NOTES FOR THE COURSE:

SYSTEM IDENTIFICATION AND THE LIMITS OF LEARNING FROM DATA

M.C. Campi

printed September 12, 2006

Contents

Chapter 1

ASYMPTOTIC CONVERGENCE OF PEM METHODS IN A STATIONARY ENVIRONMENT

The model that is found when applying an identification procedure is stochastic since it depends on the probabilistic components entering the identification problem. In the long run, however, the probabilistic components average out and the random fluctuations disappear. Then, what we are left with is an asymptotic model which only depends on intrinsic characteristics of the identification problem.

This chapter studies the properties of such an asymptotic model for prediction error minimization (PEM) methods in a stationary environment. First, we consider the case of linear models and characterize the limit point θ^* to which the parameter estimate $\hat{\theta}_t$ converges. Extending these results to a nonlinear contexts presents certain difficulties that are discussed in turn.

The asymptotic theory of PEM methods pertains to a thought experiment, since in the real-world we never have an infinite number of data. Nevertheless, the asymptotic analysis help gain insight and develop confidence in the identification methods.

1.1 A preview example: least squares estimate

Consider the system

$$y_t = \theta^\circ u_t + w_t, \tag{1.1}$$

where $\theta^\circ \in [0, 1]$, u is an independent sequence taking on value -1 or 1 with probability 0.5 each and w is a white Gaussian noise with zero mean and σ_w^2 variance.

Parameter θ° is unknown and an estimate $\hat{\theta}_t$ is obtained by minimizing the quadratic cost

$$\frac{1}{t}\sum_{k=1}^{t}\left[y_k - \hat{y}_k(\theta)\right]^2, \tag{1.2}$$

where $\hat{y}_t(\theta) = \theta u_t$, is the predictor.

The parameter which is optimal according to the criterion of minimizing the prediction error variance is given by

$$\theta^* := \arg\min_{\theta\in[0,1]} E\left[\left(y_k - \hat{y}_k(\theta)\right)^2\right], \tag{1.3}$$

and is easily seen to coincide with θ°.

Now, the question of interest here is: is it true that $\hat{\theta}_t \to \theta^*$, as $t \to \infty$?

The answer is readily found by a direct computation. The minimizer of the quadratic cost (1.2) is $\hat{\theta}_t = \frac{1}{t}\sum_{k=1}^{t} u_k y_k$, so that $\hat{\theta}_t - \theta^* = \frac{1}{t}\sum_{k=1}^{t} u_k w_k$. Appealing to the law of large numbers (see Theorem A.13 in Appendix A) we then see that $\hat{\theta}_t - \theta^* \to 0$, almost surely, so that $\hat{\theta}_t$ does converge to θ^*.

In the present example, we were able to directly verify convergence since in the least squares case the estimate $\hat{\theta}_t$ admits a simple closed-form solution. In more general situations, though this route can in principle still be followed, it is very difficult to carry through the computations along this line and different routes have to be found. In the following, we indicate a different way of proceeding which lends itself to be applied to more general contexts as well. In fact, this appraoch will be followed in the subsequent sections of this chapter.

Let $\mathcal{J}_t(\theta) := \frac{1}{t}\sum_{k=1}^{t}\left[y_k - \hat{y}_k(\theta)\right]^2$ and $\bar{\mathcal{J}}(\theta) := E\left[\left(y_k - \hat{y}_k(\theta)\right)^2\right]$ and add the additional assumption that the estimate $\hat{\theta}_t$ is sought in the feasibility set $[0, 1]$. We no longer compute the estimate $\hat{\theta}_t$. Instead, we compare the empirical cost $\mathcal{J}_t(\theta)$ with its theoretical counterpart $\bar{\mathcal{J}}(\theta)$ over the entire $[0, 1]$ interval and, from this, draw conclusions on the convergence $\hat{\theta}_t \to \theta^*$.

A simple computation shows that

$$\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta) = (\theta^* - \theta)^2\frac{1}{t}\sum_{k=1}^{t}(u_t^2 - 1) + (\theta^* - \theta)\frac{2}{t}\sum_{k=1}^{t} u_k w_k + \frac{1}{t}\sum_{k=1}^{t}(w_k^2 - 1), \tag{1.4}$$

from which

$$\sup_{\theta\in[0,1]} \left|\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right| \leq \sup_{\theta\in[0,1]} |\theta^* - \theta| \left|\frac{2}{t}\sum_{k=1}^{t} u_k w_k\right| + \left|\frac{1}{t}\sum_{k=1}^{t}(w_k^2 - 1)\right|. \tag{1.5}$$

In the right hand side, $\sup_{\theta\in[0,1]} |\theta^* - \theta| \leq 1$ (note that here the boundedness of the set to which θ belongs is essential), while the other two absolute values tend to zero by an application of the law of large numbers. Thus, we conclude that $\mathcal{J}_t(\theta)$ tends almost surely to $\bar{\mathcal{J}}(\theta)$, uniformly in θ, that is, for any $\epsilon > 0$, with probability 1 we can find an instant t large enough such that the maximal distance between $\mathcal{J}_t(\theta)$ and $\bar{\mathcal{J}}(\theta)$ does not exceed ϵ. This is illustrated in Figure 1.1.

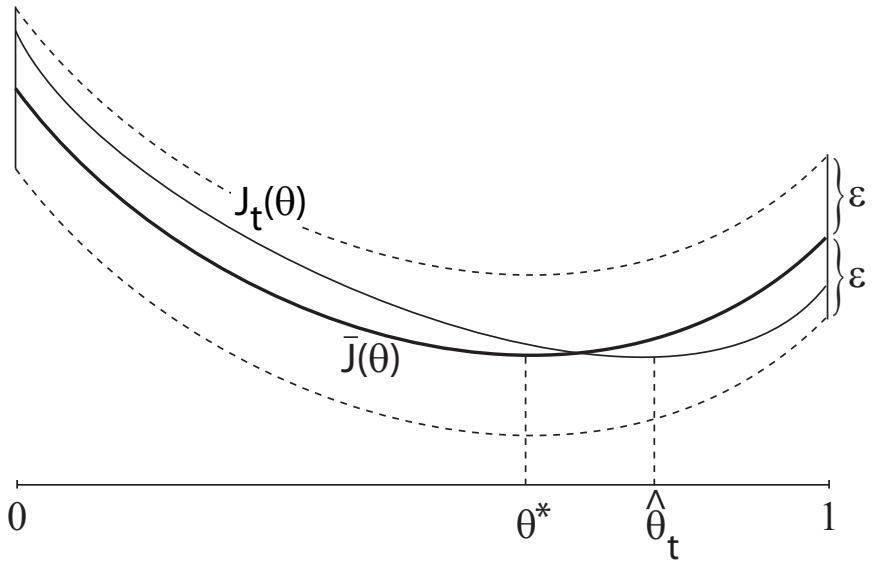

Figure 1.1: $\mathcal{J}_t(\theta)$ versus $\bar{\mathcal{J}}(\theta)$

Using the uniform conergence is then not difficult to conclude that $\hat{\theta}_t \to \theta^*$.

The interest in this second approach is that it can be extended to general linear (but not necessarily linear in the parameter) predictors, so that it basically covers all situations in linear system identification. We start in the next section by formalizing the idea that uniform convergence of $\mathcal{J}_t(\theta)$ to $\bar{\mathcal{J}}(\theta)$ implies $\hat{\theta}_t \to \theta^*$ and then apply this result to system identification in the subsequent section.

1.2 A lemma of general use

Before introducing the lemma, let us recapitulate the problem under study in this chapter. Let $\mathcal{J}_t(\theta)$ be the identification cost at time t and note that it is a random variable. Take

the expectation $E[\mathcal{J}_t(\theta)]$ and assume that it does not depend on time t (this is true under stationarity assumptions) and introduce the shorthand $\bar{\mathcal{J}}(\theta)$ for $E[\mathcal{J}_t(\theta)]$. The question we intend to answer is: is it true that, for t large, minimizing $\mathcal{J}_t(\theta)$ is equivalent to minimizing $\bar{\mathcal{J}}(\theta)$? As we shall see, this is indeed the case under weak conditions. Moreover, this result will permit us to draw interesting conclusions about the characteristics of the asymptotically estimated model.

In this section, a simple lemma that can be applied to different contexts to prove convergence of $\mathcal{J}_t(\theta) \to \bar{\mathcal{J}}(\theta)$ is presented. This lemma is instrumental to our derivations in later sections.

The lemma is deterministic. In a stochastic setting, it can be applied pathwise.

LEMMA 1.1 *Consider a family of functions $\mathcal{J}_t(\theta) : \Theta \to R$ and let $\hat{\theta}_t := \arg\min_{\theta\in\Theta} \mathcal{J}_t(\theta)$ (we assume that such a minimizer exists. If not, the theorem still holds with obvious changes; if more than one minimizer occurs, $\hat{\theta}_t$ can be any one of them.) If $\mathcal{J}_t(\theta)$ tends to a deterministic limit $\bar{\mathcal{J}}(\theta)$ uniformly in θ, i.e.*

$$\sup_{\theta\in\Theta} \left|\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right| \to 0, \tag{1.6}$$

then:

i) $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$;
ii) in i), Θ can be any set. Assume now that $\Theta \subset R^p$ be compact and that $\bar{J}(\cdot)$ is continuous over Θ. Then, letting $\Theta^* := \{\theta \in \Theta \; minimizing \; \bar{\mathcal{J}}(\theta)\}$,

$$\hat{\theta}_t \to \Theta^*, \tag{1.7}$$

in the sense that $\inf_{\theta\in\Theta^*} \|\hat{\theta}_t - \theta\| \to 0$.

PROOF. **i)** We have:

$$\begin{aligned}
\bar{\mathcal{J}}(\hat{\theta}_t) &= \mathcal{J}_t(\hat{\theta}_t) + \left(\bar{\mathcal{J}}(\hat{\theta}_t) - \mathcal{J}_t(\hat{\theta}_t)\right) && (1.8)\\
&\leq \mathcal{J}_t(\hat{\theta}_t) + \sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right| && (1.9)\\
&= \inf_{\theta\in\Theta} \mathcal{J}_t(\theta) + \sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right| && (1.10)\\
&= \inf_{\theta\in\Theta} \left[\bar{\mathcal{J}}(\theta) + \left(\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right)\right] + \sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right| && (1.11)\\
&\leq \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) + 2\sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right| && (1.12)\\
&\to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta). && (1.13)
\end{aligned}$$

ii) Suppose that the claim is false. From $\hat{\theta}_t$ extract a subsequence that is ϵ apart from Θ^*, for some $\epsilon > 0$, and from this sequence extract a convergent subsequence $\hat{\theta}_{t_k}$, $k = 1, 2, \ldots$ (such a subsequence exists since Θ is compact.) Thus, $\hat{\theta}_{t_k} \to \hat{\theta}_\infty \notin \Theta^*$. By continuity of $\bar{\mathcal{J}}$, $\bar{\mathcal{J}}(\hat{\theta}_{t_k}) \to \bar{\mathcal{J}}(\hat{\theta}_\infty) > \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$, but this contradicts i). □

The following exercise shows that the conditions in part ii) of the lemma cannot be relaxed.

EXERCISE 1.2 *Work out situations where either the continuity of $\bar{\mathcal{J}}$ or the compacteness of Θ is not met, and result ii) in Lemma 1.1 fails to hold.*

[HINT: As for continuity, let $\Theta = [0, 1]$ and consider

$$\bar{\mathcal{J}}(\theta) = \begin{cases} 1, & \theta = 0 \\ 0, & \theta = 1 \\ \theta, & \theta \in (0, 1), \end{cases} \tag{1.14}$$

so that $\Theta^* = \{1\}$. Construct a $\mathcal{J}_t(\theta)$ converging to $\bar{\mathcal{J}}(\theta)$ uniformly such that $\hat{\theta}_t \not\to 1$. As for compactness, let $\Theta = [0, \infty)$ and consider $\bar{\mathcal{J}}(\theta) = \theta/(\theta^2 + 1)$, so that $\Theta^* = \{0\}$. Again, construct a $\mathcal{J}_t(\theta)$ converging to $\bar{\mathcal{J}}(\theta)$ uniformly such that $\hat{\theta}_t \not\to 0$.] □

The key-property in (1.6) is that convergence takes place uniformly in θ. Should the convergence be nonuniform, the theorem would be false in general.

1.3 Linear predictors and quadratic cost

In this section, we are concerned with linear predictors. The following example serves to illustrate the type of questions we want to address.

EXAMPLE 1.3 *Consider the system depicted in Figure 1.2, where u and w are white Gaussian processes, independent to each other, and the true parameter value is $a^\circ = -0.5$, $b^\circ = 1$.*

We identify the system by means of two different predictors.

CASE 1)

Consider the predictor

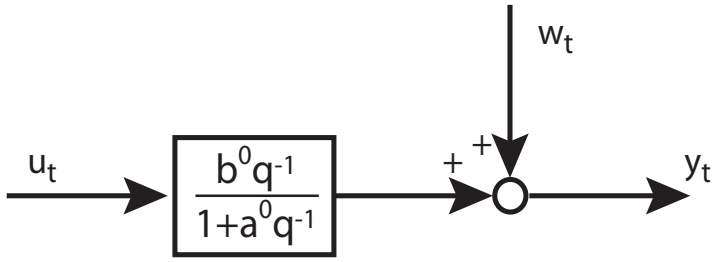

Figure 1.2: Data generation mechanism.

$$\hat{y}_t(\theta) = \frac{bz^{-1}}{1 + az^{-1}} u_t. \tag{1.15}$$

with $|a| < 1$. *This is in fact the predictor associated to the true system, that has the Output Error structure. Parameters* a *and* b *are estimated by minimizing the cost*

$$\frac{1}{t} \sum_{k=1}^{t} \left[y_k - \hat{y}_k(\theta) \right]^2. \tag{1.16}$$

Note that minimizing (1.16) involves an iterative procedure since this cost is not quadratic in θ.

CASE 2)

Since the true system can also be written as

$$y_t = a^{\circ} y_{t-1} + b^{\circ} u_{t-1} + term\ of\ disturbance, \tag{1.17}$$

we can try to use

$$\hat{y}_t(\theta) = a y_{t-1} + b u_{t-1} \tag{1.18}$$

as a predictor class which has the advantage that leads to a quadratic minimization. This is the predictor associated to an ARX model.

In Figures 1.3 and 1.4, the estimates are represented as a function of t.

Observing these figures, we can now make the following comments:

– the estimates seem to converge as $t \to \infty$;
– while the estimates convergence to the true value in the OE case, this seems not to happen in the ARX case.

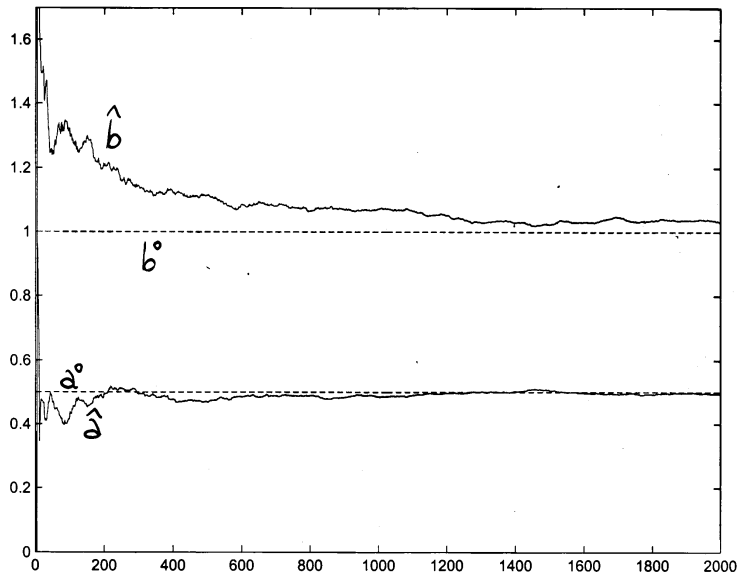

Figure 1.3: Case 1 (OE-predictor)

Note also that we can make this second observation because we are dealing with a simulation example where the "true system" is known. Should we be considering a real identification problem, the true system would not be available and the bias in the estimate could not be directly observed. Then, the question is: is there a theory that might have told us beforehand that, using an ARX predictor, we would have obtained a biased model? This theory does exist and it is the asymptotic theory studied in this section. □

The rest of this section is structured as follows. Our mathematical setting is introduced first and, then, the main convergence result is stated as Theorem 1.6. Theorem 1.6 can be generalized in may ways, as discussed after the theorem. The proof of Theorem 1.6 is delayed until Subsection 1.3.1.

The setting

Consider a predictor class of the form

$$\hat{y}_t(\theta) = W_u(z^{-1}, \theta)u_t + W_y(z^{-1}, \theta)y_t, \tag{1.19}$$

where transfer functions W_u and W_y satisfies the following assumption.

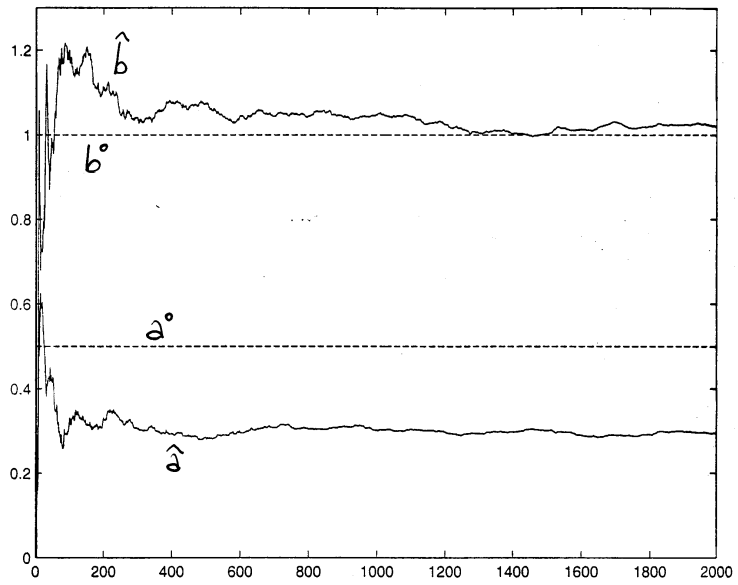

Figure 1.4: Case 2 (ARX-predictor)

ASSUMPTION 1.4 (uniform stability and continuity of the predictor) *$W_u(z^{-1},\theta)$ and $W_y(z^{-1},\theta)$ are rational transfer functions whose coefficients are continuous functions of the parameter $\theta \in \Theta \subset R^p$, Θ compact. Moreover, $W_u(z^{-1},\theta)$ and $W_y(z^{-1},\theta)$ are asymptotically stable, $\forall\theta \in \Theta$.* □

Processes u and y are generated according to the following scheme.

ASSUMPTION 1.5 (data generation mechanism) *Processes u and y are given by*

$$u_t = G_u(z^{-1})w_t \tag{1.20}$$

$$y_t = G_y(z^{-1})w_t, \tag{1.21}$$

where $G_u(z^{-1})$ and $G_y(z^{-1})$ are asymptotically stable rational transfer functions and w is a zero mean independent process with constant variance and such that $\sup_t E[w_t^4] < \infty$. □

According to Assumption 1.5, processes u and y are obtained by filtering a white process w. The fact that we consider one single remote process w is in order to keep the notations as compact as possible and considering more than one remote process does not change the analysis.

The introduced framework can for instance accomodate the situations depicted in Figures 1.5 and 1.6. Figure 1.5 represents an open-loop configuration where $G(z^{-1})$ has to be interpreted as the system and $F(z^{-1})$ is only introduced for the purpose of describing the correlation pattern in the u signal. Here, $G_u(z^{-1}) = F(z^{-1})$ and $G_y(z^{-1}) = G(z^{-1})F(z^{-1})$, which we assume to be stable transfer functions. Similar considerations apply to the feedback interconnnection of Figure 1.6.

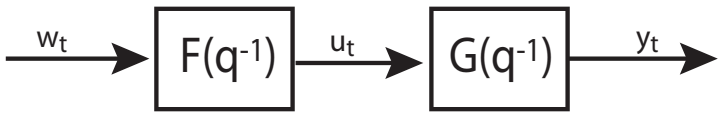

Figure 1.5: An open-loop data generation mechanism.

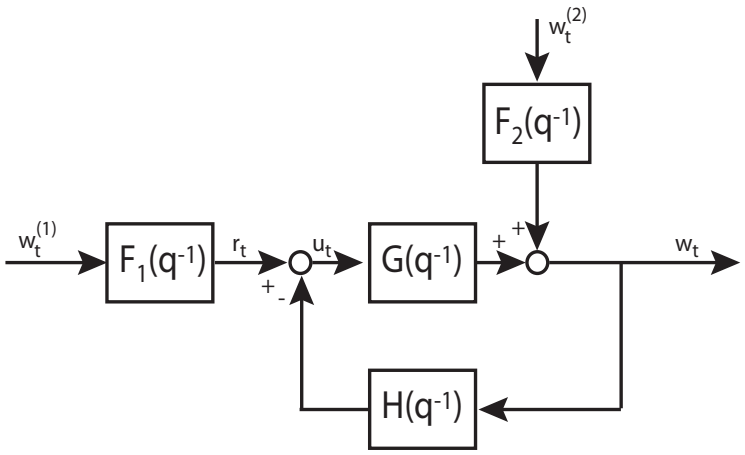

Figure 1.6: A closed-loop data generation mechanism.

More general data generation mechanisms are considered in the generalizations after Theorem 1.6.

Parameter θ is estimated by minimizing a quadratic criterion:

$$\hat{\theta}_t = \arg\min_{\theta \in \Theta} \mathcal{J}_t(\theta), \tag{1.22}$$

where

$$\mathcal{J}_t(\theta) := \frac{1}{t}\sum_{k=1}^{t}\left[y_k - \hat{y}_k(\theta)\right]^2. \tag{1.23}$$

The convergence result

We have now the following fundamental convergence result.

THEOREM 1.6 (Convergence of PEM methods) *Suppose that Assumptions 1.4 and 1.5 hold. Then, letting* $\bar{\mathcal{J}}(\theta) := E\left[(y_t - \hat{y}_t(\theta))^2\right]$ *and* $\Theta^* := \{\theta \in \Theta \text{ minimizing } \bar{\mathcal{J}}(\theta)\}$, *we have*

i) $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta}\bar{\mathcal{J}}(\theta)$, *almost surely;*
ii) $\hat{\theta}_t \to \Theta^*$, *almost surely, in the sense that* $\inf_{\theta\in\Theta^*}\|\hat{\theta}_t - \theta\| \to 0$ *almost surely.*

PROOF. See Section 1.3.1. □

A comment on the results of Theorem 1.6

Theorem 1.6 states two results i) and ii) which have very different interpretations.

Convergence i) pertains to the predictive capabilities of the identified predictor and says that they are asymptotically optimal. Point ii) deals with the parameter convergence. In certain problems, θ has a physical interpretation and the predictor structure is even more important than its predictive capabilities; in these cases, the convergence properties of $\hat{\theta}_t$ are of special importance. To put the result ii) under the correct light, however, it has to be mentioned that it does not pertain to the convergence of $\hat{\theta}_t$ to some "true parameter" θ° (provided that it exists.) In fact, the limiting point is characterized only in terms of providing optimal prediction and θ^* can as well be different from θ°.

Generalizations

The setting of Theorem 1.6 can be generalized in many directions.

1. The stationarity assumption can be relaxed to the so-called quasi-stationarity condition where $\bar{\mathcal{J}}$ is defined by taking time averaging besides ensamble averaging. This extension permits to consider deterministic signals (which may make much sense in certain cases, think of the reference signal in a feedback interconnection) and has been adopted e.g. in [15]);
2. the fact that all transfer functions are rational can be relaxed provided that a uniform stability condition is introduced (see the derivations of Section 1.3.1);
3. Assumptions 1.4 and 1.5 are twofold: i) the stability assumption guarantees that data samples far apart are little correlated. In this way, a law of large number applies; ii) assuming linearity constrains the problem so that the law of large numbers

applies uniformly in θ. Both these two facts are essential to prove that the uniform condition (1.6) in Lemma 1.1 holds true. However, facts i) and ii) are also met in more general settings allowing for nonlinear and/or time-varying models and data generation mechanisms and for nonquadratic cost functions. Early references in this directions are [8], [14], and [4]. The reader is also referred to Section 1.4 for a study in a nonlinear setting;

4. the theory holds unaltered if data u and y are filterd by a stable tranfer function $L(z^{-1})$ before they are processed in the identification algorithm. In this case the identification cost writes

$$\frac{1}{t}\sum_{k=1}^{t}\left[L(z^{-1})y_k - L(z^{-1})\hat{y}_k(\theta)\right]^2, \tag{1.24}$$

and, under the usual assumptions, $\hat{\theta}_t$ asymptotically minimizes the cost $E\left[\left(L(z^{-1})y_t - L(z^{-1})\hat{y}_t(\theta)\right)^2\right]$.

1.3.1 Proof of the convergence theorem

Proof of Theorem 1.6

The theorem is proven by resorting to Lemma 1.1. To this purpose, let

$$V_r^p := \sup_{\theta\in\Theta}\left|\sum_{k=r}^{p}\left(\epsilon_k(\theta)^2 - E\left[\epsilon_k(\theta)^2\right]\right)\right| \tag{1.25}$$

where $\epsilon_t(\theta) := y_t - \hat{y}_t(\theta)$, and note that $\frac{1}{t}V_1^t = \sup_{\theta\in\Theta}\left|\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right|$. Signal $\epsilon_t(\theta)$ is generated according to the scheme

$$\begin{aligned}\epsilon_t(\theta) &= y_t - W_u(z^{-1},\theta)u_t - W_y(z^{-1},\theta)y_t &(1.26)\\ &= G_y(z^{-1})w_t - W_u(z^{-1},\theta)G_u(z^{-1})w_t - W_y(z^{-1},\theta)G_y(z^{-1})w_t, &(1.27)\end{aligned}$$

and it is easy to verify that $\epsilon_t(\theta)$ satisfies all assumptions imposed to $z_t(\theta)$ and $v_t(\theta)$ in the technical Lemma 1.7 given below. Thus, letting $z_t(\theta) = \epsilon_t(\theta)$, Lemma 1.7 yields

$$\begin{aligned}E\left[(V_r^p)^2\right] &= E\left[\left(\sup_{\theta\in\Theta}\left|\sum_{k=r}^{p}\left(\epsilon_k(\theta)^2 - E\left[\epsilon_k(\theta)^2\right]\right)\right|\right)^2\right] &(1.28)\\ &= E\left[\sup_{\theta\in\Theta}\left(\sum_{k=r}^{p}\left(\epsilon_k(\theta)^2 - E\left[\epsilon_k(\theta)^2\right]\right)\right)^2\right] &(1.29)\\ &\leq C(p+1-r), &(1.30)\end{aligned}$$

where C is a suitable constant. In addition, it is immediately verified that

$$|V_1^t| \leq |V_1^m| + |V_{m+1}^t| \tag{1.31}$$

for $m < t$. Having proved (1.30) and (1.31), we can now appeal to Lemma A.14 in Appendix A to conclude that

$$\frac{1}{t}V_1^t = \sup_{\theta\in\Theta} \left|\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right| \to 0, \quad almost\ surely, \tag{1.32}$$

which is condition (1.6) in Lemma 1.1.

Next, we show that $\bar{\mathcal{J}}(\theta)$ is a continuous function of θ as required in point ii) of Lemma 1.1.

To this purpose, it suffices to show that $\hat{y}_t(\theta) \to \hat{y}(t,\bar{\theta})$ in L^2 when $\theta \to \bar{\theta}$. Indeed, considering the expression of $\bar{\mathcal{J}}(\theta) = E[(y_t - \hat{y}_t(\theta))^2]$, it is then immediate to conclude that $\bar{\mathcal{J}}(\theta) \to \bar{\mathcal{J}}(\bar{\theta})$ by use of the triangular inequality in L^2.

In order to prove that $\hat{y}_t(\theta) \to \hat{y}(t,\bar{\theta})$ in L^2, start by recalling that $\hat{y}_t(\theta) = W_u(z^{-1},\theta)u_t + W_y(z^{-1},\theta)y_t$. Thus, we need to show that $W_u(z^{-1},\theta)u_t \to W_u(z^{-1},\bar{\theta})u_t$ in L^2 and $W_y(z^{-1},\theta)y_t \to W_y(z^{-1},\bar{\theta})y_t$ in L^2. We only show the first convergence, the other one being entirely analogous. From the continuity in θ of the coefficients of $W_u(z^{-1},\theta)$ we have:

$$\sup_{\omega} \left|W_u(e^{-i\omega},\theta) - W_u(e^{-i\omega},\bar{\theta})\right| \to 0, \quad as\ \theta \to \bar{\theta}. \tag{1.33}$$

Then, conclusion $W_u(z^{-1},\theta)u_t \to W_u(z^{-1},\bar{\theta})u_t$ in L^2 follows by recalling that the L^2-norm of $W_u(z^{-1},\theta)u_t - W_u(z^{-1},\bar{\theta})u_t$ is the integral of its spectral density and that the spectral density is given by $\left|W_u(e^{-i\omega},\theta) - W_u(e^{-i\omega},\bar{\theta})\right|^2 f^{(u)}(\omega)$.

Thus, all assumptions in Lemma 1.1 are met and the thesis now readily follows from the Lemma. □

LEMMA 1.7 *Consider the process*

$$z_t(\theta) = \frac{B(z^{-1},\theta)}{A(z^{-1},\theta)}w_t = \sum_{k=0}^{\infty}\alpha_k(\theta)w(t-k), \tag{1.34}$$

where $B(z^{-1},\theta)$ and $A(z^{-1},\theta)$ are polynomials whose coefficients are continuous functions of a parameter $\theta \in \Theta \subset R^p$, Θ compact, $A(z^{-1},\theta)$ is asymptotically stable, $\forall\theta \in \Theta$, and w is an zero mean independent process such thak $C_w := \sup_t E[w_t^4] < \infty$. Then,

$$E\left[\sup_{\theta\in\Theta}\left(\sum_{k=r}^{p}\left(z_k(\theta)^2 - E[z_k(\theta)^2]\right)\right)^2\right] \leq 4C^4C_w(p+1-r) \tag{1.35}$$

where $C := \sup_{\theta\in\Theta}\sum_{k=0}^{\infty}|\alpha_k(\theta)|$.

PROOF. First of all, we note that, being Θ compact, $|\alpha_k(\theta)| := \alpha_k \leq H\rho^k, \forall\theta\in\Theta$, for some H and $\rho < 1$ (in fact this claim requires a bit of care. See e.g. [20] for a precise study of this issue), so that $C < \infty$.

Define

$$\gamma(k,i,j) := w(k-i)w(k-j) - E[w(k-i)w(k-j)]; \tag{1.36}$$

$$\eta(i,j) := \sum_{k=r}^{p}\gamma(k,i,j). \tag{1.37}$$

The first step in the proof consists in bounding $E[\eta(i,j)^2]$. To this end, note first that, by use of the Schwarz inequality in L^2, we have

$$\begin{aligned} E\left[\gamma(k,i,j)^2\right] &= E\left[(w(k-i)w(k-j) - E[w(k-i)w(k-j)])^2\right] && (1.38)\\ &\leq E\left[w(k-i)^2w(k-j)^2\right] && (1.39)\\ &\leq E\left[w(k-i)^4\right]^{\frac{1}{2}}E\left[w(k-j)^4\right]^{\frac{1}{2}} && (1.40)\\ &\leq C_w, && (1.41)\end{aligned}$$

from which, using again the Schwarz inequality,

$$\begin{aligned} E[\gamma(k_1,i,j)\gamma(k_2,i,j)] &\leq E\left[\gamma(k_1,i,j)^2\right]^{\frac{1}{2}}E\left[\gamma(k_2,i,j)^2\right]^{\frac{1}{2}} && (1.42)\\ &\leq C_w. && (1.43)\end{aligned}$$

Then, write

$$\begin{aligned} E\left[\eta(i,j)^2\right] &= E\left[\left(\sum_{k_1=r}^{p}\gamma(k_1,i,j)\right)\left(\sum_{k_2=r}^{p}\gamma(k_2,i,j)\right)\right] && (1.44)\\ &= \sum_{k_1=r}^{p}\sum_{k_2=r}^{p}E\left[\gamma(k_1,i,j)\gamma(k_2,i,j)\right], && (1.45)\end{aligned}$$

and note that most expectations in (1.45) are null. Precisely, a simple inspection shows that $E\left[\gamma(k_1,i,j)\gamma(k_2,i,j)\right]=0$ unless at least one of the two instant points k_1-i and k_1-j equals one of the other two instant points k_2-i, and k_2-j. Thus, for given k_1,i,j, there are at most 4 values of k_2 such that the expectation is nonzero. If we bound these 4 expectations by means of (1.42), we then find

$$\begin{aligned} E\left[\eta(i,j)^2\right] &\leq \sum_{k_1=r}^{p} 4C_w && (1.46)\\ &= 4C_w(p+1-r), && (1.47)\end{aligned}$$

which is the desired bound for $E\left[\eta(i,j)^2\right]$.

Finally, we have

$$\begin{aligned} & E\left[\sup_{\theta\in\Theta}\left(\sum_{k=r}^{p}\left(z_k(\theta)^2-E[z_k(\theta)^2]\right)\right)^2\right] && (1.48)\\ &= E\left[\sup_{\theta\in\Theta}\left(\sum_{k=r}^{p}\left(\left(\sum_{i=0}^{\infty}\alpha_i(\theta)w(k-i)\right)^2-E\left[\left(\sum_{i=0}^{\infty}\alpha_i(\theta)w(k-i)\right)^2\right]\right)\right)^2\right] && (1.49)\\ &= E\left[\sup_{\theta\in\Theta}\left(\sum_{i=0}^{\infty}\sum_{j=0}^{\infty}\alpha_i(\theta)\alpha_j(\theta)\eta(i,j)\right)^2\right] && (1.50)\\ &= E\left[\sup_{\theta\in\Theta}\sum_{i_1=0}^{\infty}\sum_{j_1=0}^{\infty}\sum_{i_2=0}^{\infty}\sum_{j_2=0}^{\infty}\alpha_{i_1}(\theta)\alpha_{j_1}(\theta)\alpha_{i_2}(\theta)\alpha_{j_2}(\theta)\eta(i_1,j_1)\eta(i_2,j_2)\right] && (1.51)\\ &\leq \sum_{i_1=0}^{\infty}\sum_{j_1=0}^{\infty}\sum_{i_2=0}^{\infty}\sum_{j_2=0}^{\infty}\alpha_{i_1}\alpha_{j_1}\alpha_{i_2}\alpha_{j_2}E[|\eta(i_1,j_1)\eta(i_2,j_2)|] && (1.52)\\ &\leq \sum_{i_1=0}^{\infty}\sum_{j_1=0}^{\infty}\sum_{i_2=0}^{\infty}\sum_{j_2=0}^{\infty}\alpha_{i_1}\alpha_{j_1}\alpha_{i_2}\alpha_{j_2}E[\eta(i_1,j_1)^2]^{\frac{1}{2}}E[\eta(i_2,j_2)^2]^{\frac{1}{2}} && (1.53)\\ &\leq C^4 4C_w(p+1-r), && (1.54)\end{aligned}$$

where (1.46) has been used in the last step. This concludes the proof of the lemma. □

1.4 Nonlinear predictors

In this section, we consider again the problem of estimating a parameter θ by minimizing a quadratic criterion of the form (1.23), viz.

$$\mathcal{J}_t(\theta) := \frac{1}{t}\sum_{k=1}^{t}\left[y_k - \hat{y}_k(\theta)\right]^2, \tag{1.55}$$

and ask whether the convergence $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta}\bar{\mathcal{J}}(\theta)$ takes place, where $\hat{\theta}_t$ is the minimizer of $\mathcal{J}_t(\theta)$ and $\bar{\mathcal{J}}(\theta) := E[\mathcal{J}_t(\theta)]$. The novel element with respect to the previous section is that we no longer require here that the predictor be a linear system.

Our main tool for proving convergence is the fundamental Lemma 1.1. This lemma rests on the assumption that the function sequence $\mathcal{J}_t(\theta)$ converges to $\bar{\mathcal{J}}(\theta)$ uniformly in θ, and, indeed, the uniformity of the convergence is crucial for the validity of the lemma's results.

In the previous section, Lemma 1.1 has been applied to a linear context. There, two facts played a fundamental role in proving that $\mathcal{J}_t(\theta) \to \bar{\mathcal{J}}(\theta)$ uniformly. The first fact is that the predictor was stable. This entails that data samples far apart are little correlated, so that a law of large number applies. The second fact was linearity of the predictor. Due to linearity, the law of large numbers applies uniformly in θ.

If we now turn to a more general context allowing for nonlinear predictors, a new problem pops up. In fact, function $\mathcal{J}_t(\theta)$ may now be a fairly complex function of θ since linearity is no longer there to moderate its complexity. As a consequence, $\mathcal{J}_t(\theta)$ still converges to $\bar{\mathcal{J}}(\theta)$ for a fixed θ, but uniform convergence in θ may get lost in general. In turn, this makes the results of Lemma 1.1 false. The main objective of this section is to discuss under what conditions uniform convergence of $\mathcal{J}_t(\theta)$ is preserved in a nonlinear context so that asymptotic convergence continues to hold.

The section is structured as follows. In the next subsection, examples are provided showing where the trouble may come from in a nonlinear context. Subsection 1.4.2 presents the Vapnik-Chervonenkis theory. This theory gives conditions for the uniform convergence of $\mathcal{J}_t(\theta)$ to $\bar{\mathcal{J}}(\theta)$ and thereby, by virtue of Lemma 1.1, for asymptotic convergence of $\bar{\mathcal{J}}(\hat{\theta}_t)$ to $\inf_{\theta\in\Theta}\bar{\mathcal{J}}(\theta)$. In these two subsections we are concerned with a special setting where the system output is binary (either 0 or 1) and no noise is present. Though very specific, this setting is of great importance in certain applications. The case of outputs taking value in a continuum is discussed in the last subsection. Throughout, we concentrate on the case when the system input forms an independent sequence, as a theory for the dependent case is still lacking.

1.4.1 Preliminary examples

We start by providing an example of a nonlinear context where convergence takes place.

EXAMPLE 1.8 (A nonlinear context where convergence takes place) *A system is fed by a real input u_t and outputs either 0 or 1 according to an unknown deterministic rule. Precisely, letting A° be a measurable set on the real line, we have:*

$$y_t = \begin{cases} 1, & if\ u_t \in A^\circ, \\ 0, & otherwise. \end{cases} \tag{1.56}$$

The system input u_t is an independent process that takes on value on the real line according to some probability distribution μ_U.

The predictor is

$$\hat{y}_t(\theta) = \begin{cases} 1, & if\ u_t \in (a,b], \\ 0, & otherwise. \end{cases} \tag{1.57}$$

and is parameterized by $\theta = [a\ b]^T$ (the fact that the interval $(a,b]$ is open to the left and closed to the right is only for convenience and has no deep significance.)

The quadratic criterion used to estimate θ in this case writes:

$$\mathcal{J}_t(\theta) := \frac{1}{t}\sum_{k=1}^{t} [y_k - \hat{y}_k(\theta)]^2 = \frac{1}{t}\sum_{k=1}^{t} 1\left(y_k \neq \hat{y}_k(\theta)\right). \tag{1.58}$$

We want to prove that

$$\sup_{\theta \in R^2} \left| \mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta) \right| \to 0, \quad almost\ surely, \tag{1.59}$$

where $\bar{\mathcal{J}}(\theta) := E\left[\frac{1}{t}\sum_{k=1}^{t} 1\left(y_k \neq \hat{y}_k(\theta)\right)\right] = E[1\left(y_t \neq \hat{y}_t(\theta)\right)]$, from which, thanks to Lemma 1.1, we have that

$$\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta \in R^2} \bar{\mathcal{J}}(\theta), \quad almost\ surely. \tag{1.60}$$

(1.60) has to be interpreted that, in the long run, the selected predictor gives the smallest possible probability of misclassifying an unseen output y.

Convergence (1.59) is a direct consequence of a classical result in probability theory known as Glivenko-Cantelli convergence theorem. We state and prove this theorem before proceeding. □

Remark 1.9 *The reader may have noticed that in (1.60) the focus is on the convergence of the $\bar{\mathcal{J}}$ cost (and not the convergence of $\hat{\theta}_t$.) The reason is that in the nonlinear context of the present section it is often the case that what really matters are the predictive capabilities of the model while parameter θ plays a marginal, mostly instrumental, role.* □

THEOREM 1.10 (Glivenko-Cantelli theorem) *Consider an independent sequence of random variables $\{v_t\}$ with common probability distribution function F. Let*

$$\hat{F}_t(x) = \frac{1}{t}\sum_{k=1}^{t} 1(v_k \leq x), \tag{1.61}$$

where $1(v_k \leq x) = 1$ if $v_k \leq x$ and 0 otherwise, be the corresponding empirical distribution function. Then,

$$\sup_{-\infty<x<\infty} \left|\hat{F}_t(x) - F(x)\right| \to 0, \quad \textit{almost surely.} \tag{1.62}$$

PROOF. Given $\epsilon > 0$, define $x_1, x_2, \ldots$ as the real numbers such that $F(x_j) \geq j\frac{\epsilon}{3}$ and $F(x) < j\frac{\epsilon}{3}$ for $x < x_j$ (see Figure 1.7.)

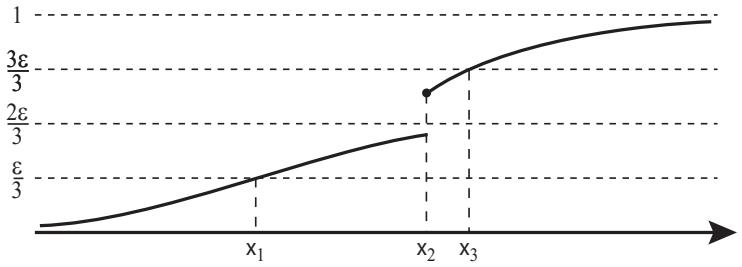

Figure 1.7: Construction of the x_k's in the proof of the Glivenko-Cantelli theorem

For any fixed x_k, the law of large numbers (see Theorem A.13 in Appendix A) yields

$$\left|\hat{F}_t(x_j) - F(x_j)\right| = \left|\frac{1}{t}\sum_{k=1}^{t} 1(v_k \leq x_j) - E[1(v_k \leq x_j)]\right| \to 0, \quad \textit{almost surely;} \tag{1.63}$$

moreover, if F is discountinuous in correspondence of x_j, by letting $\Delta F(x_j)$ be the jump in F (and similarly for $\Delta\hat{F}_t(x_j)$), we also have

$$\left|\Delta\hat{F}_t(x_j) - \Delta F(x_j)\right| = \left|\frac{1}{t}\sum_{k=1}^{t} 1(v_k = x_j) - E[1(v_k = x_j)]\right| \to 0, \quad almost\ surely. \tag{1.64}$$

Observing now that the points $x_1, x_2, \ldots$ are only finitely many, from (1.63) and (1.64) we can conclude that, with probability one, the inequalities

$$\left|\hat{F}_t(x_j) - F(x_j)\right| \leq \frac{\epsilon}{3} \tag{1.65}$$

and

$$\left|\Delta\hat{F}_t(x_j) - \Delta F(x_j)\right| \leq \frac{\epsilon}{3} \tag{1.66}$$

are satisfied simultaneously for all x_j when t is larger than a certain $\bar{t}$ ($\bar{t}$ depends on the probabilistic outcome, i.e. it is a random variable, but this is unimportant here.)

Consider now any real number x and the interval $[x_{j-1}, x_j)$ such that $x \in [x_{j-1}, x_j)$ (when $x < x_1$ or $x >$ largest x_j, the derivation is similar, though the notations are slightly different, and is not reported here in detail.) For $t > \bar{t}$ we then have that:

$$\begin{aligned}
\hat{F}_t(x) &\geq \hat{F}_t(x_{j-1}) && (1.67)\\
&\geq F(x_{j-1}) - \frac{\epsilon}{3} \quad (using\ (1.65)) && (1.68)\\
&\geq F(x) - \frac{2}{3}\epsilon \quad (because\ of\ how\ the\ x_j's\ have\ been\ selected); && (1.69)
\end{aligned}$$

and that

$$\begin{aligned}
\hat{F}_t(x) &\leq \hat{F}_t(x_j^-) \quad (where\ \hat{F}_t(x_j^-) = \lim_{x \to x_j} \hat{F}_t(x)\ from\ the\ left) && (1.70)\\
&= \hat{F}_t(x_j) - \Delta\hat{F}_t(x_j) && (1.71)\\
&\leq F(x_j) + \frac{\epsilon}{3} - \Delta F(x_j) + \frac{\epsilon}{3} \quad (using\ (1.65)\ and\ (1.66)) && (1.72)\\
&= F(x_j^-) + \frac{2}{3}\epsilon && (1.73)\\
&\leq F(x) + \epsilon \quad (because\ of\ how\ the\ x_j's\ have\ been\ selected), && (1.74)
\end{aligned}$$

showing that

$$\sup_{-\infty<x<\infty} \left| \hat{F}_t(x) - F(x) \right| \leq \epsilon \tag{1.75}$$

for $t > \bar{t}$. Due to the arbitrariness of ϵ, this implies the thesis of the theorem. □

EXAMPLE 1.11 (Example 1.8 continued) *(1.59) is now proven by resorting to the Glivenko-Cantelli theorem.*

As a matter of fact, we are going to use a simple generalization of the Glivenko-Cantelli theorem, that we state first. As in Theorem 1.10, consider a sequence of independent random variables $\{v_t\}$. Given a Borel set $B \subset R$ define $F(x) := \mu_U(B \cap (-\infty, x])$ and $\hat{F}_t(x) = \frac{1}{t}\sum_{k=1}^{t} 1(\{v_k \in B \cap (\infty, x])$. The interpretation here is that we only count the x points that lie in the set B, while disregarding the others. Then, (1.62) holds with the new definitions of F and $\hat{F}$. The proof of this fact is the same as the proof of the Glivenko-Cantelli theorem.

Now, rewrite $\mathcal{J}_t(\theta)$ as follows (^{c} stands for complement):

$$\begin{aligned} \mathcal{J}_t(\theta) &= \frac{1}{t}\sum_{k=1}^{t} 1\left(y_k \neq \hat{y}_k(\theta)\right) && (1.76)\\ &= \frac{1}{t}\sum_{k=1}^{t} 1\left(u_k \in (A^\circ)^c \cap (a,b]\right) + \frac{1}{t}\sum_{k=1}^{t} 1\left(u_k \in A^\circ \cap (a,b]^c\right). && (1.77) \end{aligned}$$

We show how to handle the first term, the second can be treated similarly. We have:

$$\begin{aligned} &\frac{1}{t}\sum_{k=1}^{t} 1\left(u_k \in (A^\circ)^c \cap (a,b]\right) && (1.78)\\ &= \frac{1}{t}\sum_{k=1}^{t} 1\left(u_k \in (A^\circ)^c \cap (-\infty,b]\right) - \frac{1}{t}\sum_{k=1}^{t} 1\left(u_k \in (A^\circ)^c \cap (-\infty,a]\right) && (1.79) \end{aligned}$$

According to the extended Glivenko-Cantelli theorem, the first term in this last expression tends uniformly in b to $\mu((A^\circ)^c \cap (-\infty, b])$, while the second one tends uniformly in a to $\mu((A^\circ)^c \cap (-\infty, a])$. Hence, the entire expression tends uniformly in a and b to $\mu((A^\circ)^c \cap (a,b])$. Likewise, the second term in (1.77) converges uniformly in a and b to $\mu_U(A^\circ \cap (a,b]^c)$. Thus, from (1.77) we conclude that

$$\mathcal{J}_t(\theta) \to \mu((A^\circ)^c \cap (a,b]) + \mu(A^\circ \cap (a,b]^c) = \bar{\mathcal{J}}(\theta), \tag{1.80}$$

uniformly in a and b, that is (1.59). □

In the previous example, we have seen that uniform convergence of $\mathcal{J}_t(\theta)$ to $\bar{\mathcal{J}}(\theta)$ takes place when θ parameterizes intervals on the real line. In turn, this entails that selecting an interval by minimizing the sample identification criterion permits one to push the probability of misclassification as close as desired to the minimal probability of misclassification.

The reason for the above result is that the class of candidate classifiers (the intervals) is quite restricted (and, indeed, this is the the reason for the convergence $\mathcal{J}_t(\theta) \to \bar{\mathcal{J}}(\theta)$ to hold uniformly in θ) so that, based on data, the predictive capabilities of all classifiers can be simultaneously estimated. Then, maximizing the estimated predictive capability corresponds to maximizing the true predictive capability. It should also be noted that A° can instead be any complex: nature can be complicated, the way we describe it has to be simple.

For more general classes of classifiers, it may be impossible to simultaneously estimate from data the predictive capabilities of all the classifiers (technically, uniform convergence is lost). Thus, no matter how large the sample size is, there can be a classifier with poor predictive capability which is ranked at the highest reliability level by the sample criterion. This leads to inability of performing a proper selection of the classifier.

The bottom-line of this discussion is that we do not have to push our luck too much: larger classes of classifiers lead to a better classification provided that we are able to select the best classifier within the class, but, since we only rely on partial information based on data, if the class is too large we can be unable to properly select the classifier. This issue is now illustrated by an example.

EXAMPLE 1.12 (A nonlinear context where convergence does not take place) *Consider again the same data generation mechanism as in Example 1.8, but this time we enlarge the predictor class by adding to the predictors associated to intervals all functions mapping u_t into $\{0,1\}$ such that 1 is achieved only in correspondance of a finite number of values of u_t. In formal terms, this second group of predictors contains all functions defined as*

$$\hat{y}_t(\theta) = \begin{cases} 1, & if\ u_t \in \theta, \\ 0, & otherwise, \end{cases} \tag{1.81}$$

where $\theta \subset R$ is formed by a finite number of isolated points. So, a predictor is either a map that returns 1 for u_t in a given interval $(a,b]$, or a map that gives 1 for all (finite) points in a set θ. It is clear that this class of predictors is fairly rich.

Now, given t data points $\{(u_k, y_k), k = 1, \ldots, t\}$, the predictor of the form (1.81) associated to the set θ that contains all and only the u_k points such that the corresponding

output $y_k = 1$ achieves an empirical identification cost of 0. *On the other hand, if the probability μ according to which points u_k's are extracted has no concentrated mass, the corresponding theoretical prediction error is*

$$E[\mathcal{J}_t(\theta)] = E[1(y_t \neq \hat{y}_t(\theta))] = \mu_U(A^\circ). \tag{1.82}$$

Thus, no uniform convergence of $\mathcal{J}_t(\theta)$ to $E[\mathcal{J}_t(\theta)]$ takes place here and, in fact, minimizing $\mathcal{J}_t(\theta)$ does not lead asymptotically to select the best predictor. □

1.4.2 The Vapnik-Chervonenkis theory

In this subsection we consider the problem of estimating nonlinear predictors for $\{0,1\}$-valued functions.

In the literature, such a problem is known under the names of pattern recognition, or pattern classification, or discrimination. It is about predicting the unknown nature (between two possibilities) of an observation, where one is given certain measurements (the u value) and is asked to predict the corresponding output (the unknown nature 0 or 1.) It has applications in a number of different contexts, where 0 and 1 may stand for sick or healthy, real or fake, correct or wrong, etc.. In pattern recognition, a predictor is also called a classifier since it classifies the u's in terms of two possible outcomes.

Considering $\{0,1\}$-valued functions is restrictive. However, simple models are easier to understand and amenable to an immediate interpretation. Moreover, the extension to functions taking value in a finite set is rather straigthforward. More general nonlinear settings are discussed in Section **??**.

In the previous subsection, we have seen by way of examples that the uniform convergence of $\mathcal{J}_t(\theta)$ to $\bar{\mathcal{J}}(\theta)$ takes place in certain cases but not in others. Moreover, failure of convergence leads to an inability to select the best predictor through the minimization of the sample identification criterion.

In this subsection, we derive general conditions for $\mathcal{J}_t(\theta) \to \bar{\mathcal{J}}(\theta)$ uniformly in θ for $\{0,1\}$-valued predictors. As expected, it turns out that the predictor class must be not too rich for this convergence to hold. The results provided here are of wide breath and comprises the Glivenko-Cantelli theorem as a particular case. Moreover, as we shall see, they are necessary and sufficient in a sense that we shall precisely state.

We commence by studying the uniform convergence of empirical probabilities to their statistical value in a general setting. Then, the results developed will be applied to the uniform convergence of $\mathcal{J}_t(\theta)$ to $\bar{\mathcal{J}}(\theta)$.

Uniform convergence of empirical probabilities

We take here a general point of view by considering measurable spaces $(E, \mathcal{E})$. However, for a more concrete understanding, just think of $(R^n, \mathcal{B}(R^n))$, or $(R, \mathcal{B}(R))$ as in the Glivenko-Cantelli theorem.

Given a measurable space $(E, \mathcal{E})$ endowed with a unknown probability measure μ, suppose that we are given t independent random extractions $e_1, \ldots, e_t$ out of the set E according to probability μ. Suppose also that, for a given set $A \in \mathcal{E}$, we are asked to provide an estimate of its probability $\mu(A)$.

A natural way to do so is to count how many times the random extractions $e_1, \ldots, e_t$ fall into A and divide this number by the total number of extractions. This is the so-called empirical mean of A:

$$\hat{\mu}_t(A) := \frac{1}{t} \sum_{k=1}^{t} 1(e_k \in A), \tag{1.83}$$

where $1(e_k \in A) = 1$ if $e_k \in A$ and 0 otherwise. Now, a natural question is: is it true that $\hat{\mu}_t(A) \to \mu(A)$ almost surely when $t \to \infty$? The answer is yes from a straightforward application of the law of large numbers (Theorem A.13 in Appendix A.)

Next, we make the problem more difficult by considering many sets A simultaneously. Specifically, let $\mathcal{A}$ be a collection of sets $A \in \mathcal{E}$. We ask whether

$$\sup_{A \in \mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| \to 0, \quad \textit{almost surely.} \tag{1.84}$$

Note also that this is exactly the problem solved by the Glivenko-Cantelli theorem in the particular case where $\mathcal{A} = \{(-\infty, x], x \in R\}$. Here, the uniform result (1.84) is studied in general.

When $\mathcal{A}$ contains only a finite number of elements, result (1.84) is trivially true. The problems becomes challenging when the cardinality of $\mathcal{A}$ is infinite. Intuitively, the uniform convergence (1.84) holds in this case provided that collection $\mathcal{A}$ is not too rich. The notion of richness of a collection of sets is formalized in the following definition.

DEFINITION 1.13 (shatter function and VC-dimension) *Given a collection $\mathcal{A}$ of subsets of E, we define:*

i) the shatter function of $\mathcal{A}$ as

$$S_{\mathcal{A}}(n) := \max_{e_1, \ldots, e_n \in E} \#\left\{\{e_1, \ldots, e_n\} \cap A; A \in \mathcal{A}\right\}, \tag{1.85}$$

(# denotes "number of elements"; the empty set is counted as a set.) Thus, $S_{\mathcal{A}}(n)$ is the maximal number of different subsets which can be obtained by intersecting a set of n points with elements of $\mathcal{A}$. Clearly, the shatter function is such that $S_{\mathcal{A}}(n) \leq 2^n$, $\forall n$);

ii) the Vapnik-Chervonenkis dimension (VC-dimension) $V_{\mathcal{A}}$ *is the largest* n *such that* $S_{\mathcal{A}}(n) = 2^n$. *If such a largest* n *does not exist we let* $V_{\mathcal{A}} = \infty$. □

When n points are such that we can form 2^n different subsets by intersecting them with elements of $\mathcal{A}$, we say that these points are shattered by $\mathcal{A}$. So, we can also say that the VC-dimension of $\mathcal{A}$ is the cardinality of largest set of points that is shatterd by $\mathcal{A}$.

The notion of VC-dimension is illustrated by an example.

EXAMPLE 1.14 (VC-dimension) *Consider the collection* $\mathcal{A}$ *of semi-infinite intervals* $\{(-\infty, x], x \in R\}$, *as in the Glivenko-Cantelli theorem. Given one point* x_1, *it is clear that we can take two sets* $(-\infty, x^{'}]$ *and* $(-\infty, x^{''}]$ *such that* $x_1 \notin (-\infty, x^{'}]$ *and* $x_1 \in (-\infty, x^{''}]$. *To this purpose, it is enough to select* $x^{'}$ *and* $x^{''}$ *such that* $x^{'} < x_1 < x^{''}$. *Thus, the set* $\{x_1\}$ *formed by only one point is shatterd by* $\{(-\infty, x], x \in R\}$.

On the other hand, it is easy to convince oneself that a set of two points $\{x_1, x_2\}$ *cannot be shattered by* $\mathcal{A}$. *The reason is that if* $x_1 < x_2$, *we cannot obtain the set* $\{x_2\}$ *by intersecting* $\{x_1, x_2\}$ *with a semi-infinite interval. From this, we see that* $V_{\mathcal{A}} = 1$ *in this case.*

Consider instead the collection $\mathcal{A}$ *of all sets formed by an arbitrary, though finite, number of points, as in Example 1.12. In this case, given a set of* n *points* $\{e_1, \ldots, e_n\}$, *no matter how large* n *is, it is immediate to conclude that it is shattered by* $\mathcal{A}$. *Thus,* $V_{\mathcal{A}} = \infty$. □

EXERCISE 1.15 *Show that the VC-dimension of half planes in* R^2 *is* 3. *Similarly, show that the VC-dimension of disks in* R^2 *is again* 3. □

The VC-dimension of $\mathcal{A}$ is a purely combinatorial measure and has nothing to do with probability.

We now derive a fundamental inequality that connects the shatter function of $\mathcal{A}$ with the probability of the set $\{\sup_{A \in \mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| > \epsilon\}$. This result is at the basis of the theory of uniform convergence of empirical means.

Before stating the inequality, for the sake of clarity we deem advisable to make an observation regarding the probability space where the different random variables are hosted. Up to now, we have spoken of random independent extractions out of the set $(E, \mathcal{E})$ according to a probability μ. We now better specify this by saying that $e_1, \ldots, e_t$ are nothing but independent random elements defined on some underlying probability space $(\Omega, \mathcal{F}, P)$ taking value in E. Then, μ is the image probability measure of P through one

e_k and it is the same for all e_k's. It is perhaps worth mentioning that an alternative point of view can be taken where all random variables are hosted in the product measurable space $(E^t, \mathcal{E}^t)$ with the probability μ^t.

THEOREM 1.16 (Vapnik-Chervonenkis inequality) *With all symbols as above, we have*

$$P\left\{\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| > \epsilon\right\} \leq 2S_{\mathcal{A}}(2t)e^{-2t\epsilon^2/5}. \tag{1.86}$$

PROOF. We start by showing that

$$E\left[\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|\right] \leq \sqrt{\frac{2\ln(2S_{\mathcal{A}}(2t))}{t}}. \tag{1.87}$$

Introduce an independent copy $e_1^{'}, \ldots, e_t^{'}$ of $e_1, \ldots, e_t$ (that is the copy has the same distribution as the original multisample) and t independent and identically distributed sign random variables $s_1, \ldots, s_t$ with $P(s_k = -1) = P(s_k = 1) = 0.5$, independent of $e_1, \ldots, e_t$ and $e_1^{'}, \ldots, e_t^{'}$. Then, we can write:

$$\begin{aligned}
&E\left[\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|\right] && (1.88)\\
&= E\left[\sup_{A\in\mathcal{A}} \left|E\left[\hat{\mu}_t(A) - \hat{\mu}^{'}(A) \mid \sigma(e_1, \ldots, e_t)\right]\right|\right] \quad (where\ \hat{\mu}_t^{'}(A) := \frac{1}{t}\sum_{k=1}^{t} 1(e_k^{'} \in A)) && (1.89)\\
&\leq E\left[\sup_{A\in\mathcal{A}} E\left[\left|\hat{\mu}_t(A) - \hat{\mu}^{'}(A)\right| \mid \sigma(e_1, \ldots, e_t)\right]\right] && (1.90)\\
&\leq E\left[\sup_{A\in\mathcal{A}} \left|\hat{\mu}_t(A) - \hat{\mu}^{'}(A)\right|\right] \quad (since\ \sup E \leq E \sup) && (1.91)\\
&= E\left[\sup_{A\in\mathcal{A}} \left|\frac{1}{t}\sum_{k=1}^{t}\left(1(e_k \in A) - 1(e_k^{'} \in A)\right)\right|\right] && (1.92)\\
&= E\left[\sup_{A\in\mathcal{A}} \left|\frac{1}{t}\sum_{k=1}^{t} s_k\left(1(e_k \in A) - 1(e_k^{'} \in A)\right)\right|\right] && (1.93)\\
&\quad (since\ e_1, \ldots, e_t, e_1^{'}, \ldots, e_t^{'}\ are\ independent\ and\ identically\ distributed) && (1.94)\\
&= \frac{1}{t}E\left[E\left[\sup_{A\in\mathcal{A}} \left|\sum_{k=1}^{t} s_k\left(1(e_k \in A) - 1(e_k^{'} \in A)\right)\right| \mid \sigma(e_1, \ldots, e_t, e_1^{'}, \ldots, e_t^{'})\right]\right] && (1.95)
\end{aligned}$$

Now, because of the independence of the s_k's of the rest of the variables, the inner conditional expectation is given by

$$E\left[\sup_{A\in\mathcal{A}}\left|\sum_{k=1}^{t}s_k\left(1(e_k\in A)-1(e_k^{'}\in A)\right)\right|\right], \tag{1.96}$$

where the e_k's and $e_k^{'}$'s are fixed and expectation is taken with respect to the s_k's. We show that (1.96) is bounded by $\sqrt{2t\ln(2S_{\mathcal{A}}(2t))}$, independently of the value of the e_k's and $e_k^{'}$'s, so concluding from (1.95) that (1.87) holds.

The first thing to note is that the qualifier $\sup_{A\in\mathcal{A}}$ in (1.96) can be substituted with a $\max$ over a finite set. In fact, for given e_k's and $e_k^{'}$'s, let $\bar{\mathcal{A}}\subset\mathcal{A}$ be a collection of sets such that any two sets in $\bar{\mathcal{A}}$ have different intersections with $\{e_1,\ldots,e_t,e_1^{'},\ldots,e_t^{'}\}$ and every possible intersection is represented once. Then, $\#\bar{\mathcal{A}}\leq S_{\mathcal{A}}(2t)$ and

$$E\left[\sup_{A\in\mathcal{A}}\left|\sum_{k=1}^{t}s_k\left(1(e_k^{'}\in A)-1(e_k\in A)\right)\right|\right]=E\left[\max_{A\in\bar{\mathcal{A}}}\left|\sum_{k=1}^{t}s_k\left(1(e_k^{'}\in A)-1(e_k\in A)\right)\right|\right]. \tag{1.97}$$

Next, let

$$v_k(A):=s_k\left(1(e_k^{'}\in A)-1(e_k\in A)\right). \tag{1.98}$$

We show that, for any $r>0$,

$$E[e^{rv_k(A)}]\leq e^{\frac{r^2}{2}},\forall k,A. \tag{1.99}$$

Quantity $1(e_k^{'}\in A)-1(e_k\in A)$ is either 0 or -1 or $+1$. If $1(e_k^{'}\in A)-1(e_k\in A)=0$, then $v_k(A)=0$ and (1.99) is trivially satisfied. If instead $1(e_k^{'}\in A)-1(e_k\in A)$ is -1 or $+1$, then $E[e^{rv_k(A)}]=\frac{1}{2}e^{r}+\frac{1}{2}e^{-r}=e^{\ln(\frac{1}{2}e^{r}+\frac{1}{2}e^{-r})}$. Function $\Phi(r):=\ln(\frac{1}{2}e^{r}+\frac{1}{2}e^{-r})$ is 0 for $r=0$. Its derivative $\frac{e^{r}-e^{-r}}{e^{r}+e^{-r}}$ is zero too in $r=0$, while the second derivative $\frac{4}{(e^{r}+e^{-r})^2}$ is less than 1, $\forall r$. Thus, by Taylor series expansion, for some $\xi\in[0,r]$:

$$\Phi(r)=\Phi(0)+\Phi^{'}(0)r+\frac{1}{2}\Phi^{''}(\xi)r^2\leq\frac{r^2}{2}, \tag{1.100}$$

so showing (1.99) in this case too.

We then have

$$e^{rE\left[\max_{A\in\bar{\mathcal{A}}}\left|\sum_{k=1}^{t}v_k(A)\right|\right]} \tag{1.101}$$

$$\leq\ E\left[e^{r\max_{A\in\bar{\mathcal{A}}}\left|\sum_{k=1}^{t}v_k(A)\right|}\right]\quad(by\ Jensen's\ inequality...) \tag{1.102}$$

$$= E\left[\max_{A\in\bar{\mathcal{A}}} e^{r\left|\sum_{k=1}^{t} v_k(A)\right|}\right] \tag{1.103}$$

$$\leq \sum_{A\in\bar{\mathcal{A}}} E\left[e^{r\sum_{k=1}^{t} v_k(A)} + e^{-r\sum_{k=1}^{t} v_k(A)}\right] \tag{1.104}$$

$$\leq 2\sum_{A\in\bar{\mathcal{A}}} E\left[e^{r\sum_{k=1}^{t} v_k(A)}\right] \quad (since\ \sum_{k=1}^{t} v_k(A)\ has\ symmentric\ distribution) \tag{1.105}$$

$$\leq 2\sum_{A\in\bar{\mathcal{A}}} \Pi_{k=1}^{t} E\left[e^{r v_k(A)}\right] \quad (since\ the\ v_k(A)'s\ are\ independent) \tag{1.106}$$

$$\leq 2S_{\mathcal{A}}(2t)\Pi_{k=1}^{t} e^{\frac{r^2}{2}} \quad (using(1.99)) \tag{1.107}$$

$$= 2S_{\mathcal{A}}(2t)e^{\frac{tr^2}{2}}. \tag{1.108}$$

Taking logarithm, we obtain

$$rE\left[\max_{A\in\bar{\mathcal{A}}}\left|\sum_{k=1}^{t} v_k(A)\right|\right] = \ln(2S_{\mathcal{A}}(2t)) + \frac{tr^2}{2}, \tag{1.109}$$

from which the sought result that (1.96) is bounded by $\sqrt{2t\ln(2S_{\mathcal{A}}(2t))}$ is achieved by taking $r = \sqrt{2\ln(2S_{\mathcal{A}}(2t))/t}$. This concludes the proof of (1.87).

Having proven the bound (1.87) on the expected value of $\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|$, we now appeal to the bounded difference inequality (Theorem A.12 in Appendix A) to draw conclusions on $\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|$ itself.

Note that $\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|$ is a function of $e_1, \ldots, e_t$ that satisfies the condition (A.46) of the bounded difference inequality with $\gamma_k = 1/t$, $\forall k$. Hence, the bounded difference inequality gives

$$P\left\{\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| - E\left[\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|\right] \geq \rho\right\} \leq e^{-2t\rho^2} \tag{1.110}$$

((1.110) expresses the somehow surprising result that the deviation of $\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|$ from its mean is small for almost all extractions of $e_1, \ldots, e_t$ (the exceptional set has probability as small as $e^{-2t\rho^2}$.) Roughly speaking, we could say that $\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)|$ is "almost deterministic".)

The conclusion of the theorem is now proven by the joint use of (1.87) and (1.110).

Inserting the bound (1.87) into (1.110), and letting $\epsilon := \rho + \sqrt{\frac{2\ln(2S_{\mathcal{A}}(2t))}{t}}$, we see that $P\{\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| > \epsilon\}$ is bounded by

$$e^{-2t\left(\epsilon - \sqrt{\frac{2\ln(2S_{\mathcal{A}}(2t))}{t}}\right)^2}. \tag{1.111}$$

Letting $a := \sqrt{\frac{2\ln(2S_{\mathcal{A}}(2t))}{t}}$, the exponent can be handled as follows:

$$\begin{aligned}
-2t(\epsilon - a)^2 &= -2t\left(\frac{1}{5}\epsilon^2 + \left[\frac{4}{5}\epsilon^2 + a^2 - 2\epsilon a\right]\right) && (1.112)\\
&= -2t\left(\frac{1}{5}\epsilon^2 + \left[\frac{2}{\sqrt{5}}\epsilon - \frac{\sqrt{5}}{2}a\right]^2 - \frac{1}{4}a^2\right) && (1.113)\\
&\leq -2t\left(\frac{1}{5}\epsilon^2 - \frac{1}{4}a^2\right) && (1.114)\\
&= -2t\left(\frac{1}{5}\epsilon^2 - \frac{1}{4}\frac{2\ln(2S_{\mathcal{A}}(2t))}{t}\right) && (1.115)\\
&= -\frac{2t\epsilon^2}{5} + \ln(2S_{\mathcal{A}}(2t)) && (1.116)
\end{aligned}$$

so that

$$e^{-2t\left(\epsilon - \sqrt{\frac{2\ln(2S_{\mathcal{A}}(2t))}{t}}\right)^2} \leq 2S_{\mathcal{A}}(2t)e^{-2t\epsilon^2/5}, \qquad (1.117)$$

so concluding the proof. □

The merit of the Vapnik-Chervonenkis inequality is that it converts the problem of evaluating the probability that the empirical mean deviates from the true mean by more than ϵ for some set A in $\mathcal{A}$ into the combinatorial problem of computing the shatter function of $\mathcal{A}$. If $S_{\mathcal{A}}(2t)$ grows at a subexponential rate, then $\sum_{t=1}^{\infty} P\left\{\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| > \epsilon\right\} < \infty$ and, by virtue of the Theorem A.4 in Appendix A, we can conclude that $\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| \to 0$, almost surely.

It turns out that $S_{\mathcal{A}}(2t)$ either grows very fast or as slowly as polynomially. In fact, if $V_{\mathcal{A}} = \infty$, then $S_{\mathcal{A}}(2t) = 2^{2t}$ and the subexponential growth fails. On the other hand, if $V_{\mathcal{A}} < \infty$, the polynomial bound given in the next theorem holds true.

THEOREM 1.17 (Sauer's theorem) *For all n,*

$$S_{\mathcal{A}}(n) \leq \sum_{k=0}^{\min\{V_{\mathcal{A}}, n\}} \binom{n}{k} \leq (n+1)^{V_{\mathcal{A}}}, \qquad (1.118)$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$ are the bynomial coefficients.

PROOF. Given $n > 0$, consider a set $\{e_1, \ldots, e_n\}$ such that $\#\{\{e_1, \ldots, e_n\} \cap A; A \in \mathcal{A}\} = S_{\mathcal{A}}(n)$ and let

$$H_0 := \{h = (h_1, \ldots, h_n) \in \{0,1\}^n \; such \; that, \; for \; some \; A \in \mathcal{A}, \tag{1.119}$$

$$h_k = 1(e_k \in A), k = 1, \ldots, n\}. \tag{1.120}$$

Clearly, $\#H_0 = S_{\mathcal{A}}(n)$. The thesis is proven by showing that $\#H_0 \leq \sum_{k=0}^{min\{V_{\mathcal{A}}, n\}} \binom{n}{k}$. (The reader may find it convenient to follow the rest of the proof on an example. Consider e.g. $H_0 = \{(0,0,0), (0,1,0), (1,0,0), (1,1,0), (1,1,1)\}$.)

We say that a set $H \subset \{0,1\}^n$ shatters a set of integers $I = \{i_1, \ldots, i_m\} \subset \{1, \ldots, n\}$ if the restriction of H to the components $i_1, \ldots, i_m$ is the full m-dimensional binary hypercube, namely

$$\{(h_{i_1}, \ldots, h_{i_m}) = restriction \; of \; some \; h \in H \; to \; I\} = \{0,1\}^m. \tag{1.121}$$

Notice that the set H_0 cannot shatter a set of integers with cardinality $m > V_{\mathcal{A}}$, since otherwise the VC-dimension of $\mathcal{A}$ would be larger than $V_{\mathcal{A}}$.

The rest of the proof proceeds as follows. We iteratively modify H_0 into new subsets $H_1, \ldots, H_n$ of $\{0,1\}^n$ in such a way that:

i) $\#H_0 = \#H_1 = \cdots = \#H_n$; and
ii) if a set of integers is shatterd by H_{k+1}, then it was also shatterd by H_k.

The thesis is then proven by counting the elements of H_n.

In order to construct H_1, proceed as follows. For any vector of $h = (h_1, \ldots, h_n)$ of H_0, if $h_1 = 1$, then flip it to 0, unless $(0, h_2, \ldots, h_n) \in H_0$. If instead $h_1 = 0$, then keep the vector unchanged. The collection of the so-obtained vectors is H_1. Clearly, $\#H_1 = \#H_0$. Moreover, we claim that if I is shatterd by H_1, it is shattered by H_0 as well. This last claim needs a bit of inspection. Suppose first that $1 \notin I$. Then, I is certainly shattered by H_0 since the restriction to I of vectors in H_0 has remained unchanged when moving on to H_1. Suppose instead that $1 \in I$ and I is shattered by H_1, that is

$$\{(h_1, h_{i_2}, \ldots, h_{i_m}) = restriction \; of \; some \; h \in H_1 \; to \; I\} = \{0,1\}^m, \tag{1.122}$$

where $I = \{1, i_2, \ldots, i_m\}$. We show that all the restrictions $(h_1, h_{i_2}, \ldots, h_{i_m})$ of H_1 are also restrictions of H_0, so proving that H_0 shatters I too. If $h_1 = 1$, then the restriction belongs to H_0 since the generating vector has not been modified. If $h_1 = 0$ and the generating vector has not been modified, again the restriction belongs to H_0 too. Finally,

if the restriction is $(0, h_{i_2}, \ldots, h_{i_m})$ and indeed the first 0 has been obtained by flipping to 0 a 1, we claim that $(0, h_{i_2}, \ldots, h_{i_m})$ was already present as a restriction of H_0. To see this, note that, since I is shattered by H_1, even the restrictions $(1, h_{i_2}, \ldots, h_{i_m})$ belongs to H_1. But, then, $(0, h_{i_2}, \ldots, h_{i_m})$ is a restriction of H_0 since, otherwise, $(1, h_{i_2}, \ldots, h_{i_m})$ would have been converted into $(0, h_{i_2}, \ldots, h_{i_m})$.

Now, repeat the same procedure, but this time by flipping the second component of each vector. The so-obtained H_2 has the same cardinality as H_1 and, thereby, as H_0. Moreover, if I is shatterd by H_2, then I is also shattered by H_1 and, in turn, by H_0. Proceeding similarly for all components, we arrive to the subset H_n such that i) and ii) are satisifed.

The thesis is now completed by counting the number of elements in H_n.

To this purpose, note first that if H_n contains a vector $(h_1, \ldots, h_n)$, then it also contains all vectors obtained by this by flipping to 0 as many 1's as we want (this fact is proven by a bit of inspection using the way the H_k's have been constructed.) This implies that if H_n contains a vector with p 1's, then it shatters the set I of the indeces of these 1's. But then, because of ii) even H_0 shatters I. Recalling that H_0 cannot shatter set of integers with cardinality bigger than $V_{\mathcal{A}}$, we conclude that $p \leq V_{\mathcal{A}}$, that is all vectors in H_n have at most $V_{\mathcal{A}}$ 1's. Then, the question we need to answer is: how many vectors with at most $V_{\mathcal{A}}$ 1's exist in $\{0,1\}^n$? There is just $1 = \binom{n}{0}$ vector with all 0's; there are $\binom{n}{1}$ vectors with only one 1 and so on. Thus,

$$\#H_n \leq \binom{n}{0} + \binom{n}{1} + \binom{n}{2} + \cdots + \binom{n}{\min\{V_{\mathcal{A}}, n\}} = \sum_{k=0}^{\min\{V_{\mathcal{A}},n\}} \binom{n}{k}, \quad (1.123)$$

so proving the first inequality in the statement of the theorem. The second inequality is obtained by simple boundings:

$$\sum_{k=0}^{\min\{V_{\mathcal{A}},n\}} \binom{n}{k} \leq \sum_{k=0}^{\min\{V_{\mathcal{A}},n\}} \frac{n^k}{k!} \leq \sum_{k=0}^{V_{\mathcal{A}}} \frac{n^k}{k!} = \sum_{k=0}^{V_{\mathcal{A}}} \frac{n^k V_{\mathcal{A}}!}{k!(V_{\mathcal{A}}-k)!} = \sum_{k=0}^{V_{\mathcal{A}}} n^k \binom{V_{\mathcal{A}}}{k} = (n+1)^{V_{\mathcal{A}}}. \quad (1.124)$$

□

Now, consider a collection $\mathcal{A}$ of sets such that $V_{\mathcal{A}} < \infty$. Sauer's theorem tells us that $S_{\mathcal{A}}(2t) \leq (2t+1)^{V_{\mathcal{A}}}$, which, used in the Vapnik-Chervoenenkis inequality, gives

$$P\left\{\sup_{A \in \mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| > \epsilon\right\} \leq 2(2t+1)^{V_{\mathcal{A}}} e^{-2t\epsilon^2/5}. \quad (1.125)$$

Then, since $\sum_{t=1}^{\infty}(2t+1)^{V_{\mathcal{A}}}e^{-2t\epsilon^2/5} < \infty$, $\forall \epsilon > 0$, Theorem A.4 permits us to conclude that $\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| \to 0$, almost surely.

As an example of application of this result, we can recover the Glivenko-Cantelli theorem. In fact, in Example 1.14 we have seen that the set of semi-infinite intervals $\{(-\infty, x], x \in R\}$ has VC-dimension 1. Hence,

$$\sup_{x\in R} |\hat{\mu}_t(-\infty, x] - \mu(-\infty, x]| = \sup_{x\in R} \left|\hat{F}_t(x) - F(x)\right| \to 0, \quad almost\ surely. \tag{1.126}$$

On the other hand, it should be noted that the results of this section are very general and can be applied to many more situations than the Glivenko-Cantelli theorem.

For easy reference, the result is summarized in the following theorem.

THEOREM 1.18 (Uniform convergence of empirical probabilities) *In a measurable space $(E, \mathcal{E})$, consider a collection $\mathcal{A}$ of sets $A \in \mathcal{E}$. Given a sequence $\{e_t\}$ of independent random elements taking value in E with a common probability measure μ, consider the empirical mean of the sets A, that is $\hat{\mu}_t(A) := \frac{1}{t}\sum_{k=1}^{t} 1(e_k \in A)$. Then, if the VC-dimension of $\mathcal{A}$ is finite, $\hat{\mu}_t(A)$ converges uniformly to $\mu(A)$ with probability 1:*

$$\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| \to 0, \quad almost\ surely. \tag{1.127}$$

□

Asymptotic convergence for $\{0, 1\}$-valued predictors

Here, we study the asymptotic convergence of PEM identification methods for nonlinear predictors taking value in $\{0, 1\}$.

We take a somehow broad point of view by assuming that the system input u_t takes value in a measurable space $(U, \mathcal{U})$. Clearly, as a particular case we can consider $(R^n, \mathcal{B}(R^n))$ or even go down to $(R, \mathcal{B}(R))$, that is to real inputs. We are motivated to consider $(U, \mathcal{U})$ by the fact that in pattern recognition one is often called to consider input spaces more general that R^n and by the fact that, with the theory as far developed in our hands, such a generalization is attained at no additional cost.

Suppose that the u_t's are independently extracted in U according to a probability μ_U. The independence assumption means that we are considering static problems and dynamics is not allowed. The system output is

$$y_t = \begin{cases} 1, & if\ u_t \in A^\circ, \\ 0, & otherwise. \end{cases} \tag{1.128}$$

where $A^\circ \in \mathcal{U}$ is a set unknown to us.

We look for a predictor in a prespecified class that minimizes the probability of misclassifying an unseen output y_t. Precisely, given a collection of sets $\Theta \subset \mathcal{U}$, we want to minimize

$$\bar{\mathcal{J}}(\theta) = E[1(y_t \neq \hat{y}_t(\theta))], \tag{1.129}$$

where $\hat{y}_t(\theta)$ is the predictor given by

$$\hat{y}_t(\theta) = \begin{cases} 1, & if\ u_t \in \theta, \\ 0, & otherwise. \end{cases} \tag{1.130}$$

As usual, θ is selected by minimizing the quadratic sample cost

$$\mathcal{J}_t(\theta) := \frac{1}{t}\sum_{k=1}^{t} 1(y_k \neq \hat{y}_k(\theta)) \tag{1.131}$$

and the corresponding minimizer is named $\hat{\theta}_t$ (we assume that such a minimizer exists. This assumption is motivated by convenience in the notation but is not essential. If more than one minimizer occurs, $\hat{\theta}_t$ can be any one of them.) Hence, our goal is to verify whether $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$ with probability 1.

Now, based on Lemma 1.1, $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$ with probability 1 provided that

$$\sup_{\theta\in\Theta} \left|\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right| \to 0, \quad almost\ surely. \tag{1.132}$$

In the following, we show that this latter condition is in fact satisfied if the VC-dimension of Θ is finite. The idea underlying this result is that we are guaranteed to select the best predictor provided that the class of predictors is not too rich (i.e. it has finite VC-dimension.)

The derivation of the result is almost immediate using the uniform convergence Theorem 1.18. All we need to do is to recast the problem so as it matches the framework of the theorem.

First of all, introduce the set $E := U\times\{0,1\}$ with the σ-algebra $\mathcal{E} := \mathcal{U}\otimes\sigma\{0,1\}$ formed by all subsets of the form $A_1\times\{0\} \cup A_2\times\{1\}$ where $A_1, A_2 \in \mathcal{U}$. Probability μ is defined as: $\mu(A_1\times\{0\} \cup A_2\times\{1\}) = \mu_U((A_1 \cap (A^\circ)^c) \cup (A_2 \cap A^\circ)))$. Finally, the collection of sets $\mathcal{A}$ is given by

$$\mathcal{A} = \{A(\theta) := \theta\times\{0\} \cup \theta^c\times\{1\}, \quad \theta \in \Theta\}. \tag{1.133}$$

Note that, for any $\theta \in \Theta$, we have

$$\mu(A(\theta)) = \mu_U((\theta \cap (A^\circ)^c) \cup (\theta^c \cap A^\circ))) = E[1(y_t \neq \hat{y}_t(\theta))] = \bar{\mathcal{J}}(\theta), \tag{1.134}$$

and

$$\hat{\mu}_t(A(\theta)) = \frac{1}{t}\sum_{k=1}^{t} 1\left(u_k \in (\theta \cap (A^\circ)^c) \cup (\theta^c \cap A^\circ))\right) = \frac{1}{t}\sum_{k=1}^{t} [y_k - \hat{y}_k(\theta)]^2 = \mathcal{J}_t(\theta). \tag{1.135}$$

Thus, convergence (1.132) can be rewritten in the new notations as $\sup_{\theta\in\Theta} |\hat{\mu}_t(A(\theta)) - \mu(A(\theta))| \to 0$, almost surely. But this is exactly the thesis of Theorem 1.18, so that what is left to verify is whether the assumption of this theorem that the VC-dimension of $\mathcal{A}$ is finite is satisfied in the present context. This matter is taken care of in the next lemma.

LEMMA 1.19 *With all notations as above, we have* $V_{\mathcal{A}} = V_\Theta$.

PROOF. Given any set $\{(u_1, y_1), \ldots, (u_n, y_n)\} \in U \times \{0, 1\}$, consider the "projection" set of points $\{u_1, \ldots, u_n\} \in U$. We show that the number of different sets that can be achieved by intersecting $\{(u_1, y_1), \ldots, (u_n, y_n)\}$ with sets in $\mathcal{A}$ equals the number of sets that can be achieved by intersecting $\{u_1, \ldots, u_n\}$ with sets in Θ.

Note that, this suffices to prove the lemma. In fact, this result implies in particular that the largest number of different sets that can be formed by intersecting a set of n points in $U \times \{0, 1\}$ with sets in $\mathcal{A}$ equals the largest number of different sets that can be formed by intersecting a set of n points in U with sets in Θ, so that $S_{\mathcal{A}}(n) = S_\Theta(n)$. This immediately implies that $V_{\mathcal{A}} = V_\Theta$, that is the thesis of the lemma.

Letting

$$H := \{h = (h_1, \ldots, h_n) \in \{0, 1\}^n \text{ such that, for some } \theta \in \Theta, \tag{1.136}$$

$$h_k = 1(u_k \in \theta), k = 1, \ldots, n\} \tag{1.137}$$

and

$$M := \{m = (m_1, \ldots, m_n) \in \{0, 1\}^n \text{ such that, for some } A(\theta) \in \mathcal{A}, \tag{1.138}$$

$$m_k = 1((u_k, y_k) \in A(\theta)), k = 1, \ldots, n\}, \tag{1.139}$$

what we need to prove is that $\#M = \#H$.

For any given θ, we obtain a vector $h \in H$ such that $h_k = 1(u_k \in \theta), k = 1, \ldots, n$ and a vector $m \in M$ such that $m_k = 1((u_k, y_k) \in A(\theta)), k = 1, \ldots, n$. A bit of inspection reveals that $m_k = h_k \Delta y_k$, where Δ is the "exclusive or", that is m_k is 0 if and only if h_k and y_k are both 0 or both 1. Thus, $m = h\Delta y$, where Δ is the "exclusive or" applied componentwise and $M = \{h\Delta y, h \in H\}$. Observing that the map $h \to h\Delta y$ is one-to-one, we conclude that $\#M = \#H$. □

Summarizing the discussion we can say that if V_Θ is finite, than by virtue of Lemma 1.19, the VC-dimansion of $\mathcal{A}$ is finite too and hence, by Theorem 1.18, we conclude that (1.132) is met. But then Lemma 1.1 guarantees that $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$ with probability 1.

For easy reference, this result is stated as a theorem.

THEOREM 1.20 (Convergence of PEM methods for $\{0,1\}$-valued predictors) *Suppose that the system output is given by (1.128) where the system input u_t is generated in an independent fashion. If the VC-dimension of Θ is finite, then the probability of misclassification $\bar{\mathcal{J}}(\hat{\theta}_t)$ (see (1.129)) associated to the set $\hat{\theta}_t$ obtained by minimizing the sample cost (1.131) is asymptotically optimal in the sense that*

$$\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta), \quad almost\ surely. \tag{1.140}$$

□

Remark 1.21 *The attentive reader may have noticed that the complexity of the nature, expressed by the fact that A° can be any measurable set, has been absorbed in the definition of the probability μ defined over $\mathcal{U} \otimes \sigma\{0,1\}$. This is achieved at no extra cost since Theorem 1.18 holds with no assumptions at all on the underlying probability μ.* □

Uniform convergence for $\{0,1\}$-valued predictors

In the previous point, we have shown that the best $\{0,1\}$-valued predictor can be asymptotically estimated through the minimization of the empirical identification cost if the predictor class is not too rich. Precisely, Theorem 1.20 proves that the estimate

$$\hat{\theta}_t := \arg\min_{\theta\in\Theta} \mathcal{J}_t(\theta), \tag{1.141}$$

where $\mathcal{J}_t(\theta)$ is the empirical identification cost, is such that

$$\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta), \quad almost\ surely, \tag{1.142}$$

where $\bar{\mathcal{J}}(\theta)$ is the probability of misclassification associated to θ, provided that the set Θ has finite VC-dimension.

We now observe that a much stronger result of uniform convergence can in fact be achieved from the Vapnik-Chervonenkis theory of uniform convergence of empirical means.

We start by recalling the inequality (1.125):

$$P\left\{\sup_{A\in\mathcal{A}} |\hat{\mu}_t(A) - \mu(A)| > \epsilon\right\} \leq 2(2t+1)^{V_{\mathcal{A}}} e^{-2t\epsilon^2/5}, \tag{1.143}$$

which bounds the probability that the maximal deviation of the empirical mean from the corresponding theoretical value is greater than ϵ. If applied in the context of $\{0,1\}$-valued predictors of the previous point, this inequality writes

$$P\left\{\sup_{\theta\in\Theta} \left|\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right| > \epsilon\right\} \leq 2(2t+1)^{V_\Theta} e^{-2t\epsilon^2/5}. \tag{1.144}$$

On the other hand, the following chain of inequalities (extracted from the proof of Lemma 1.1) holds true

$$\begin{aligned}
\bar{\mathcal{J}}(\hat{\theta}_t) &= \mathcal{J}_t(\hat{\theta}_t) + \left(\bar{\mathcal{J}}(\hat{\theta}_t) - \mathcal{J}_t(\hat{\theta}_t)\right) && (1.145)\\
&\leq \mathcal{J}_t(\hat{\theta}_t) + \sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right| && (1.146)\\
&= \inf_{\theta\in\Theta} \mathcal{J}_t(\theta) + \sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right| && (1.147)\\
&= \inf_{\theta\in\Theta} \left[\bar{\mathcal{J}}(\theta) + \left(\mathcal{J}_t(\theta) - \bar{\mathcal{J}}(\theta)\right)\right] + \sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right| && (1.148)\\
&\leq \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) + 2\sup_{\theta\in\Theta} \left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right|, && (1.149)
\end{aligned}$$

showing that

$$\bar{\mathcal{J}}(\hat{\theta}_t) - \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) \leq 2\left|\bar{\mathcal{J}}(\theta) - \mathcal{J}_t(\theta)\right|. \tag{1.150}$$

Using this inequality in (1.144) finally gives

$$P\left\{\bar{\mathcal{J}}(\hat{\theta}_t) - \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) > 2\epsilon\right\} \leq 2(2t+1)^{V_\Theta} e^{-2t\epsilon^2/5}. \tag{1.151}$$

The fundamental observation is that the right-hand-side depends neither on the probability μ_U according to which the inputs u to the system are drawn, nor on the underlying "true set" A°. In other word, the bound is completely independent of the data generation mechanism. Thus, we see that

$$\sup_{\mu_U} \sup_{A^\circ} P\left\{\bar{\mathcal{J}}(\hat{\theta}_t) - \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) > 2\epsilon\right\} \to 0, \tag{1.152}$$

that is uniform convergence in probability to 0 takes place.

More importantly, for given $\epsilon > 0$ and $\delta > 0$, based on (1.151) we can determine an integer $t(\epsilon, \delta)$ such that

$$\sup_{\mu_U} \sup_{A^\circ} P\left\{\bar{\mathcal{J}}(\hat{\theta}_t) - \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) > 2\epsilon\right\} \leq \delta, \quad for\ t \geq t(\epsilon, \delta). \tag{1.153}$$

The interpretation of $t(\epsilon, \delta)$ is very meaningful. It is the number of data points that guarantee, with probability $1 - \delta$, that a set $\hat{\theta}_t$ is selected whose probability of missclassification is no more than 2ϵ apart from the lowest possible probability of missclassification.

Function $t(\epsilon, \delta)$ has been widely studied in the statistical learning literature under the name of "sample complexity function". In this literature, an acronym that is often used is PAC-learning (Probably Approximately Correct - learning) to signify that $t(\epsilon, \delta)$ measuraments are enough to only gurantee that the selected $\hat{\theta}_t$ is approximately (up to 2ϵ) correct (i.e. optimal) with high $(1 - \delta)$ probability, or confidence. Here, it should be noted that the presence of a confidence parameter δ is essential. In fact, for any large (though finite) number of data it is impossible to aiming for results that hold with probability 1 because data are randomly selected so that there is always a chance that they are little informative. The reader is referred to the books [13], [29], [25] for thourough presentations of PAC-learning.

EXAMPLE 1.22 (selecting a disk in $\mathbf{R^2}$) *Suppose we are given data u_k and $y_k = 1(u_k \in A^\circ)$, $k = 1, \ldots, t$, where A° is an underlying unknown set and $u_k \in R^2$ are independently extracted according to a probability μ_U.*

We want to select a disk $\theta \subset R^2$ so that the probability of misclassification is minimized.

To this aim, we select a disk $\hat{\theta}_t$ by minimizing

$$\frac{1}{t}\sum_{k=1}^{t} 1(y_k \neq \hat{y}_k(\theta)). \tag{1.154}$$

How many samples do we need to extract so that relation

$$probability\ of\ misclassification \tag{1.155}$$
$$\leq\ minimal\ probability\ of\ misclassification + 2\epsilon \tag{1.156}$$

is satisfied with probability greater that δ? (note that Statement (1.156) is stochastic since the "probability of misclassification" in the left-hand-side depends on the u_k extractions; yet, the right-hand-side id fully deterministic.)

The answer is easily found by recalling from Example 1.15 that $V_\Theta = 3$, so that, by virtue of (1.151), all we need to do is to make explicit relation

$$2(2t+1)^3 e^{-2t\epsilon^2/5} = \delta \tag{1.157}$$

with respect to t. If, for instance, $2\epsilon = 0.1$ and $\delta = 0.1$, we obtain: $t = 36600$. □

Bias versus variance

Equation (1.153) provides a bound for the actually achieved performance with respect to the optimal performance achievable within the selected class of predictors. Suboptimality is due to the dispersion of $\hat{\theta}_t$ around the optimal parameter value.

For different classes of predictors, both dispersion and optimal performance within the class, as measured by $\inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$, change. Tipically, $\inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$ decreases for larger classes, while dipsersion increases. Thus, in the selection of the predictor class, it is important to compromise between two contrasting objectives: taking a class rich enough so that the best performance within the class is satisfactory and keeping low the effect of dispersion within the class. This duality is known as the bias versus variance tradeoff.

In general, an evaluation of how $\inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$ changes for different classes needs relying on some a-priori assumption on the data generation mechanism.

Necessity of the condition $\mathbf{V_\Theta < \infty}$

In Theorem 1.20, we have seen that:

$$V_\Theta < \infty \implies \bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta),\ almost\ surely. \tag{1.158}$$

We now ask whether the condition $V_\Theta < \infty$ is necessary for the conclusion to the right to hold. The answer is no from a trivial example.

EXAMPLE 1.23 ($\mathbf{V_\Theta < \infty}$ is not necessary for $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$, almost surely)
Let Θ be the collection of all sets formed by an arbitrary, though finite, number of points.

In Example 1.14, we have seen that $V_\Theta = \infty$. Suppose that μ_U has concentrated mass 1 in 0, that is any set that contains 0 has μ_U probability 1, while all other sets have probabiloity 0 and take $A^\circ = \{0\}$. It is immediate to conclude that $\bar{\mathcal{J}}(\theta) = 1$ if $0 \in \theta$ and 0 otherwise. Moreover, since with probability 1 $u = 0$ for all extractions, we see that, for all t, $\mathcal{J}_t(\theta) = 1$ if $0 \in \theta$ and 0 otherwise with probability 1, so concluding that $\mathcal{J}_t(\theta) = \bar{\mathcal{J}}(\theta)$, $\forall\theta$, with probability 1. Thus, uniform convergence takes place and $\bar{\mathcal{J}}(\hat{\theta}_t) \to \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta)$, almost surely. □

The previous example is trivialized by the fact that probability μ_U concentrates all mass in just one point, so rendering ineffective the richness of Θ. Other less trivial situations can be found as well.

Now, let us now strengthen our requirements and consider uniform convergence. Precisely, from point "Uniform convergence for $\{0,1\}$-valued predictors" we know that

$$V_\Theta < \infty \quad \Longrightarrow \quad \forall\epsilon,\delta > 0, \textit{there exists } t(\epsilon,\delta) \textit{ such that} \tag{1.159}$$

$$\sup_{\mu_U}\sup_{A^\circ} P\left\{\bar{\mathcal{J}}(\hat{\theta}_t) - \inf_{\theta\in\Theta}\bar{\mathcal{J}}(\theta) > 2\epsilon\right\} \le \delta, \quad \textit{for } t \ge t(\epsilon,\delta) \tag{1.160}$$

Now we ask whether $V_\Theta < \infty$ is required for the right-hand-side to hold.

Not only the answer is positive this time, but an even stronger statement holds true. In fact, in (1.160) $\hat{\theta}_t$ has been determined through a specific algorithm, that is the minimization of the empirical cost $\mathcal{J}_t(\theta)$. We show that no algorithm for determining $\hat{\theta}_t$ exists such that the right-hand-side of (1.160) holds true if $V_\Theta = \infty$. Thus, in particular the right-hand-side of (1.160) is not true if $\hat{\theta}_t$ is obtained by minimizing the empirical cost $\mathcal{J}_t(\theta)$. First, a precise definition of algorithm is given and then the result is proven.

DEFINITION 1.24 (Algorithm for PAC-learning) *A learning algorithm is an indexed family of maps $A_t : (U \times \{0,1\})^t \to \Theta$. For a given data set $(u_1, y_1, \ldots, u_t, y_t)$, the algorithm outputs an estimate $\hat{\theta}_t := A_t((u_1, y_1, \ldots, u_t, y_t)) \in \Theta$.*
A learning algorithm $\{A_t\}$ is probably approximately correct (PAC) if, $\forall\epsilon,\delta > 0$, the right-hand-side of (1.160) holds true. □

THEOREM 1.25 ($V_\Theta = \infty \Rightarrow$ no algorithm is PAC) *Suppose that the system output is given by (1.128) where the system input u_t is generated in an independent fashion with probability μ_U. If the VC-dimension of Θ is infinite, then no PAC algorithm (see Definition 1.24) exists.*

PROOF. Fix $\epsilon = \frac{1}{16}$ and $\delta = \frac{1}{8}$. We show that no finite $t(\epsilon, \delta) = t\left(\frac{1}{16}, \frac{1}{8}\right)$ exists that makes true the right-hand-side of (1.160).

Given any t, consider a set $\tilde{U}$ of $2t$ points in U that are shattered by Θ (such a set exists since $V_\Theta = \infty$) and let $\bar{\mu}_U$ be the probability that assigns to each one of the $2t$ points probability $\frac{1}{2t}$.

Moreover, let $\mathcal{A}^\circ = \{A^\circ\}$ be a collection of 2^{2t} sets such that each set in $\mathcal{A}^\circ$ has a different intersection with the set of $2t$ points.

Take any algorithm, and let $\hat{\theta}_t$ be the corresponding estimate at time t. We show that a set $\bar{A}^\circ$ can be found in $\mathcal{A}^\circ$ such that

$$P\left\{\bar{\mathcal{J}}(\hat{\theta}_t) - \inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) > 2\epsilon = 2\frac{1}{16}\right\} > \delta = \frac{1}{8}, \tag{1.161}$$

for $\mu_U = \bar{\mu}_U$ and $A^\circ = \bar{A}^\circ$, so violating the right-hand-side of (1.160).

Given a random extraction of points in U, $(u_1, \ldots, u_t)$, group the sets of $\mathcal{A}^\circ$ in equivalence classes that agree over $(u_1, \ldots, u_t)$. All sets A° in the same equivalence class ouputs the same data set $(u_1, y_1, \ldots, u_t, y_t)$ and, therefore, the same $\hat{\theta}_t$. Consider now a given equivalence class and

$$\frac{1}{\#(eq.\ class)} \sum_{A^\circ \in\ eq.\ class} \bar{\mu}_U(A^\circ \Delta\ \hat{\theta}_t), \tag{1.162}$$

where Δ is the symmetric difference and $\#$ denotes "number of elements". We claim that (1.162) $\geq \frac{1}{4}$. In fact, for every $u \in \tilde{U}$ that is not in $(u_1, \ldots, u_t)$, there are half sets in the equivalence class that agree with $\hat{\theta}_t$ and half that do not. Noting that each one of these u's has associated a $\mu_U(u) = \frac{1}{2t}$ and that these u's are at least t (at least because in the set $(u_1, \ldots, u_t)$ there can as well be repetitions), the result is easily achieved.

Since the above result holds for any equivalence class, we also have:

$$\frac{1}{\#\mathcal{A}^\circ} \sum_{A^\circ\in\mathcal{A}^\circ} \bar{\mu}_U(A^\circ \Delta\ \hat{\theta}_t) \geq \frac{1}{4}, \tag{1.163}$$

where the expression holds for every random extraction of points $(u_1, \ldots, u_t)$. Taking expectation, we then obtain

$$E\left[\frac{1}{\#(\mathcal{A}^\circ)} \sum_{A^\circ\in\mathcal{A}^\circ} \bar{\mu}_U(A^\circ \Delta\ \hat{\theta}_t)\right] \geq \frac{1}{4}, \tag{1.164}$$

from which we conclude that there exists at least a set $\bar{A}^\circ \in \mathcal{A}^\circ$ such that

$$E\left[\bar{\mu}_U(\bar{A}^\circ \Delta\ \hat{\theta}_t)\right] \geq \frac{1}{4}. \tag{1.165}$$

Now, let $p := P(\bar{\mu}_U(\bar{A}^\circ \Delta\ \hat{\theta}_t) > \frac{1}{8})$. Observing that $\bar{\mu}_U(\bar{A}^\circ \Delta\ \hat{\theta}_t) \leq 1$, we can write

$$\frac{1}{4} \leq E\left[\bar{\mu}_U(\bar{A}^\circ \Delta\ \hat{\theta}_t)\right] \leq p + (1-p)\frac{1}{8}, \tag{1.166}$$

where the first inequality is (1.165), so obtaining the lower bound $p \geq \frac{1}{7}$.

Finally, observing that, for $\mu_t = \bar{\mu}_t$ and $A^\circ = \bar{A}^\circ$, $\bar{\mathcal{J}}(\hat{\theta}_t) = \bar{\mu}_U(\bar{A}^\circ \Delta\ \hat{\theta}_t)$ and $\inf_{\theta\in\Theta} \bar{\mathcal{J}}(\theta) = 0$, because there certainly exists a $\theta \in \Theta$ that agrees with $\bar{A}^\circ$ over the set of $2t$ points in U since Θ shatters this set of points, we have

$$\sup_{\mu_U}\sup_{A^\circ} P\left\{\bar{\mathcal{J}}(\hat{\theta}_t) - \inf_{\theta\in\Theta}\bar{\mathcal{J}}(\theta) > 2\frac{1}{16}\right\} \geq P\left\{\bar{\mu}_U(\bar{A}^\circ \Delta\ \hat{\theta}_t) > 2\frac{1}{16}\right\} = p \geq \frac{1}{7}, \tag{1.167}$$

so proving equation (1.161). □

1.4.3 Complements and Bibliographical notes

The Glivenko-Cantelli theorem was first proven in R and very early on extended to R^n. A significant generalization was obtained by Ranga Rao, [21], who considered the uniform convergence of empirical probability for families of convex sets. Interestingly enough, his results are based on topological conditions, and not on combinatorial assumptions as in the Vapnik-Chervoenenkis theory. The uniform convergence of empirical probability theory presented herein has been settled by Vapnik and Chervonenkis in the seminal contributions [26], [27], and [28].

The notion of VC-dimension appears for the first time in [26] (though without referring to it as VC-dimension.) Sauer's theorem is due to Sauer, [22]. The constant that multiplies $-t\epsilon^2$ in the exponent of (1.125) has been the object of several improvements through time. Up to date, the best known constant is 1, see [18]. Most literature on the uniform convergence of empirical means focuses on convergence in probability. Steele, [23], considers almost sure convergence, as it is done here (see Theorem 1.18.)

The definition of PAC-learnability was introduced by Valiant, [24], though the terminology "probably approximately correct" has been coined later, probably by Angluin, [1]. Valiant, [24], only considers Boolean formulae, for which the underlying U set has finite cardinality, but adds computational requirements to his definition of learnability. The connection between the VC-dimension and PAC-learnability has been highlighted by Blumer

et al., [7], and Anthony et al., [2]. Theorem 1.25 showing that the finiteness of the VC-dimension is necessary for PAC-learning is due to Benedek and Itai, [5].

The P-dimension has been introduced by Pollard, [19]. Haussler, [11], studies the connections between the P-dimension and PAC-learnability of function classes, and proves, among other results, that the finiteness of the P-dimension implies PAC-learnability. While it is known that the converse is not true in general (see e.g. Example 7.1 in [29]), Bartlett et al., [3], prove that PAC-learnability does require the finiteness of the P-dimension in presence of noise. Alternative definitions of the richness of a function class are given by [10] and [17].

The definition of PAC-learnability given here (see Definition 1.24) requires that learnability takes place uniformly with respect to all probabilities μ_U. In certain instances, μ_U is known in advance, so that uniformity with respect to μ_U can be dropped in (1.160). Benedek and Itai, [5], give necessary and sufficient conditions for PAC-learnability in this case.

1.5 Main points of the chapter

1. In this chapter we have studied the conditions under which the predictor which is selected by minimizing the empirical quadratic identification criterion $\mathcal{J}_t(\theta)$ tends (in a suitable sense) to the optimal predictor minimizing the theoretical criterion $\bar{\mathcal{J}}(\theta)$. We have seen that the crucial condition for this to hold is the uniform convergence of $\mathcal{J}_t(\theta)$ to $\bar{\mathcal{J}}(\theta)$.

2. In a linear setting, the main issue is the uniform stability of the dynamical systems in play. If stability holds, signals stay bounded (in a stochastic sense) and they are long-term independent. This implies, by a law of large numbers, that $\mathcal{J}_t(\theta)$ tends to $\bar{\mathcal{J}}(\theta)$ for a fixed θ. Uniform convergence then follows from the fact that linear predictors have simple enough structure.

3. In a nonlinear setting, the complexity of the predictor structure can hinder the uniform convergence of $\mathcal{J}_t(\theta)$ to $\bar{\mathcal{J}}(\theta)$, so resulting in an inability of selecting a proper predictor through the minimization of the empirical identification criterion.

4. For $\{0,1\}$-valued predictors, if the VC-dimension of the set class is finite, uniform convergence takes place. Moreover, results concerning the sample size needed to achieve certain quality levels of the predictor can be worked out.

Appendix A

STOCHASTIC CONVERGENCE

A.1 Probabilistic notions of convergence

We present a number of probabilistic notions of convergence of a sequence of random variables $\{v_n\}$ to a limit random variable v and relate each one to the others.

DEFINITION A.1 (stochastic convergence) *Given a sequence of random variables $\{v_n\}$ and an additional random variable v defined on a probability space $(\Omega, \mathcal{F}, P)$, we say that $v_n \to v$*

(a) uniformly, if $\sup_{\omega\in\Omega} |v_n(\omega) - v(\omega)| \to 0$*;*
(b) surely, if $v_n(\omega) - v(\omega) \to 0$, $\forall\omega \in \Omega$*;*
(c) almost surely, if $P\{\omega \text{ such that } v_n(\omega) - v(\omega) \to 0\} = 1$ *(when we want to emphasize probability P we write P-almost surely.) Another expression equivalent to 'almost surely' is 'with probability 1';*
(d) in L^2*, if* $E[(v_n - v)^2] \to 0$*;*
(e) in L^1*, if* $E[|v_n - v|] \to 0$*;*
(f) in probability, if $\forall\epsilon > 0$, $P\{\omega \text{ such that } |v_n(\omega) - v(\omega)| \geq \epsilon\} \to 0$*;*
(g) weakly, if, for any continuous and bounded function $f : \mathbb{R} \to \mathbb{R}$, $E[f(v_n)] \to E[f(v)]$. *'$v_n \to v$ weakly' is also expressed as '$v_n \to v$ in distribution'. When v_n converges in distribution to a variable with distribution F, we also write $v_n \sim AsF$.* □

Definitions (a)-(f) regards the behavior of $v_n - v$ and requires that this difference goes somehow to zero as $n \to \infty$. Thus, for instance, v_n tends to v almost surely if $v_n - v$ tends to zero almost surely. In contrast, the fact that $v_n \to v$ weakly in no ways implies that $v_n - v \to 0$. Suppose for instance that $v_n = \xi$, $n = 1, 2, \ldots$, where ξ is a fixed

random variable different from v but sharing with v the same distribution. Then clearly $E[f(v_n)] = E[f(\xi)] = E[f(v)]$, $\forall n$, so that $v_n \to v$ weakly, but $v_n - v = \xi - v$ does not converge in any sense.

Weak convergence is in fact a property of the image probabilities. Re-writing $E[f(v_n)] \to E[f(v)]$ as $\int_{\mathbb{R}} f dF_n \to \int_{\mathbb{R}} f dF$ (where F_n is the distribution of v_n and F that of v), we see that $v_n \to v$ weakly corresponds to say that F_n soemhow approaches F.

The different notions of convergence are related to each other by the following theorem.

THEOREM A.2 *The implications shown in Figure A.1 hold true.*

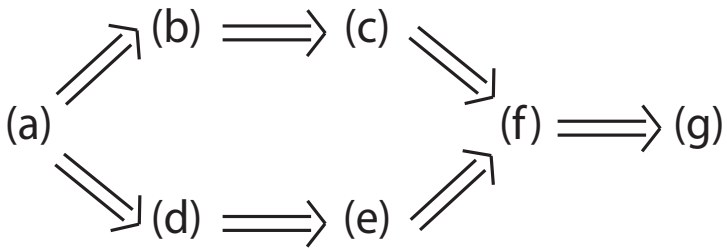

Figure A.1: Implications among convergence properties

PROOF. Implications $(a) \Rightarrow (b) \Rightarrow (c)$ and $(a) \Rightarrow (d)$ are obvious.

Let $\xi_n := v_n - v$.

(d) $\Rightarrow$ (e)) By Schwarz inequality: $E[|\xi_n|] = E[1 \cdot |\xi_n|] \leq (E[1^2])^{1/2}(E[\xi_n^2])^{1/2} = (E[\xi_n^2])^{1/2}$, so that $E[\xi_n^2] \to 0$ implies $E[|\xi_n|] \to 0$.
(c) $\Rightarrow$ (f)) Let $A_n^\epsilon := \{\omega \; such \; that \; |\xi_k| < \epsilon, \forall k \geq n\}$. From (c) we have that $A_n^\epsilon \uparrow A^\epsilon$ with $P(\Omega - A^\epsilon) = 0$. Since $\{\omega \; such \; that \; |\xi_n| \geq \epsilon\} \subseteq \Omega - A_n^\epsilon$, it follows that $P\{\omega \; such \; that \; |\xi_n| \geq \epsilon\} \to 0$.
(e) $\Rightarrow$ (f)) From (e), $\epsilon P\{\omega \; such \; that \; |\xi_n| \geq \epsilon\} \leq E[|\xi_n|] \to 0$ and (f) follows.
(f) $\Rightarrow$ (g)) Given $\epsilon_1, \epsilon_2 > 0$, fix M and ϵ such that $P\{\omega \; such \; that \; |v| \geq M\} \leq \epsilon_1$ and $|f(x) - f(y)| \leq \epsilon_2$ for $|y| < M$ and $|x - y| < \epsilon$ (ϵ exists since a continuous function is uniformly continuous in a bounded set.) Then,

$$
\begin{aligned}
|E[f(v_n)] - E[f(v)]| &\leq E\left[|f(v_n) - f(v)|\right] && \text{(A.1)} \\
&\leq (2 \max_x |f(x)|)\epsilon_1 && \text{(A.2)} \\
&\quad + (2 \max_x |f(x)|) P\{\omega \; such \; that \; |\xi_n| > \epsilon\} && \text{(A.3)} \\
&\quad + \epsilon_2. && \text{(A.4)}
\end{aligned}
$$

Since ϵ_1 and ϵ_2 are arbitrarily chosen and $P\{\omega \; such \; that \; |\xi_n| > \epsilon\} \to 0$ by assumption, (g) follows. □

No other implications than the ones stated in Theorem A.2 hold true. In particular, almost sure convergence does not imply and is not implied by L^2-convergence, as the following example shows.

EXAMPLE A.3 *Consider the following sequence of random variables defined on the probability space* $([0,1], \mathcal{B}[0,1], \lambda)$*:*

$$v_n = \begin{cases} \sqrt{n}, & on \; [0, 1/n] \\ 0, & otherwise. \end{cases} \tag{A.5}$$

Clearly, $v_n \to v = 0$ *almost surely, but* $E[(v_n - v)^2] = \frac{1}{n} n = 1 \not\to 0$.

Instead, consider the sequence $v_1^1, v_2^1, v_2^2, v_3^1, v_3^2, v_3^3, \ldots$ *with*

$$v_n^k = \begin{cases} 1, & on \; [\frac{k-1}{n}, \frac{k}{n}] \\ 0, & otherwise. \end{cases} \tag{A.6}$$

This sequence is L^2*-convergent to zero, but does not converge at any point of* $[0,1]$. □

In (A.6) the idea is to construct a sequence of intervals with two properties: i) their size shrinks (so that L^2 convergence to zero holds); and ii) each point in $[0,1]$ falls infinitely many times in the intervals (and, thus, almost sure convergence fails.) The latter property is possible since the sum of the interval sizes $1, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{1}{3}, \frac{1}{3}, \ldots$ where variables v_n^k takes on value 1 diverges. The following theorem, which is important to assess almost sure convergence in many contexts, shows that, when this sum is finite, almost sure convergence does take place.

THEOREM A.4 *Let* v_n*,* $n = 1, 2, \ldots$*, and* v *be random variables. Suppose that for any* $\epsilon > 0$*,*

$$\sum_{n=1}^{\infty} P\{\omega \; such \; that \; |v_n - v| \geq \epsilon\} < \infty, \tag{A.7}$$

then $v_n \to v$ *almost surely.*

PROOF. Letting

$$A_j^k := \{\omega \; such \; that \; |v_n - v| \geq \frac{1}{k}, \quad \forall n \geq j\}, \tag{A.8}$$

we have $\{\omega \; such \; that \; v_n - v \not\to 0\} = \cup_{k=1}^{\infty} \cap_{j=1}^{\infty} A_j^k$. Thus,

$$P\{\omega \; such \; that \; v_n - v \not\to 0\} \leq \sum_{k=1}^{\infty} \lim_{j \to \infty} \sum_{n=j}^{\infty} P\{\omega \; such \; that \; |v_n - v| \geq \frac{1}{k}\}. \tag{A.9}$$

Since (A.7) holds, each single term $\lim_{j\to\infty} \sum_{n=j}^{\infty} P\{\omega \; such \; that \; |v_n - v| \geq \frac{1}{k}\}$ is zero and the right hand side of the previous inequality is null, so proving that $v_n \to v$ almost surely. □

Finally, we give a result relating the measurability properties of a sequence of random variables to those of its limit.

THEOREM A.5 *Let $\{v_n\}$ be a sequence of random variables on $(\Omega, \mathcal{F}, P)$ (i.e. each v_n is $\mathcal{F}$-measurable) and let v be an additional variable that is not required to be $\mathcal{F}$-measurable by assumption.*

(i) if $v_n(\omega) \to v(\omega)$, $\forall \omega \in \Omega$ (up to this point, this is not exactly sure convergence as it is given in Definition A.1 since here v is not required to be a random variable), then v is a random variable (and, thus, $v_n \to v$ surely, according to Definition A.1);

(ii) if $\{v_n \not\to v\}$ is measurable and $P\{v_n \not\to v\} = 0$ (similarly to (i), this is not exactly almost sure convergence since v is not required to be a random variable), then v need not be a random variable. However, $\bar{v}$ so defined

$$\bar{v} = \begin{cases} v, & where \; v_n \to v \\ 0, & otherwise \end{cases} \tag{A.10}$$

is a random variable;

(iii) if v is a random variable, $v_n \to v$ in probability (or, of course, in any other sense that implies convergence in probability, e.g. L^2-convergence), and v_n are $\mathcal{G}$-measurable for some $\mathcal{G} \subseteq \mathcal{F}$, then there exist a $\mathcal{G}$-measurable $\bar{v}$ such that $P\{\bar{v} \neq v\} = 0$.

The idea underlying (i) is that v has to behave well (i.e. it need be $\mathcal{F}$-measurable) since it is the limit of $\mathcal{F}$-measurable variables. In point (ii), $P\{v_n \not\to v\} = 0$ leaves open the possibility that v exhibits a wired behavior on the set where $v_n \not\to v$, so losing the measurability property of v. As an alternative to construct $\bar{v}$, one can consider complete σ-algebras (a σ-algebra $\mathcal{F}$ is complete when all subsets of sets in $\mathcal{F}$ with zero probability are also in $\mathcal{F}$), for which v itself is measurable. Considering complete σ-algebras is not restrictive since, given a σ-algebra, one can always make it complete by augmenting it with all subsets of measurable sets with zero probability. We do not dwell further on this and refer the interested reader to standard textbooks on probability for details. In (iii), v is required to be a random variable in the beginning since, otherwise, $P\{\omega \; such \; that \; |v_n - v| \geq \epsilon\}$ that appears in the definition of convergence in probability may not be well defined. Note that, since L^2-convergence and almost sure convergence imply convergence in probability, result (iii) can be applied to these types of convergence as well.

PROOF. (i) For any $a, b \in \mathbb{R}$,

$$\{\omega \; such \; that \; v \in (a,b)\} = \cup_{p=1}^{\infty} \cap_{n \geq p} \{\omega \; such \; that \; v_n \in (a,b)\}. \tag{A.11}$$

Since $\{\omega \; such \; that \; v_n \in (a,b)\} \in \mathcal{F}$ and a σ-algebra is closed under countable intersection and union, we have that $\{\omega \; such \; that \; v \in (a,b)\} \in \mathcal{F}$, from which the measurability of v follows.

(ii) It is a simple variant of the proof of (i) and is left as an exercise.

(iii) Fix a sequence of real numbers $\epsilon_k > 0$ such that $\epsilon_k \to 0$ and extract from $\{v_n\}$ a subsequence $\{v_{n_k}\}$ such that

$$\sum_{k=1}^{\infty} P\{\omega \; such \; that \; |v_{n_k} - v| \geq \epsilon_k\} < \infty, \tag{A.12}$$

(such a sequence exists since $v_n \to v$ in probability.) By applying Theorem A.4, we then conclude that $v_{n_k} \to v$ almost surely and the thesis follows along the line of point (ii). □

A.2 Limit under the sign of expectation

Suppose that $v_n \to v$ almost surely. Under what conditions is it true that $E[v_n] \to E[v]$? The following theorems provide an answer.

THEOREM A.6 (monotone convergence) *Let $v_n, n = 1, 2, \ldots$, and v be random variables such that $v_n \uparrow v$ almost surely (i.e. $v_n(\omega)$ is increasing and tends to $v(\omega)$ for*

almost all ω's), and $v_n \geq z, n = 1, 2, \ldots$, for some random variable z with $E[z] > -\infty$. Then,

$$E[v_n] \uparrow E[v]. \tag{A.13}$$

□

THEOREM A.7 (dominated convergence) *Let $v_n, n = 1, 2, \ldots$, and v be random variables such that $v_n \to v$ almost surely, and $|v_n| \leq z, n = 1, 2, \ldots$, for some random variable z with $E[z] < \infty$. Then,*

$$E[v_n] \to E[v]. \tag{A.14}$$

□

A proof of these theorems can be found in any textbook on probability.

In the statements of the theorems, two types of conditions are present: v_n is required to approach v; and v_n is somehow uniformly bounded. The latter condition serves the purpose to limit the importance of the mismatch between v_n and v on events of small probability. An example clarifies the matter.

EXAMPLE A.8 (Example A.3 continued) *Consider again the v_n's in (A.5). Clearly, $v_n^2 \to 0$ almost surely, but $E[v_n^2] = 1 \not\to E[0] = 0$. Here, no dominating z exists with $E[z] < \infty$, so that the conditions of Theorem A.7 are violated.* □

A.3 Convergence results for independent random variables

We commence by proving probabilistic inequalities. Besides being in use later in this section when proving convergence results, they bear an interest in their own right.

MARKOV'S INEQUALITY

For any nonnegative random variable v and $\epsilon > 0$,

$$P\{v \geq \epsilon\} \leq \frac{E[v]}{\epsilon}. \tag{A.15}$$

□

Markov's inequality is elementary to prove: $E[v] = \int_\Omega v dP \geq \int_{\{v \geq \epsilon\}} v dP \geq \epsilon P\{v \geq \epsilon\}$. An application of Markov's inequality gives

CHEBYSHEV'S INEQUALITY

For any $\epsilon > 0$,

$$P\{|v| \geq \epsilon\} \leq \frac{E[v^2]}{\epsilon^2}. \tag{A.16}$$

PROOF.

$$\begin{aligned} P\{|v| \geq \epsilon\} &= P(v^2 \geq \epsilon^2) && \text{(A.17)} \\ &\leq \frac{E[v^2]}{\epsilon^2} \quad (using\ (A.15)) && \text{(A.18)} \end{aligned}$$

□

In Markov's inequality, the idea is to lowerbound $E[v]$ by squeezing the tail ov v to the value ϵ. Thus, the bound is tight only when the tail rapidly vanishes after ϵ. A similar observation applies to Chebyshev's inequality. Better bounds can be found by redressing the random variable distribution through some transformation before Markov's inequality is applied. One such example is given by the following inequality of Chernoff. In this inequality, s is a free parameter that can be used to tune the distribution redressing and an example of use of s is found in the proof of Hoeffding's inequality (Theorem A.10.)

CHERNOFF'S INEQUALITY

For any $s > 0$ and $\epsilon > 0$,

$$P\{v \geq \epsilon\} \leq \frac{E[e^{sv}]}{e^{s\epsilon}}. \tag{A.19}$$

PROOF.

$$\begin{aligned} P\{v \geq \epsilon\} &= P\{e^{sv} \geq e^{s\epsilon}\} && \text{(A.20)} \\ &\leq \frac{E[e^{sv}]}{e^{s\epsilon}}. && \text{(A.21)} \end{aligned}$$

□

Concentration inequalities

Consider a sequence of independent random variables $v_k, k = 1, 2, \ldots$. Concentration inequalities study how a function $f(v_1, v_2, \ldots, v_n)$ concentrates around its expected value $E[f(v_1, v_2, \ldots, v_n)]$. Here, we are mainly concerned with the deviation of sums of random variables from their means, that is our interest is on function $f(v_1, v_2, \ldots, v_n) = \frac{1}{n}\sum_{k=1}^{n} v_k$ and we study the behavior of

$$S_n := \frac{1}{n}\sum_{k=1}^{n} v_k - E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right]. \tag{A.22}$$

A first bound is obtained by means of Chebyshev's inequality:

$$P\{|S_n| \geq \epsilon\} \leq \frac{E[S_n^2]}{\epsilon^2} = \frac{\frac{1}{n^2}\sum_{k=1}^{n} var(v_k)}{\epsilon^2}. \tag{A.23}$$

EXAMPLE A.9 *For an independent and identically distributed sequence of Bernoulli random variables (i.e. $P\{v_k = 1\} = 1 - P\{v_k = 0\} = p$), from (A.23) we have*

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} v_k - p\right| \geq \epsilon\right\} \leq \frac{p(1-p)}{n\epsilon^2}. \tag{A.24}$$

□

Do we expect that bound (A.23) is tight? Remember that Chebishev's inequality is tight when the distribution tail vanishes rapidly after ϵ. On the other hand, applying the central limit theorem leads to the conclusion that, under mild assumptions, the distribution of S_n tends weakly to a Gaussian, a long-tailed distribution. For example, in the case of the Bernulli sequence of Example A.9, letting $\Phi(x) = \int_{-\infty}^{x}(2\pi)^{-1/2}e^{-r^2/2}dr$, the central limit theorem states that

$$P\left\{\sqrt{\frac{n}{p(1-p)}}\left(\frac{1}{n}\sum_{k=1}^{n} v_k - p\right) \geq x\right\} \to 1 - \Phi(x) \leq \frac{e^{-x^2/2}}{x\sqrt{2\pi}}, \tag{A.25}$$

from which we would expect something like

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} v_k - p\right| \geq \epsilon\right\} \sim e^{-\frac{n\epsilon^2}{2p(1-p)}}, \tag{A.26}$$

that is the probability decays exponentially with n. The gap between linear (see (A.23)) and exponential convergence in n can be filled in by resorting to Chernoff's inequality. This leads to Hoeffding's inequality.

THEOREM A.10 (Hoeffding's inequality) *Let $v_k, k = 1, 2, \ldots$, be independent bounded random variables taking value in $[\alpha_k, \beta_k]$ and let S_n be defined as in (A.22). Then,*

$$P\{S_n \geq \epsilon\} \leq e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^{n}(\beta_k - \alpha_k)^2}}; \tag{A.27}$$

and

$$P\{S_n \leq -\epsilon\} \leq e^{-\frac{2n^2\epsilon^2}{\sum_{k=1}^{n}(\beta_k - \alpha_k)^2}}. \tag{A.28}$$

PROOF. For ease of notation, we assume $[\alpha_k, \beta_k] = [0, 1]$, in which case we prove that

$$P\{S_n \geq \epsilon\} \leq e^{-2n\epsilon^2}; \tag{A.29}$$

and

$$P\{S_n \leq -\epsilon\} \leq e^{-2n\epsilon^2}. \tag{A.30}$$

The extension is easy.

We start by observing that for any random variable v with $E[v] = 0$ and $\alpha \leq v \leq 1 + \alpha$, and any $h > 0$, we have

$$E[e^{hv}] \leq e^{\frac{h^2}{8}}. \tag{A.31}$$

In fact, by convexity of the exponential function, $e^{hv} \leq (v - \alpha)e^{(1+\alpha)h} + (1 + \alpha - v)e^{\alpha h}$, so that

$$\begin{aligned} E[e^{hv}] &\leq E\left[(v - \alpha)e^{(1+\alpha)h} + (1 + \alpha - v)e^{\alpha h}\right] && \text{(A.32)} \\ &= E\left[-\alpha e^{(1+\alpha)h} + (1 + \alpha)e^{\alpha h}\right] \quad (since\ E[v] = 0) && \text{(A.33)} \\ &= -\alpha e^{(1+\alpha)h} + (1 + \alpha)e^{\alpha h} && \text{(A.34)} \\ &= e^{\Gamma(h)}, && \text{(A.35)} \end{aligned}$$

where $\Gamma(h) = \alpha h + \ln(1 + \alpha - \alpha e^h)$. The derivative of $\Gamma(h)$ is $\Gamma'(h) = \alpha - \alpha/[(1 + \alpha)e^{-h} - \alpha]$, so that $\Gamma'(0) = 0$. Moreover,

$$\begin{aligned}
\Gamma''(h) &= \frac{-\alpha(1+\alpha)e^{-h}}{[(1+\alpha)e^{-h}-\alpha]^2} && \text{(A.36)}\\
&= \frac{ba}{[a+b]^2} \quad (with\ a=(1+\alpha)e^{-h}, b=-\alpha) && \text{(A.37)}\\
&\leq \frac{1}{4}, \quad \forall h. && \text{(A.38)}
\end{aligned}$$

Thus, by Taylor series expansion, for some $\xi \in [0,h]$:

$$\Gamma(h) = \Gamma(0) + \Gamma'(0)h + \frac{1}{2}\Gamma''(\xi)h^2 \leq \frac{h^2}{8}, \tag{A.39}$$

which, used in (A.35), yields (A.31).

Thanks to (A.31), (A.29) is now easily obtained from Chernoff's inequality:

$$\begin{aligned}
P\{S_n \geq \epsilon\} &\leq \frac{E[e^{sS_n}]}{e^{s\epsilon}} \quad (using\ Chernoff's\ inequality) && \text{(A.40)}\\
&= \frac{E[e^{s\frac{1}{n}\sum_{k=1}^{n}(v_k - E[v_k])}]}{e^{s\epsilon}} && \text{(A.41)}\\
&= \frac{\Pi_{k=1}^{n} E[e^{s\frac{1}{n}(v_k - E[v_k])}]}{e^{s\epsilon}} \quad (by\ the\ independence\ of\ the\ v_k's) && \text{(A.42)}\\
&\leq \frac{\Pi_{k=1}^{n} e^{\frac{s^2}{8n^2}}}{e^{s\epsilon}} \quad (using\ (A.31)\ with\ h=s/n) && \text{(A.43)}\\
&\leq e^{-2n\epsilon^2}. \quad (by\ choosing\ s=4\epsilon n) && \text{(A.44)}
\end{aligned}$$

Equation (A.30) is obtained in the same way by considering $P\{S_n \leq -\epsilon\} = P\{-S_n \geq \epsilon\}$ instead of $P\{S_n \geq \epsilon\}$. □

EXAMPLE A.11 (Example A.9 continued) *Using Hoeffding's inequality yields*

$$P\left\{\left|\frac{1}{n}\sum_{k=1}^{n} v_k - p\right| \geq \epsilon\right\} \leq 2e^{-2n\epsilon^2} \tag{A.45}$$

(compare with (A.26).) □

Hoeffding's inequality deals specifically with empirical means, showing that the empirical mean concentrates around the true mean value. The reason why this is so is that in the empirical mean each single variable has a moderate impact in determining the empirical mean value and, moreover, different variables do not cooperate because they are independent.

It is a fact that this result can be extendend to general functions provided that each variable has a marginal importance in determining the total value of the function, as stated in the following theorem (for a proof of the theorem, see the references given after its statement.)

THEOREM A.12 (the bounded difference inequality) *Let $v_k, k = 1, 2, \ldots$, be independent random variables taking value in a set A and assume that $\forall x_1, \ldots, x_n \in A$, $x'_k \in A$, and $\forall k \in [1, n]$, the measurable function $f : \mathbb{R}^n \to \mathbb{R}$ satisfies the condition:*

$$|f(x_1, \ldots, x_{k-1}, x_k, x_{k+1}, \ldots, x_n) - f(x_1, \ldots, x_{k-1}, x'_k, x_{k+1}, \ldots, x_n)| \leq \gamma_k. \quad \text{(A.46)}$$

Then,

$$P\{f(v_1, v_2, \ldots, v_n) - E[f(v_1, v_2, \ldots, v_n)] \geq \epsilon\} \leq e^{-\frac{2\epsilon^2}{\sum_{k=1}^n \gamma_k^2}}; \quad \text{(A.47)}$$

and

$$P\{f(v_1, v_2, \ldots, v_n) - E[f(v_1, v_2, \ldots, v_n)] \leq -\epsilon\} \leq e^{-\frac{2\epsilon^2}{\sum_{k=1}^n \gamma_k^2}}; \quad \text{(A.48)}$$

□

Note that (A.47) and (A.48) reduce to (A.27) and (A.28) when we consider empirical means.

Bibliographical notes

Hoeffding's inequality has been originally proven in [12], while, for Bernoulli sequences, it appeard early in [9]. An elegant proof of the bounded difference inequality using martingale techniques is found in [16] while a proof can be also found e.g. in [13].

Laws of large numbers

The laws of large numbers study the convergence of $\frac{1}{n}\sum_{k=1}^n v_k$ to $E\left[\frac{1}{n}\sum_{k=1}^n v_k\right]$. This is probably the most studied problem in probability theory and the literature offers an abundant supply of results under varying assumptions and according to different notions of convergence. Here, we only present a standard result and prove it by means of concentration inequalities.

THEOREM A.13 (law of large numbers) *Let $v_k, k = 1, 2, \ldots$, be independent random variables with uniformly bounded variance: $var(v_k) \leq C, \forall k$. Then,*

$$\frac{1}{n}\sum_{k=1}^{n} v_k \to E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right], \quad almost\ surely. \tag{A.49}$$

Before proving the theorem, we would like to note that if $v_k \in [\alpha, \beta]$, $\forall k$ (that is the v_k's are uniformly bounded), then the result is immediate. In fact, from Hoeffding's inequality,

$$P\{|S_n| \geq \epsilon\} \leq 2e^{-\frac{2n\epsilon^2}{(\beta-\alpha)^2}}, \tag{A.50}$$

where $S_n := \frac{1}{n}\sum_{k=1}^{n} v_k - E\left[\frac{1}{n}\sum_{k=1}^{n} v_k\right]$, so that $S_n \to 0$ almost surely follows from Theorem A.4. On the other hand, if we only assume the boundedness of the variance of the v_k's (as is the case in the theorem), Hoeffding's inequality does not apply. Resorting instead to (A.23), we only conclude that

$$P\{|S_n| \geq \epsilon\} \leq \frac{C}{n\epsilon^2}, \tag{A.51}$$

which is not enough to prove that $S_n \to 0$ almost surely by way of Theorem A.4 since $\sum_{n=1}^{\infty} \frac{C}{n\epsilon^2} = \infty$. The proof of the theorem given below suggests a way to get around this difficulty.

Instead of directly proving the theorem, we prefer to state the following lemma, from which the theorem immediately follows, because the lemma is useful in other contexts as well.

LEMMA A.14 *Consider the doubly indexed set of random variables S_r^p such that $S_r^p = 0$ for $r > p$ and assume that $E[(S_r^p)^2] \leq C(p + 1 - r)$ for some constant C and that, for $m < n$, $|S_1^n| \leq |S_1^m| + |S_{m+1}^n|$. Then,*

$$\frac{1}{n}S_1^n \to 0, \quad almost\ surely. \tag{A.52}$$

Note that Theorem A.13 immediately follows from the lemma by the position $S_r^p := \sum_{k=r}^{p}(v_k - E[v_k])$.

PROOF OF THE LEMMA. Given an integer n, let N be the integer such that $N^2 \leq n < (N+1)^2$ and write:

$$\left|\frac{1}{n}S_1^n\right| \leq \frac{1}{N^2}\left|S_1^{N^2}\right| + \frac{1}{N^2}\left|S_{N^2+1}^n\right|. \tag{A.53}$$

The lemma is proven by showing that both terms in the last expression go to zero almost surely.

As for the first term, by the Chebishev's inequality we have

$$\sum_{N=1}^{\infty} P\left\{\frac{1}{N^2}\left|S_1^{N^2}\right| \geq \epsilon\right\} \leq \sum_{N=1}^{\infty} \frac{var\left(\frac{1}{N^2}\left|S_1^{N^2}\right|\right)}{\epsilon^2} \tag{A.54}$$

$$\leq \sum_{N=1}^{\infty} \frac{CN^2}{N^4\epsilon^2} \tag{A.55}$$

$$< \infty, \tag{A.56}$$

so that almost sure convergence to zero follows from Theorem A.4. Turn now to the second term in (A.53). We have

$$\sum_{n=1}^{\infty} P\left\{\frac{1}{N^2}\left|S_{N^2+1}^{n}\right| \geq \epsilon\right\} \leq \sum_{n=1}^{\infty} \frac{var\left(\frac{1}{N^2}\left|S_{N^2+1}^{n}\right|\right)}{\epsilon^2} \tag{A.57}$$

$$\leq \sum_{n=1}^{\infty} \frac{C(n-N^2)}{N^4\epsilon^2} \tag{A.58}$$

$$\leq \sum_{N=1}^{\infty} \sum_{n=N^2}^{(N+1)^2-1} \frac{C(n-N^2)}{N^4\epsilon^2} \tag{A.59}$$

$$\leq \sum_{N=1}^{\infty} ((N+1)^2 - N^2) \frac{C((N+1)^2 - 1 - N^2)}{2N^4\epsilon^2} \tag{A.60}$$

$$\leq \sum_{N=1}^{\infty} (2N+1) \frac{C2N}{2N^4\epsilon^2} \tag{A.61}$$

$$< \infty \tag{A.62}$$

and, again, Theorem A.4 can be resorted to to prove almost sure convergence to zero. □

Bibliography

[1] D. Angluin. Queries and concept learning. *Machine Learning*, 2:319–342, 1992.

[2] M. Anthony, N. Biggs, and J. Shawe-Taylor. The learnability of formal concepts. In *Proc. 3rd Workshop on Computational Learning Theory*, pages 246–257, 1990.

[3] P.L. Bartlett, P.M. Long, and R.C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proc. 7th ACM Conf. on Computational Learning Theory*, pages 299–310, 1994.

[4] J.B. Moore B.D.O. Anderson and R.M. Hawkes. Model approximation via prediction error identification. *Automatica*, 14:615–622, 1978.

[5] G.M. Benedek and A. Itai. Learnability by fixed distribution. In *Proc. 1st Workshop on Computational Learning Theory*, pages 80–90, 1988.

[6] P. Billingsley. *Probability and measure.* John Wiley and Sons, 1995.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[8] P.E. Caines. Stationary linear and nonlinear system identification and predictor set completeness. *IEEE Trans. on Automatic Control*, AC-23:583–594, 1978.

[9] H. Chernoff. A measure of asymptotic efficiency of tests of a hypothesis based on the sum of observations. *Annal of Mathematical Statistics*, 23:493–507, 1952.

[10] R.M. Dudley. Universal donsker classes and metric entropy. *Annals of Probability*, 15:1306–1326, 1984.

[11] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.

[12] W. Hoeffding. Probability inequalities for sums of bounded random variables. *J. of American Statistical Association*, 58:13–30, 1963.

[13] L.Devroye, L.Gyorfi, and G.Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New York, 1996.

[14] L. Ljung. Convergence analysis of parametric identification methods. *IEEE Trans. on Automatic Control*, AC-23:770–783, 1978.

[15] L. Ljung. *System identification. Theory for the user.* Prentice-Hall, Englewood Cliffs, NJ, 2nd edition, 1999.

[16] C. McDiarmid. *On the method of bounded differences. In: Surveys in Conbinatorics.* Cambridge University Press, 1989.

[17] B.K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[18] J.M. Parrondo and C. van den Broeck. Vapnik-chervonenkis bounds for generalization. *J. Phys.*, 26:2211–2223, 1993.

[19] D. Pollard. *Empirical processes: theory and applications.* Regional Conference Series in Probability and Statistics. Institute of Mathematical Statistics, Volume 2, 1990.

[20] M. Prandini and M.C. Campi. Adaptive lqg control of input-output ystems - a cost biased approach. *SIAM J. Control and Optim.*, 36:1890–1907, 1998.

[21] R. Ranga Rao. Relations between weak and uniform convergence of measures with applications. *Annal of Mathematical Statistics*, 33:659–680, 1962.

[22] N. Sauer. On the densities of family of sets. *J. of Combin. Theory*, 13:145–147, 1972.

[23] J.M. Steele. Empirical discepancies and subadditive processes. *J. Phys.*, 6:118–127, 1978.

[24] L.G. Valiant. A theory of the learnable. *Comm. ACM*, 27:1134–1142, 1984.

[25] V.N. Vapnik. *Statistical learning theory.* Wiley, New York, 1998.

[26] V.N. Vapnik and A.Y. Chervonenkis. Uniform convergence of the frequencies of occurrence of events to their probabilities. *Soviet Math. Doklady*, 9:915–918, 1968.

[27] V.N. Vapnik and A.Y. Chervonenkis. On the uniform convergence of relative frequencies to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.

[28] V.N. Vapnik and A.Y. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, 26:532–553, 1981.

[29] M. Vidyasagar. *A theory of learning and generalization with applications to neural networks and control systems.* Springer-Verlag, London, 1997.