

and the estimate $\hat{\theta}$ is given as the solution to

$$\left(\frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^T \right) \hat{\theta} = \left(\frac{1}{n} \sum_{t=1}^{n} y_t \varphi_t \right), \tag{6.5}$$

which are known as the normal equations associated to problem (6.4). If the matrix

$$\Phi = \left(\frac{1}{n} \sum_{t=1}^{n} \varphi_t \varphi_t^T \right), \tag{6.6}$$

is of full rank the estimate is unique and is given as

$$\hat{\theta} = \Phi^{-1} \left(\frac{1}{n} \sum_{t=1}^{n} y_t \varphi_t \right). \tag{6.7}$$

Such approach is also related as an 'equation error method' since the errors we minimize derive directly from $\{e_t\}_t$ which occur as equation errors in (6.1).

The normal equations can readily be solved with the numerical tools described in Chapter 1. For the statistical properties it is of crucial importance which setup is assumed. We will work with

the assumption that $\{e_t\}_t$ are modeled as random variables, and hence so are $\{y_t\}_t$ and $\{\varphi_t\}_{t=1}^n$. This is an important difference with a classical analysis of a LS approach as given in Section ... as there one assumes φ_t is deterministic. The reason that this difference is important is that when taking expectations various quantities, it is no longer possible to treat Φ nor Φ^{-1} as a constant matrix.

The common statistical assumptions used to model and analyze this problem go as follows. Formally, let the signals $\{U_t\}_t$ and $\{Y_t\}_t$ be stationary stochastic processes related as

$$Y_t = \varphi_t^T \theta_0 + V_t, \ \forall t = \ldots, 0, 1, , 2, \ldots, \tag{6.8}$$

where $\theta_0 \in \mathbb{R}^{n_a+n_b}$ is the fixed but unknown 'true' parameter vector, the vector $\varphi_t = (-Y_{t-1}, \ldots, Y_{t-n_a}, U_{t-1}, \ldots, U_{t-n_b})^T$ which takes values in $\mathbb{R}^{n_a+n_b}$, and where we assume that $\{V_t\}_t$ is a stationary stochastic process independent of the input signal $\{U_t\}_t$. If an estimate $\hat{\theta}$ is 'good', it should be in some sense 'close' to θ_0. Lets examine then how good the LS estimator is. From the normal equations one gets

$$\begin{aligned}\hat{\theta} - \theta_0 &= \left(\frac{1}{n}\sum_{t=1}^n \varphi_t\varphi_t^T\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^n \varphi_t Y_t\right) - \left(\frac{1}{n}\sum_{t=1}^n \varphi_t\varphi_t^T\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^n \varphi_t\varphi_t^T\right)\theta_0 \\ &= \left(\frac{1}{n}\sum_{t=1}^n \varphi_t\varphi_t^T\right)^{-1}\left(\frac{1}{n}\sum_{t=1}^n V_t\varphi_t\right). \end{aligned} \tag{6.9}$$

Under weak conditions, the normalized sums tend to their expected values when n tends to infinity. Hence $\hat{\theta} \to \theta_0$, or $\hat{\theta}$ is consistent if

$$\begin{cases} \mathbb{E}\left[\varphi_t\varphi_t^T\right] \text{ is nonsinguar} \\ \mathbb{E}[\varphi_t V_t] = 0. \end{cases} \tag{6.10}$$

The first condition ('nonsingular') is often satisfied, but there are a few important exceptions:

- The inputs $\{U_t\}$ is not sufficiently rich: it is not PE of order n_b.
- The data is noise-free (i.e. $V_t = 0$ for all t), and the model orders are chosen too high: this implies that $A_0(q^{-1})$ and $B_0(q^{-1})$ associated with θ_0 have common factors (are not coprime).
- The input signal $\{U_t\}_t$ is generated by a linear low-order feedback law from the output $\{Y_t\}_t$.

Unlike the 'nonsingular' condition, the requirement $\mathbb{E}[\varphi_t V_t] = 0$ is in general *not* satisfied. An important exception is when $\{V_t\}_t$ is white noise, i.e. is a sequence of uncorrelated random variables. In such case, $\{V_t\}_t$ will be uncorrelated with all past data, and in particular V_t will be uncorrelated with φ_t, implying the condition.

The LS estimation technique is certainly simple to use. In case those requirements are not at all satisfied, we need modifications to the LS estimate to make it 'work', i.e. make the estimate consistent or at least not too biased. We will study two such modifications.

- Minimization of the prediction error for 'more detailed' model structures. This idea leads to the class of Prediction Error Methods (PEM) dealt with in this chapter.
- Modification of the normal equations associated to the LS estimator. This idea leads to the class of Instrumental Variables dealt with in Chapter ... .

6.2 Optimal Prediction and PEM

A model obtained by identification can be used in many ways depending on the purpose of modeling. In may applications the aim of the model is prediction in some way. It therefore makes sense to determine the model such that the prediction error would be minimal. Let us consider the SISO case at first. We denote the model prediction error here as

$$\epsilon_t(\theta) = y_t - f_{t|t-1}(\theta),\ \forall t = 1, 2, \ldots, \tag{6.11}$$

where θ represents the parameters of the current model, and $f_{t|t-1}(\theta)$ represents the prediction of the outcome y_t using all past information and the model determined by θ. In case of an ARX model as described in the previous chapter, we have obviously that

$$f_{t|t-1}(\theta) = \varphi_t^T \theta. \tag{6.12}$$

In the context of PEM methods one is in general interested in more general models. Suppose a general LTI describes the signals $\{u_t\}_t$ and $\{y_t\}_t$ as

$$y_t = G(q^{-1}, \theta)u_t + H(q^{-1}, \theta)V_t,\ \forall t = \ldots, 0, 1, \ldots, \tag{6.13}$$

where we assume that $\{V_t\}_t$ is a stochastic process with $\mathbb{E}[V_s V_t^T] = \sigma^2 \delta_{s,t}$ with $\delta_{s,t} = 1$ if $s = t$, and zero otherwise. For notational convenience, assume that $G(0; \theta) = 0$, i.e. that the model has at least one pure delay from input to output. Then, the optimal predictor can be written as

$$f_{t|t-1}(\theta) = L_1(q^{-1}, \theta)y_t + L_2(q^{-1}, \theta)u_t,\ \forall t = \ldots, 0, 1, \ldots, \tag{6.14}$$

which is a function of the past data only if $L_1(0, \theta) = L_2(0, \theta) = 0$. Suppose we have for our model (H, G) corresponding mappings (L_1, L_2). Now, a PEM method will estimate the parameter vector θ by optimizing the prediction performance, i.e.

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}} \frac{1}{n} \sum_{t=1}^{n} \ell(\epsilon_t(\theta)), \tag{6.15}$$

where $\ell : \mathbb{R} \to \mathbb{R}$ is a loss-function. E.g. $\ell(e) = e^2$.

Now we study the question how to go from an LTI model (G, H) to the corresponding predictors $(L_1, L2)$. Again let us introduce ideas using a series of elementary examples.

Example 41 (White Noise) *Assume a realization $\{e_1, e_2, \ldots, e_t\}$ of zero mean white (uncorrelated) noise. Given the values of $(e_1, e_2, \ldots, e_{t-1})$, the best estimate of e_t in L_2 sense is then $\hat{e}_t = 0$, That is*

$$\hat{e}_t = \underset{\sum_{\tau=1}^{t+1} h_\tau q^{-\tau}}{\operatorname{argmin}} \mathbb{E}\left(e_t - \sum_{\tau=1}^{t+1} h_\tau e_{t-\tau}\right)^2 = \underset{\sum_{\tau=1}^{t+1} h_\tau q^{-\tau}}{\operatorname{argmin}} \mathbb{E}[e_t^2] + \sum_{\tau=1}^{t-1} h_\tau \mathbb{E}[e_{t-\tau}], \tag{6.16}$$

and the minimum is clearly achieved when $h_1 = \cdots = h_\tau = 0$.

Example 42 (FIR(d)) *Given a deterministic sequence $\{u_t\}_{t=1}^n$, and given a realization $\{y_t\}_t$ of a process $\{Y_t\}_{t=1}^n$ which satisfies a FIR system, or*

$$Y_t = b_1 u_{t-1} + \cdots + b_d u_{t-d} + D_t, \tag{6.17}$$

where $\{D_1, \dots, D_n\}$ is a zero mean white noise sequence with bounded variance. Then the optimal prediction at instance $t+1$ is clearly

$$\hat{y}_{t+1} = b_1 u_t + \cdots + b_d u_{t-d+1}, \tag{6.18}$$

for any $t = d, \dots, n-1$.

Example 43 (AR(d)) *Given a realization $\{y_t\}_t$ of a process $\{Y_t\}_{t=1}^n$ which satisfies a AR(d) system, or*

$$Y_t + a_1 Y_{t-1} + \dots a_d Y_{t-d} = D_t, \tag{6.19}$$

where $\{D_1, \dots, D_n\}$ is a zero mean white noise sequence with bounded variance. Then the optimal prediction at instance $t+1$ is clearly

$$\hat{y}_{t+1} = a_1 y_t + \cdots + a_d y_{t-d+1}, \tag{6.20}$$

for any $t = d, \dots, n-1$.

Example 44 (MA(d)) *Given a realisation $\{y_t\}_t$ of a process $\{Y_t\}_{t=1}^n$ which satisfies a MA(d) system, or*

$$Y_t = D_t + c_1 D_{t-1} + \dots c_d D_{t-d}, \tag{6.21}$$

where $\{D_1, \dots, D_n\}$ is a zero mean white noise sequence with bounded variance. Equivalently,

$$Y_t = C(q^{-1}) D_t, \tag{6.22}$$

for any $t = d, \dots, n-1$. Thus

$$\begin{cases} D_t = C^{-1}(q^{-1}) Y_t \\ Y_t = (C(q^{-1}) - 1) D_t + D_t, \end{cases} \tag{6.23}$$

where the second equality separates nicely the contribution of the past noise $D_{t-1}, D_{t-2}, \dots$ on which we have some knowledge, and the present term D_t which is entirely unknown to us. This is a consequence of the fact that C is a monomial, i.e. the zeroth order term equals 1. Then it is not too difficult to combine both equations in (6.23) and then we find the corresponding optimal predictor as

$$\hat{Y}_t = \left(C^{-1}(q^{-1}) - 1\right) Y_t. \tag{6.24}$$

Those elementary reasonings lead to the optimal predictors corresponding to more complex models, as e.g.

Example 45 (ARMAX(1,1,1) model) *Consider the stochastic signals $\{U_t\}_t$ and $\{Y_t\}_t$ both taking values in $\mathbb{R}$ which follow a fixed but unknown system*

$$Y_t + aY_{t-1} = bU_{t-1} + V_t + cV_{t-1}, \ \forall t = \dots, 0, \dots, \tag{6.25}$$

where $\{V_t\}_t$ is zero mean white noise with $\mathbb{E}[V_t V_s] = \delta_{t,s}\lambda^2$. The parameter vector is $\theta = (a, b, c)^T \in \mathbb{R}^3$. Assume V_t is independent of U_s for all $s < t$, and hence the model allows for feedback from $\{Y_t\}_t$ to $\{U_t\}_t$. The output at time t satisfies

$$y_t = (-aY_{t-1} + bU_{t-1} + cV_{t-1}) + V_t, \ \forall t = \dots, 0, \dots, \tag{6.26}$$

and the two terms on the right hand side (r.h.s.) are independent by assumption. Now let $y_t^* \in \mathbb{R}$ *be any number serving as a prediction, then one has for* t *that*

$$\mathbb{E}[Y_t - y_t^*]^2 = \mathbb{E}[-aY_{t-1} + bU_{t-1} + cV_{t-1}]^2 + \mathbb{E}[V_t]^2 \geq \lambda^2, \tag{6.27}$$

giving as such a lower-bound to the prediction error variance. An optimal predictor $\{f_{t|t-1}(\theta)\}_t$ *is one which achieves this lower-bound. This is the case for*

$$f_{t|t-1}(\theta) = -aY_{t-1} + bU_{t-1} + cV_{t-1}. \tag{6.28}$$

The problem is of course that this predictor cannot be used as it stands as the term V_{t-1} *is not measurable. However, it* V_{t-1} *may be reconstructed from past data as the residual in the previous iteration, and as such*

$$\begin{aligned} f_{t|t-1}(\theta) &= -aY_{t-1} + bU_{t-1} + cV_{t-1} \\ &= -aY_{t-1} + bU_{t-1} + c\left(Y_{t-1} + aY_{t-2} - bU_{t-2} - cV_{t-2}\right) \\ &= -aY_{t-1} + bU_{t-1} + c\left(Y_{t-1} + aY_{t-2} - bU_{t-2}\right) - c^2\left(Y_{t-2} + aY_{t-3} - bU_{t-3} - cV_{t-3}\right) \\ &= \dots \\ &= \sum_{i=1}^{t-1}(c-a)(-c)^{i-1}Y_{t-i} - a(-c)^{t-1}Y_0 + b\sum_{i=1}^{t-1}(-c)^{i-1}U_{t-i} - (-c)^t V_0. \end{aligned} \tag{6.29}$$

Under assumption that $|c| < 1$ *the last term can be neglected for large* t *as it will have an exponentially decaying transient effect. Then we get a computable predictor. However we reorder terms to get a more practical expression as*

$$f_{t|t-1}(\theta) = f(t-1|t-2,\theta) + (c-a)Y_{t-1} + bU_t, \tag{6.30}$$

which gives a simple recursion for computing the optimal prediction corresponding to past observations and the model parameter vector θ*. We can compute the corresponding prediction error* $\epsilon_t(\theta) = Y_t - f_{t|t-1}(\theta)$ *similarly as*

$$\epsilon_t(\theta) + c\epsilon_{t-1}(\theta) = Y_t + cY_{t-1} - \left((c-a)Y_{t-1} + bU_{t-1}\right) = Y_t + aY_{t-1} - bU_{t-1}, \tag{6.31}$$

for any $t = 2, \dots, n$*. This recursion needs an initial value* $\epsilon_t(\theta)$ *which is in general unknown and often set to 0. Observe that we need the statistical framework only for a definition of what an optimal predictor means exactly as in (6.27).*

The above analysis can be stated more compactly using the polynomials,

Example 46 (An ARMAX(1,1,1), bis) *Consider* $\{u_t\}_t$ *and* $\{y_t\}_t$ *obeying the system*

$$(1 + aq^{-1})y_t = (bq^{-1})u_t + (1 + cq^{-1})e_t, \tag{6.32}$$

$$\begin{cases} f(t|t-1,\theta) & = \mathbf{H}^{-1}(q^{-1},\theta)\mathbf{G}(q^{-1},\theta)U_t + \left(1 - \mathbf{H}^{-1}(q^{-1},\theta)\right) Y_t \\ \epsilon_t(\theta) = V_t & = \mathbf{H}^{-1}(q^{-1},\theta)(Y_t - \mathbf{G}(q^{-1},\theta)U_t). \end{cases} \tag{6.37}$$

Figure 6.2: The optimal expected least squares predictor for a general LTI.

for all t. Then

$$\begin{aligned} y_t &= \frac{(bq^{-1})}{(1+aq^{-1})}u_t + \frac{(1+cq^{-1})}{(1+aq^{-1})}e_t \\ &= \frac{(bq^{-1})}{(1+aq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+aq^{-1})}e_t + \frac{(1+aq^{-1})}{(1+aq^{-1})}e_t \\ &= \frac{(bq^{-1})}{(1+aq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+aq^{-1})}\left(\frac{(1+aq^{-1})y_t - (bq^{-1})u_t}{(1+cq^{-1})}\right) + e_t \\ &= \left(\frac{(bq^{-1})}{(1+aq^{-1})} - \frac{(c-a)q^{-1}}{(1+aq^{-1})}\frac{(bq^{-1})}{(1+cq^{-1})}\right)u_t + \frac{(c-a)q^{-1}}{(1+aq^{-1})}\frac{(1+aq^{-1})}{(1+cq^{-1})}y_t + e_t \\ &= \frac{(bq^{-1})}{(1+cq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+cq^{-1})}y_t + e_t, \end{aligned} \tag{6.33}$$

and again because of the noise terms e_t cannot be predicted from the past or the model parameters θ, the best any predictor can do is

$$f_{t|t-1}(\theta) = \frac{(bq^{-1})}{(1+cq^{-1})}u_t + \frac{(c-a)q^{-1}}{(1+cq^{-1})}y_t, \tag{6.34}$$

yielding the result. When working with filters in this way it is assumed that data are available from the infinite past. Since this wouldn't be the case in practical situations, one has to take into account transient effects before implementing thus predictors.

In general the derivation goes as follows. Assume the data $(y_1, y_2, \dots)$ and $(u_1, u_2, \dots)$ follows an LTI model where

$$y_{t+1} = H(q^{-1};\theta_0)u_{t+1} + G(q^{-1};\theta_0)e_{t+1}. \tag{6.35}$$

where

$$\begin{cases} H(q^{-1};\theta_0) = 1 + h_1 q^{-1} + \cdots + h_{m_h} q^{-m_h} \\ G(q^{-1};\theta_0) = 1 + g_1 q^{-1} + \cdots + g_{m_g} q^{-m_g}, \end{cases} \tag{6.36}$$

where $m_h \geq 1$ and $m_g \geq 1$ denote the orders of both monic polynomials, and $\theta_0 = (h_1, \dots, h_{m_h}, g_1, \dots, g_{m_g}) \in \mathbb{R}^{m_g+m_h-2}$. Then we face the question what value of

6.3 Statistical Analysis of PEM methods

The statistical analysis of PEM estimates starts off similar as in the least squares case. Assume that the observed signals satisfy a stochastic signal, or that

$$Y_t = G(q^{-1},\theta_0)U_t + G(q^{-1},\theta_0)D_t. \tag{6.38}$$

- The observed signals $\{u_t\}_{t=1}^n \subset \mathbb{R}$ and $\{y_t\}_{t=1}^n \subset \mathbb{R}$ are assumed to be samples from quasi-stationary stochastic processes $\{U_t\}_t$ and $\{Y_t\}_t$.
- The noise $\{D_t\}$ is assumed to be a stationary process with zero mean.
- The input is with high probability Persistently Exciting (PE) of sufficient order. that is $\mathbb{E}[V_t V_t^T] \succ 0$ where $V_t = (U_{t-1}, \ldots, U_{t-d})$, and hence $\sum_{t=1}^n \mathbf{u}_t \mathbf{u}_t^T \succ 0$ where $\mathbf{u}_t = (u_{t-1}, \ldots, u_{t-d}) \in \mathbb{R}^d$.
- The filters $G(q^{-1}, \theta)$ and $H(q^{-1}, \theta)$ are smooth (differentiable) functions of the parameters.

Then a PEM approach would be to solve for θ

$$V_n^* = \min_\theta V_n(\theta) = \frac{1}{2} \sum_{t=1} n \left(Y_t - (\mathbf{H}^{-1}(q^{-1}, \theta)\mathbf{G}(q^{-1}, \theta)U_t + \left(1 - \mathbf{H}^{-1}(q^{-1}, \theta)\right) Y_t) \right)^2 . \tag{6.39}$$

This approach is in general different from a LS estimate. We also need the following assumption, namely that

- The Hessian $V_n''(\theta)$ is non-singular at least for the parameters θ close to the true parameters θ_0. This implies that no different parameters can solve the PEM objective asymptotically, and is thus in a sense closely related to Persistency of Excitation (PE).

The proof that the PEM would result in accurate estimates in that case is quite involved, but the main reasoning is summarized in Fig. (6.3). This result is then found strong enough also to quantify the variance of the estimates if n tends to infinity. Specifically we have that

$$\sqrt{n}(\theta_n - \theta_0) \sim \mathcal{N}(0, \mathbf{P}), \tag{6.40}$$

where

$$\mathbf{P} = \mathbb{E}[D_t^2]\mathbb{E}\left[\varphi_t(\theta_0)\varphi_t(\theta_0)^T\right]^{-1}, \tag{6.41}$$

and where

$$\varphi_t(\theta_0) = \frac{d\epsilon_t(\theta)}{d\theta}\Big|_{\theta=\theta_0}. \tag{6.42}$$

That is, the estimates are asymptotically unbiased and have asymptotic variance which is given by the Fisher information matrix based on the gradients of the prediction errors evaluated at the true parameters.

6.4 Computational Aspects

The practical difference of a PEM approach to a LS approach is that the solution is not given in closed form as the normal equations before did. In general, one has to resort to numerical optimization tools to solve the optimization problem. While good software implementations exists that can do this task very accurate, it is useful to write out some common approaches for getting a feeling how to interpret results from such a software.

The prototypical approach goes as follows. Let us abstract the problem as the following optimization problem over a vector $\theta \in \mathbb{R}^d$ as

$$\theta^* = \operatorname*{argmin}_\theta J(\theta), \tag{6.43}$$

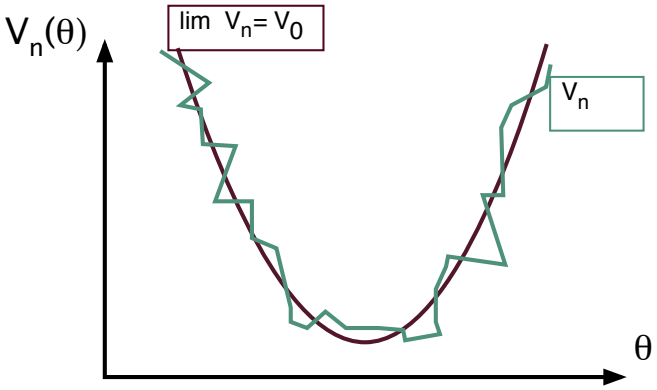

Figure 6.3: Schematic example of the PEM cost function. Here θ denotes the parameter vector to be minimized over. In case only a finite number of samples n is available, the PEM objective is a noisy version of the asymptotical loss $V_0(\theta) = \lim_{n\to\infty} V_n(\theta)$. Two results are stated then: (i) the true parameters θ_0 are the minimizer of the asymptotic objective function, and (ii) the asymptotic objective function $V_0\theta)$ differs not too much from the sample objective function $V_n(\theta)$ for *any* ('uniform) θ. Hence the minimizer θ_n to V_n is not too different from the true parameters.

where $J : \mathbb{R}^d \to \mathbb{R}$ is a proper cost function (i.e. a minimal value exists). We have an iterative regime, and in each iteration the previous estimate is refined slightly. Formally, we generate a sequence of vectors from an initial estimate $\theta^{(0)}$, obeying the recursion

$$\theta^{(k+1)} = \theta^{(k)} + \gamma \mathbf{b}(J, \theta^{(k)}), \tag{6.44}$$

where $\mathbf{b}(J, \theta^{(k)}) \in \mathbb{R}^d$ is a correction ('step') which refines the estimator. The algorithm then hopefully converges, in the sense that $\theta^{(k)} \to \theta^*$ when k increases. See Fig. (6.4.a) for an cost function in 2D, and Fig. (6.4.b) for an iterative algorithm at work in 1D. Now different algorithms specialize further using different quantities.

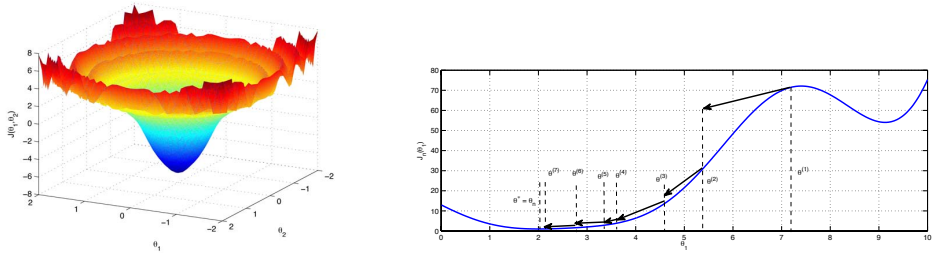

Figure 6.4: An example of an iterative optimization routine of J over a parameter θ.

The prototypical algorithm goes as follows. Here, the correction factor is determined by using a quadratic approximation of the cost function J at the current estimate $\theta^{(k)}$. The algorithm follows

the recursion

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k \left(V_n''(\theta^{(k)}) \right)^{-1} V_n'(\theta^{(k)}), \tag{6.45}$$

where

- α_k is the step size, typically taken as a positive decreasing function of k.
- $V_n'(\theta_n) \in \mathbb{R}^d$ denotes the gradient of the cost function J at $\theta^{(k)}$.
- $V_n''(\theta_n) \in \mathbb{R}^{d \times d}$ denotes the Hessian matrix of the cost function J at $\theta^{(k)}$.

This algorithm is referred to as Newton-Raphson.

In the optimal point θ^* for the PEM problem one has a simplified approximative expression for the cost function $J(\theta^*)$ given as

$$V_n''(\theta^*) \approx \frac{2}{n} \sum_{i=1}^{n} \psi_t^T(\theta^*) \mathbf{H} \psi_t(\theta^*), \tag{6.46}$$

where $\mathbf{H}$ is a given matrix, and $\psi_t(\theta^*)$ equals the (first order) influence of the tth sample on the loss function of the PEM objective. Using this approximation in an iterative optimization gives the Gauss-Newton recursive algorithm given as

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k \left(\sum_{i=1}^{n} \psi_t^T(\theta^{(k)}) \mathbf{H} \psi_t(\theta^{(k)}) \right)^{-1} \left(\sum_{i=1}^{n} \psi_t^T(\theta^*) \mathbf{H} \epsilon_t(\theta^{(k)}) \right), \tag{6.47}$$

where here $\epsilon_t(\theta^{(k)})$ denotes the prediction error on y_t using the past samples and the model with parameters $\theta^{(k)}$. When n is quite large both algorithms (6.48) and (6.45) behave quite similarly. But in general, the Newton-Raphson converges with quadratic speed $1/n^2$. The Gauss-Newton approach converges 'only' (super-) linear, but has the additional advantage that each iteration of the algorithm can be computed and stored much more efficiently.

If computational issues are even more important in the case at hand one may resort to a steepest descent algorithm, implementing the recursion

$$\theta^{(k+1)} = \theta^{(k)} - \alpha_k V_n'(\theta^{(k)}), \tag{6.48}$$

where $\theta^{(0)} \in \mathbb{R}^d$ is an appropriate initial estimate. Such algorithm is referred to as a steepest descent or gradient descent algorithm.

There are three important caveats when using such an approaches.

- The numerical software might be stuck in local minima, most likely giving parameter estimates which are not useful. This depends largely on the shape of the loss function. If V_n were (almost) a positive quadratic function, then there are no such local 'false' minima a numerical optimization routine could be stuck in. A simple way to circumvent this problem is to let the optimizer run based on a number of different starting points: if the optima are not mostly not equal, the problem of local minima is severely present. On the other hand, if the optimizers were equal for most of them, it is not too large a stretch to assume that the global minimizer were found successfully.

- In a number of cases the loss function V_n might not be differentiable at certain values for θ, or lead to large values, preventing the solver to converge properly. This is typically the case when the underlying dynamics are almost unstable, and slight deviations of the parameters might lead to unbounded predictions.
- The uncertainties computed by the software based on the discussion in the previous section is often not valid for finite n. Specifically, the derivation there assumes that θ_n is close enough to θ_0 in order to admit a quadratic expansion of the loss between either. This is clearly not valid θ_n were only a local minimizer. Hence, the estimated variances in software routines are valid conditioned on the fact that the optimizer worked well.