On Minimax Prediction for Nonparametric Autoregressive Models

Anatoli Juditsky[*] Bernard Delyon[†]

2000

Abstract

We consider the problem of nonparametric prediction for a multi-dimensional functional autoregression $y_t = f(y_{t-1}, ..., y_{t-d}) + e_t$ on the basis of N observations of y_t. In the case when the unknown nonlinear function f belongs to the Barron class, we propose an estimation algorithm which provides approximations of f with expected L_2 accuracy $O(N^{1/4}\ln^{1/4} N)$. We also show that this approximation rate cannot be significantly improved.

The proposed algorithms are "computationally efficient" – the total number of elementary computations necessary to complete the estimate grows polynomially with N.

1 Introduction

We address the following *prediction problem*: we are interested to predict the future value y_{N+1} of nonlinear autoregressive process (y_t) given N observations

$$y_t = f(y_{t-1}, ..., y_{t-d}) + e_t,\ y = (y_0, ..., y_{-d+1})^T \in \boldsymbol{R}^d,\ \ t = 1, 2, ..., N, \tag{1}$$

Here $f(x) : \boldsymbol{R}^d \to \boldsymbol{R}$ is a bounded Borel real-valued function of d real variables), $e_t \in \boldsymbol{R}$ are independent and identically distributed gaussian random variables such that

$$E\{e_1\} = 0,\ E\{e_1^2\} = \sigma_e^2 < \infty. \tag{2}$$

By analogy with the corresponding parametric model we call the above system a *non-parametric autoregression* or a *functional autoregression* of dimension d (NAR(d)). Again, as in the case of parametric autoregression, we measure the quality of an estimate $\widehat{y}_{N+1}$ of y_{N+1} by its squared L_2-error:

$$E(\widehat{y}_{N+1} - y_{N+1})^2. \tag{3}$$

Let us denote $Y_t = (y_t, ..., y_{t-d+1})^T$. Due to the independence of e_{N+1} of the totality $(y, ..., y_N)$ for any $N \geq 1$, one can easily see that this error coincides with

$$\sigma_e^2 + E(\widehat{y}_{N+1} - f(Y_N))^2.$$

Thus our initial objective – to minimize (3) with respect to $\widehat{y}_{N+1}$– can be reformulated as follows: find a $(y, ..., y_N)$-measurable function $\widehat{f}(\cdot)$ which minimizes

$$E(\widehat{f}(y, ..., y_N) - f(Y_N))^2. \tag{4}$$

For the sake of simplicity we put the dependence on previous observations in the index and use the notation $\widehat{f}_N(Y_N)$ rather than $\widehat{f}(y, ..., y_N)$. We can now call $\widehat{f}_N(\cdot)$ the estimate of f.

[*] INRIA Rhone-Alpes, 655 avenue de l'Europe, 38330 MONTBONNOT SAINT MARTIN, FRANCE

[†] IRISA-INRIA, Campus de Beaulieu, 35042 RENNES Cedex, FRANCE

1.1 Barron's class

As we shall see below the problem of minimizing (4) is fairly close to that of estimating a regression function f from the model

$$y_t = f(x_t) + e_t,\ t = 1, ..., N \tag{5}$$

where x_t are independent and identically distributed with some distribution μ. The majority of the known estimates of multivariate regression functions, see [4] and references therein, are aimed to restore smooth signals. It is well known that in this case the rates of convergence degrade rather fast when the dimensionality d of f increases[1] and become exceedingly slow when d approaches $\ln N$.

There are basically two ways to overcome the indicated difficulty (known as the "curse of dimensionality"): either to accept that a huge amount of data is necessary or to strengthen restrictions on the function class in order to bound its "effective dimension". One such class which allows to reduce the "effective dimension" has been recently introduced by A. Barron [1] and can formalized as follows:

Definition 1 *$\mathcal{F}$ consists of functions f which are the Fourier transform*

$$f(x) = \int_{\boldsymbol{R}^d} \widehat{f}(\omega) \exp\{i\omega^T x\} d\omega \tag{6}$$

of a finite complex-valued measure (a function from the centered at the origin L_1-ball of the radius $L < \infty$):

$$\int_{\boldsymbol{R}^d} |\widehat{f}(\omega)| d\omega \le L < \infty. \tag{7}$$

The interest of this definition is explained by the approximation result, proved in [1]. When reformulated for our purposes it can be stated as follows: for any positive integer n and for any probability distribution μ on $\boldsymbol{R}^d$ there exists an n-tuple $(\omega_1, ..., \omega_n)$ and coefficients $\lambda_1, ..., \lambda_n$, $\sum_{k=1}^n |\lambda_k| \le L$, such that the combination

$$f_n(x) = \sum_{k=1}^n \lambda_k \exp\{i\omega_k^T x\}$$

satisfies

$$\int |f(x) - f_n(x)|^2 \mu(dx) \le L^2/n. \tag{8}$$

Although this proposition states that the quality of a "simple" approximation of a function f from the class given by (6) and (7) admits a bound (8) independent of the dimension d, this is an existence result only. Indeed, it is unclear how to recover the frequencies ω_k and the weights λ_k when the available information on f is given by reasonably many observations (1). The following solution of this problem was proposed in [6] to solve the regression problem (5): in order to use the indicated existence theorem we can act as follows: let Ω be a "fine" grid in the space of frequencies; we consider the functional system $f_k(x) = L \exp\{i\omega_k^T x\}$, $\omega_k \in \Omega$ with the cardinality M, and use observations (1) to solve the following optimization problem:

$$\text{minimize} \int (f(x) - \sum_{k=1}^M \lambda_k f_k(x))^2 \mu(dx)$$

under constraint that $\lambda = (\lambda_1, ..., \lambda_M)^T$ belongs to the $\|\cdot\|_1$-ball Λ:

$$\Lambda = \{\{\lambda_\omega\}_{\omega\in\Omega} \mid \sum_{\omega\in\Omega} |\lambda_\omega| \le 1\}. \tag{9}$$

Surprisingly enough, this approach, under minor additional assumptions on f, allows to provide computationally efficient procedures to recover f with basically the same quality as that stated in Barron's

[1] For example, the rate is $O(N^{-1/(2+d)})$ for Lipschitz continuous functions f.

existence theorem. What we are up to do now is to show that the same technique gives optimal results in the nonparametric prediction problem above.

The paper is organized as follows: in Section 2 we present an algorithm of stochastic approximation type for estimation of multi-dimensional parameter under constraint (9). This algorithm then is used in Section 3 to estimate a nonlinear autoregressive model from observations (1). We also provide a minimax lower bound for the estimation problem in question which shows that the proposed estimate is optimal up to a multiplicative constant.

2 Stochastic approximation algorithm

In this section we present a robust algorithm to estimate a multi-dimensional regression parameter under convex constraints. This algorithm will be used in the estimation procedure in the next section, however, it is of interest by itself.

We consider the following regression model

$$y_t = \phi_t \lambda^* + e_t + b_t, \quad t = 1, \dots , \tag{10}$$

where $\phi_t \in \boldsymbol{R}^M$ and $y_t \in \boldsymbol{R}$ are the observable regressors and outputs respectively, e_t and b_t are unobservable disturbances; we suppose that $\|\lambda^*\|_1 \leq 1$.

Note that as $M \gg N$ we cannot use the least squares algorithm because the matrix $\Phi_N = \sum_{i=1}^N \phi_i \phi_i^T$ is ill-conditioned since $\text{rank}(\Phi_N) \leq N$. For the same reason the "standard" stochastic approximation procedure

$$\lambda_t = \Pi_\Lambda(\lambda_{t-1} + \gamma_t(y_t - \lambda_{t-1}^T \phi_t)\phi_t)$$

(here $\Pi_\Lambda(\cdot)$ stands for the projector onto the set $\Lambda = \{\lambda : \|\lambda\|_1 \leq 1\}$) with stepsizes $\gamma_t = O(t^{-1})$ is completely inappropriate here (the Hessian matrix $\Phi = \lim_{t\to\infty} \phi_t \phi_t^T$ is extremely ill-conditioned). However, one can use the robust version of this algorithm (cf. [7]) with the gain $\gamma_t = O(t^{-1/2})$ with the estimate $\bar{\lambda}_N$ obtained by the averaging:

$$\bar{\lambda}_N = \frac{1}{N} \sum_{t=0}^{N-1} \lambda_t.$$

The prediction error of the latter algorithm attains

$$E[\phi_N^T(\lambda^* - \bar{\lambda}_N)]^2 = O(N^{-1/2}) \tag{11}$$

for and does not depend on the conditional number of Φ. However, the constant factor in the right-hand side of (11) is proportional to the "l_2-level" of noise $E\|e_t\phi_t\|_2^2$ which is $O(M)$.

The method we use here is the non-Eucludian stochastic approximation procedure associated with the L_1-norm. It has been first introduced in [7] and is referred to as *mirror-descent algorithm.*

Let $q = 2\ln M$ and γ_t, $t = 1, ..., N$ be a positive sequence (we give the precise definition of this sequence below). We set $W(z) = \|z\|_q^2/2$. In order to obtain the estimate $\bar{\lambda}_t$ of the parameter λ^* given t observations (10) we use the following mirror-descent algorithm (cf. [7]):

$$w_t = z_{t-1} + \gamma_t(y_t - \lambda_{t-1}^T \phi_t)\phi_t, \quad z_0 \in \boldsymbol{R}^M, \;\; \|z_0\|_q \leq 1;$$

$$z_t = \left\{ \begin{array}{ccc} w_t, & \text{for} & \|w_t\|_q \leq 1, \\ \|w_t\|_q^{-1} w_t, & \text{for} & \|w_t\|_q > 1; \end{array} \right.$$

$$\begin{aligned} \lambda_t &= \nabla W(z_t); \\ \bar{\lambda}_t &= \frac{1}{m} \sum_{i=t-m}^{t-1} \lambda_i, \;\; \text{with} \;\; m = \left[\frac{t}{2}\right]. \end{aligned} \tag{12}$$

Let $\mathcal{F}_t = \sigma(z_0, \phi_1, e_1, b_1, ..., \phi_t, e_t, b_t)$. We consider the following assumptions:

Assumption 1. $\|\lambda^*\|_1 \leq 1$.

Assumption 2. For any t, and some $\kappa > 1$ and $L, K, \epsilon, \sigma_e < \infty$ the following holds:

$$\begin{aligned}
&E(e_t^2|\mathcal{F}_{t-1}, \phi_t, b_t) \leq \sigma_e^2,\\
&E|b_t|^2 \leq \epsilon^2\\
&\|\phi_t\|_\infty \leq L,\\
&E(e_t|\mathcal{F}_{t-1}, \phi_t, b_t) = 0.
\end{aligned}$$

Assumption 3. Furthermore, we require that the limit

$$\Phi = \lim_{t\to\infty} E\phi_t\phi_t^T$$

exists and for for some $K < \infty$, $\rho > 0$ and any $m \geq 1$

$$\|E\phi_{t+m}\phi_{t+m} - \Phi|\mathcal{F}_t\|_\infty \leq KL^2(1-\rho)^{m-1}.$$

Note that Assumption 3 holds for the process (ϕ_t), $t = 1, ...$ which satisfies an exponential $\phi-$mixing condition. For instance, it holds true when $\phi_t = \phi(Y_t)$ and (Y_t), $t = 1, ...$ is a Doeblin Markov chain (cf. [2], Section 5.5).

Theorem 1 *Suppose that Assumptions 1 – 3 hold and the gain coefficient is*

$$\gamma_i = \left(\sqrt{3ei\ln(M)(9K\rho^{-1}+1)}L(eL + \sigma_e + \epsilon)\right)^{-1}, \quad i = 1, ..., N. \tag{13}$$

Then for any $M \geq 2$ and $N \geq 2$

$$E[\phi_N^T(\bar{\lambda}_N - \lambda^*)^2] \quad \leq \quad \frac{12}{\gamma_N N} + \frac{60L^2\ln(KN+1)}{N\rho(1-\rho)^2} + \frac{45KL^2}{N\rho} + e\epsilon L. \tag{14}$$

3 Main result

We return now to the basic estimation problem (4). We define the following functional class $\mathcal{F}_N^d(L^*, \nu)$:

Definition 2 *Let L^* and ν be positive reals and d and N be positive integers. We associate with the tuple (L^*, ν, d, N) the class $\mathcal{F}_N^d(L^*, \nu)$ comprised of all functions $f : \boldsymbol{R}^d \to \boldsymbol{R}$ which are Fourier transforms of finite Borel complex-valued measures on $\boldsymbol{R}^d$:*

$$f(x) = \int \exp(i\omega^T x)\widehat{F}(d\omega),$$

such that

$$\begin{aligned}
\int |\widehat{F}(d\omega)| \quad &\leq \quad L^*\\
\int_{|\omega|>\rho} |\widehat{F}(d\omega)| \quad &\leq \quad \rho^{-1}N^\nu \quad \forall\rho > 0
\end{aligned}$$

Note that the classes in question grow as N grows up.

The problem is to recover the unknown function $f : \boldsymbol{R}^d \to \boldsymbol{R}$, given N observations $y_t = f(Y_{t-1}) + e_t$ of of the function (cf. (1)). We assume that we know in advance the parameters L^*, ν of the class $\mathcal{F}_N^d(L^*, \nu)$, as well as σ_e^2 from (2).

The idea of the algorithm below can be summarized as follows. We fix a large enough ball W_ρ in the space of frequencies, so that f can be properly approximated by the Fourier transform of a measure with the support contained in W_ρ. On W_ρ we define a fine ϵ-net $\Omega = \{\omega_i\}$ of cardinality $M^* = O(N^\alpha)$ for some $\alpha < \infty$. Next we denote

$$\phi(x) = \{\sqrt{2}L^* \cos(\omega^T x),\ \sqrt{2}L^* \sin(\omega^T x)\}_{\omega\in\Omega},\ \ \phi_t = \phi(Y_t) \tag{15}$$

and find the approximation $\widehat{\lambda}$ of λ^* from the observations

$$y_t = \phi_t^T \lambda^* + b_t + e_t$$

on the set $\Lambda = \{\lambda \in \boldsymbol{R}^M :\ \|\lambda\|_1 \le 1\}$. Due to Barron's approximation result we know that there exists $\lambda^* \in \Lambda$ such that the quantity

$$Eb_t^2 = \int (\phi(x)^T \lambda^* - f(x))\mu_t(dx),$$

where μ_t stands for the distribution of Y_t, is small.

The algorithm which implements the above idea is as follows:

Algorithm 3.1 1. Given N, d, L^*, ν and σ_e, we set

$$\eta = \sqrt{\frac{d \ln[N^\nu L^*(L^* + \sigma_e)\sqrt{d}]}{N}}(\sigma_e + L^*),\ \ \rho(\eta) = \frac{N^\nu}{\eta},\ \ \epsilon = \frac{\eta}{2L^*\sqrt{d}(L^* + \sigma_e)}. \tag{16}$$

2. We define an ϵ-net $\Omega = \{\omega_k\}_{k=1}^{M^*}$ on the ball

$$W_{\rho(\eta)} = \{\omega \in \boldsymbol{R}^d :\ |\omega| \le \rho(\eta)\}$$

with ϵ given by (16). The cardinality M^* of the net is assumed to satisfy the inequality

$$M^* \le (1 + 2\epsilon^{-1}\rho(\eta))^d \tag{17}$$

(such a net for sure exists).

3. Let $M = 2M^*$, $\Lambda = \{\lambda \in \boldsymbol{R}^M :\ \|\lambda\|_1 \le 1\}$ and

$$f_\lambda(x) = \sum_{k=1}^{M^*} [\lambda_{2k-1}\sqrt{2}L^* \cos(\omega_k^T x) + \lambda_{2k}\sqrt{2}L^* \sin(\omega_k^T x)].$$

Then we set ϕ_t as in (15) and use the stochastic approximation algorithm described in the previous section to obtain the estimate $\widehat{\lambda}$ of $\lambda^* \in \Lambda$ from the observations

$$y_t = \phi_t^T \lambda^* + e_t,\ \ t = 1, ..., N.$$

The convergence rate of the resulting estimate

$$\widehat{f}_N(x) = \sum_{k=1}^{M^*} [\widehat{\lambda}_{2k-1}\sqrt{2}L^* \cos(\omega_k^T x) + \widehat{\lambda}_{2k}\sqrt{2}L^* \sin(\omega_k^T x)]$$

of f is given by the following

Theorem 2 *Let $f \in \mathcal{F}_N^d(L^*, \nu)$, (2) be satisfied and the gain γ_i, $i = 1, ..., N$ be chosen as*

$$\gamma_i = \frac{1}{\sqrt{ei \ln M}(2\pi)^{1/4} \exp(\frac{(L^* + \sigma_e/2)^2}{4\sigma_e^2})L^*(eL^* + \sigma_e)}.$$

Then for all large enough N

$$E(\widehat{f}_N(Y_N) - f(Y_N))^2 \le \kappa\sqrt{\frac{\ln N}{N}}L^*(\sigma_e + L^*)d\sqrt{\nu}(2\pi)^{d/4} \exp\left(\frac{d(L^* + \sigma_e/2)^2}{4\sigma_e^2}\right), \tag{18}$$

where κ is an absolute constant.

Comments:

1. If $\sigma_e \ll L^*$, the constant in (18) which reflects the dependence on L^* and σ_e becomes really large due to the exponent term $\exp((L^*)^2/2\sigma_e^2)$. This multiplier comes out of the estimate of the "forgetting factor" $\rho = \sqrt{2\pi}\exp(-\frac{(L^*+\sigma_e/2)^2}{2\sigma_e^2})$ in the mixing inequality in Assumption 3 for our problem. Of course, this estimate is minimax and rather pessimistic. Under some mild extra assumptions on the Markov chain (Y_t) this estimate can be significantly improved. However, this study is out of the scope of the present paper.
2. One can easily recognize that the computation difficulty of the proposed algorithm is order of $N \times M$ (this is the complexity of the stochastic approximation procedure), and the memory volume which is required by the algorithm is order of M.

3.1 Lower bound

We have shown that when estimating a nonlinear function from the Barron class on the basis of N observations (1), the expected quadratic error

$$E(f(Y_t) - \widehat{f}(Y_t))^2$$

can be made $O(\sqrt{\ln N}N^{-1/2})$, with the constant factor in $O(\cdot)$ depending on the parameters L (the constant of the class) and σ_e (intensity of noise e_t in observations (1)). A natural question is whether any estimate with essentially better expected performance is possible. We are about to show that in the minimax setting the answer to the latter question is negative.

Theorem 3 *Let $L > 0$. Consider the problem of estimating a univariate function $f(x) : \boldsymbol{R} \to \boldsymbol{R}$ from observations $(x_t, y_t = f(y_{t-1}) + e_t)$, $t = 1, ..., N$, where (e_t) is a sequence of independent and identically distributed random variables, $e_1 \sim \mathcal{N}(0, \sigma_e^2)$ and y_0 has the invariant distribution. Let $\mathcal{F}_N^*(L)$ be the class $\mathcal{F}_N^1(L, 1)$. Then for some absolute constant κ and all large enough values of N for any estimate $\widehat{f}$ of $f \in \mathcal{F}_N^*(L)$ on the basis of the above observations one has*

$$E(\widehat{f}(y_t) - f(y_t))^2 \geq \kappa L \sigma_e \sqrt{\frac{\ln N}{N}}. \tag{19}$$

Comment. In the case of $L \sim 1$ and $\sigma_e \sim 1$, the lower bound (19) differs from the upper bound (18) by an absolute constant factor only.

4 Proofs

In what follows C and C' stand for positive constants which values are not important. The proofs are split into a sequence of steps.

4.1 Proof of Theorem 1

Step 1. In this step we introduce some notations and basic inequalities. Set

$$\begin{aligned} \Delta_t &= \lambda_t - \lambda^* \\ \tilde{W}(z) &= W(z) - z^T \lambda^* \\ p &= \frac{q}{q-1}. \end{aligned}$$

We shall verify that the following holds for $x \in \boldsymbol{R}^M$, $z \in \boldsymbol{R}^M$, $\|z\|_q \leq 1$, and z_t and w_t from (12):

$$\begin{aligned} \|x\|_q &\leq \sqrt{e}\|x\|_\infty \\ \|\Delta_t\|_1 &\leq e \\ x^T \nabla^2 W(z) x &\leq (q-1)\|x\|_q^2 \\ \|\nabla W(z+x) - \nabla W(z)\|_1 &\leq (2q-3)\sqrt{e}\|x\|_q^2 \end{aligned}$$

$$\begin{aligned} \tilde{W}(z_t) &\leq \tilde{W}(w_t) \\ |\tilde{W}(z_t)| &\leq 3/2 \\ \|w_t - z_{t-1}\|_q &\leq \gamma_t \sqrt{e} L \ (eL + |e_t| + |b_t|). \end{aligned} \tag{20}$$

The first relation is simply

$$\|x\|_q \quad \leq \quad M^{1/q} \|x\|_\infty = \sqrt{e} \|x\|_\infty$$

Note that

$$\nabla_i W(z) \quad = \quad \text{sign}(z_i) |z_i|^{q-1} \|z\|_q^{2-q}$$

what implies

$$\|\lambda_t\|_1 \leq M^{1/q} \|\lambda_t\|_p = \sqrt{e} \|\nabla W(z_t)\|_p = \sqrt{e} \|z_t\|_q \leq \sqrt{e}$$

and the second one results

$$\|\Delta_t\|_1 \leq 1 + \|\lambda_t\|_1 < e.$$

The third one is an immediate consequence of

$$\nabla^2 W(z) \quad = \quad -(q-2) Z Z^T + (q-1) \|z\|_q^{2-q} \text{diag}(z_i^{q-2}), \qquad Z_i = \text{sign}(z_i) |z_i|^{q-1} \|z\|_q^{1-q}.$$

The proof of the fourth one reuses the same identity, with the fact that $\|Z\|_p = 1$:

$$\begin{aligned} \|\nabla W(z+x) - \nabla W(z)\|_1 &= \|\nabla^2 W(z') x\|_1, \qquad \text{for some } z' \\ &\leq (q-2) \|Z'\|_1 |x^T Z'| + (q-1) \|z'\|_q^{2-q} \sum_i |z_i'|^{q-2} |x_i| \\ &\leq (q-2) \|Z'\|_1 \|Z'\|_p \|x\|_q + (q-1) \|x\|_{q/2} \\ &\leq (q-2) M^{1/q} \|Z'\|_p^2 \|x\|_q + (q-1) \|x\|_q M^{1/q} \\ &= (2q-3) \sqrt{e} \|x\|_q \end{aligned}$$

The fifth one is a consequence of the following monotonicity property, satisfied for $r \geq 1$, $\|z\|_q = 1$

$$\frac{d}{dr} \tilde{W}(rz) = r \|z\|_q^2 - z^T \lambda^* \geq r - \|z\|_q \|\lambda^*\|_p \geq r - \|\lambda^*\|_1 \geq 0$$

The next inequality follows from the fact that for $\|z\|_q \leq 1$

$$|\tilde{W}(z)| = | \ \|z\|_q^2/2 - z^T \lambda^* | \leq 1/2 + |z^T \lambda^*| \leq 1/2 + \|z\|_\infty \leq 1/2 + \|z\|_q \leq 3/2.$$

Finally, for the last relation we get:

$$\begin{aligned} \|w_t - z_{t-1}\|_q &\leq \sqrt{e} \|w_t - z_{t-1}\|_\infty = \gamma_t \sqrt{e} \|(y_t - \lambda_{t-1}^T \phi_t) \phi_t\|_\infty \\ &\leq \gamma_t \sqrt{e} \ (\|\phi_t \phi_t^T \Delta_{t-1}\|_\infty + \|(e_t + b_t) \phi_t\|_\infty) \\ &\leq \gamma_t \sqrt{e} \ (\|\phi_t \phi_t^T\|_\infty \|\Delta_{t-1}\|_1 + |e_t + b_t| \|\phi_t\|_\infty) \\ &\leq \gamma_t \sqrt{e} \ (eL^2 + (|e_t| + |b_t|) L) \end{aligned}$$

Step 2. We prove here the following bound:

$$\frac{1}{m} \left| E \sum_{i=t-m+1}^{t} \Delta_{i-1}^T [\phi_i \phi_i^T - \Phi] \Delta_{i-1} \right| \quad \leq \quad \frac{6eKL^2}{t\rho} \left(e + 3 \ln M (eL + \sigma_e + \epsilon)^2 \gamma_1 \sqrt{t} \right).$$

Using the mixing condition in assumption 3 we decompose $\phi_t \phi_t^T - \Phi$ into a martingale increment μ_t and a difference process

$$\phi_t \phi_t^T - \Phi = \mu_t + \nu_{t-1} - \nu_t,$$

where

$$\begin{aligned}
\nu_t &= \sum_{i=1}^{\infty} E(\phi_{t+i}\phi_{t+i}^T - \Phi|\mathcal{F}_t), \\
\mu_t &= \sum_{i=0}^{\infty} E(\phi_{t+i}\phi_{t+i}^T - \Phi|\mathcal{F}_t) - \sum_{i=0}^{\infty} E(\phi_{t+i}\phi_{t+i}^T - \Phi|\mathcal{F}_{t-1}).
\end{aligned}$$

Then

$$\begin{aligned}
&\sum_{i=t-m+1}^{t} \Delta_{i-1}^T[\phi_i\phi_i^T - \Phi]\Delta_{i-1} \\
&= \sum_{i=t-m+1}^{t} \Delta_{i-1}^T\mu_i\Delta_{i-1} + \sum_{i=t-m+1}^{t} [\Delta_{i-1}^T\nu_{i-1}\Delta_{i-1} - \Delta_{i-1}^T\nu_i\Delta_{i-1}] = I_1 + I_2.
\end{aligned}$$

However, I_1 is a martingale, thus $EI_1 = 0$. We can rewrite I_2 as follows

$$I_2 = \Delta_{t-m}^T\nu_{t-m}\Delta_{t-m} - \Delta_t^T\nu_t\Delta_t + \sum_{i=t-m+1}^{t} [\Delta_i^T\nu_i\Delta_i - \Delta_{i-1}^T\nu_i\Delta_{i-1}].$$

We have

$$\|\nu_t\|_\infty \leq KL^2\sum_{i=1}^{\infty}(1-\rho)^{i-1} = \frac{KL^2}{\rho},$$

so that

$$\begin{aligned}
|EI_2| &\leq |E\Delta_t^T\nu_t\Delta_t| + |E\Delta_{t-m}^T\nu_{t-m}\Delta_{t-m}| + \sum_{i=t-m+1}^{t} E[\|\Delta_{i-1}-\Delta_i\|_1\|\nu_i\|_\infty(\|\Delta_{i-1}\|_1 + \|\Delta_i\|_1)] \\
&\leq \frac{2eKL^2}{\rho}\left(e + \sum_{i=t-m+1}^{t} E\|\Delta_{i-1}-\Delta_i\|_1\right).
\end{aligned}$$

Next we get

$$\begin{aligned}
\|\Delta_i - \Delta_{i-1}\|_1 &\leq (2q-3)\sqrt{e}\|z_i - z_{i-1}\|_q \\
&\leq 2q\sqrt{e}\|w_i - z_{i-1}\|_q \\
&\leq 2\gamma_i qeL(eL + |b_i + e_i|) \\
E[\|\Delta_i - \Delta_{i-1}\|_1] &\leq 4\gamma_i \ln(M)eL(eL + \sigma_e + \epsilon) \leq 4\gamma_i\ln(M)(eL+\sigma_e+\epsilon)^2.
\end{aligned}$$

Finally, we obtain for $t \geq 2$ (note that $\sqrt{t-m} \geq \sqrt{(t-1)/2}$)

$$\begin{aligned}
\frac{1}{m}|EI_2| &\leq \frac{2eKL^2}{m\rho}\left(e + 4\ln(M)(eL+\sigma_e+\epsilon)^2\sum_{i=t-m+1}^{t}\gamma_i\right) \\
&\leq \frac{2eKL^2}{m\rho}\left(e + 8\ln(M)(eL+\sigma_e+\epsilon)^2\gamma_1(\sqrt{t-1}-\sqrt{t-m})\right) \\
&\leq \frac{6eKL^2}{t\rho}\left(e + 3\ln(M)(eL+\sigma_e+\epsilon)^2\gamma_1\sqrt{t}\right).
\end{aligned}$$

Step 3. Let m be chosen as in (12). We prove here that the "averaged" criterion satisfies

$$\frac{1}{m}\sum_{i=t-m+1}^{t} E[(\phi_i^T\Delta_{i-1})^2] \quad\leq\quad \frac{6}{\gamma_1\sqrt{t}} + 4L^2\ln(M)(eL+\sigma_e+\epsilon)^2\frac{\gamma_1}{\sqrt{t}} + e\epsilon L. \tag{21}$$

Note first that $\Delta_t = \nabla\tilde{W}(z_t)$. Using the result of Step 1 we get the recursive relation

$$\begin{aligned}\tilde{W}(z_t) \quad&\leq\quad \tilde{W}(w_t) \leq \tilde{W}(z_{t-1}) + \nabla^T\tilde{W}(z_{t-1})(w_t - z_{t-1}) + \frac{1}{2}\max_z(w_t - z_{t-1})^T\nabla^2\tilde{W}(z)(w_t - z_{t-1}) \\ &\leq\quad \tilde{W}(z_{t-1}) + \gamma_t\Delta_{t-1}^T((e_t+b_t)\phi_t - \phi_t\phi_t^T\Delta_{t-1}) + \frac{(q-1)}{2}\|w_t - z_{t-1}\|_q^2\end{aligned}$$

hence

$$\gamma_t(\phi_t^T\Delta_{t-1})^2 \quad\leq\quad \tilde{W}(z_{t-1}) - \tilde{W}(z_t) + \gamma_t(e_t+b_t)\Delta_{t-1}^T\phi_t + \frac{(q-1)}{2}\|w_t - z_{t-1}\|_q^2. \tag{22}$$

Then by (20)

$$\|w_t - z_{t-1}\|_q^2 \leq \gamma_t^2 e\ L^2(eL + |e_t| + |b_t|)^2$$

and by the Minkowski inequality

$$\begin{aligned}E[\|w_t - z_{t-1}\|_q^2]^{1/2} \quad&\leq\quad \gamma_t e\ L(eL+\sigma_e+\epsilon) \\ \frac{(q-1)}{2}E[\|w_t - z_{t-1}\|_q^2] \quad&\leq\quad \ln(M)e\gamma_t^2L^2(eL+\sigma_e+\epsilon)^2.\end{aligned}$$

On the other hand,

$$|Eb_t\Delta_{t-1}^T\phi_t| \leq e\epsilon L.$$

Thus, when taking expectations in equation (22),

$$E(\Delta_{t-1}^T\phi_t)^2 \quad\leq\quad \gamma_t^{-1}E[\tilde{W}(z_{t-1}) - \tilde{W}(z_t)] + e\epsilon L + \ln(M)e\gamma_tL^2(eL+\sigma_e+\epsilon)^2.$$

When summing over t:

$$\begin{aligned}&\sum_{i=t-m+1}^{t} E(\Delta_{i-1}^T\phi_i)^2 \leq E\left[\frac{\tilde{W}_{t-m}}{\gamma_{t-m}} - \frac{\tilde{W}_t}{\gamma_t} + \sum_{i=t-m}^{t-1}[\gamma_{i+1}^{-1} - \gamma_i^{-1}]\tilde{W}_i\right] + me\epsilon L \\ &\qquad + \ln(M)eL^2(eL+\sigma_e+\epsilon)^2\sum_{i=t-m+1}^{t}\gamma_t \\ &\leq\quad \frac{3}{2}\left(\gamma_{t-m}^{-1} + \gamma_t^{-1} + \sum_{i=t-m}^{t-1}[\gamma_{i+1}^{-1} - \gamma_i^{-1}]\right) + me\epsilon L + \ln(M)eL^2(eL+\sigma_e+\epsilon)^2\gamma_1\sum_{i=t-m+1}^{t}\frac{1}{\sqrt{i}} \\ &\leq\quad \frac{3}{\gamma_1}\sqrt{t} + \frac{e\epsilon Lt}{2} + \ln(M)eL^2(eL+\sigma_e+\epsilon)^2\gamma_1\sqrt{t}.\end{aligned}$$

Step 4. We show that the variable $\bar{\Delta}_t = \frac{1}{m}\sum_{i=t-m+1}^{t}\Delta_{i-1}$ satisfies

$$|E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - E(\bar{\Delta}_t^T\phi_t)^2| \leq \frac{2e^2L^2K}{t} - 8e^2L^2\ln(1-\rho)\frac{\ln t}{t}.$$

Define for some s

$$\hat{\Delta}_t = \frac{1}{m}\sum_{i=t-m+1}^{t-s}\Delta_{i-1}.$$

We have

$$\begin{aligned}|E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - E(\bar{\Delta}_t^T\phi_t)^2| &\leq |E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - \hat{\Delta}_t^T\Phi\hat{\Delta}_t| \\ &\quad +|E\hat{\Delta}_t^T\Phi\hat{\Delta}_t - E(\hat{\Delta}_t^T\phi_t)^2| \\ &\quad +|E(\bar{\Delta}_t^T\phi_t)^2 - E(\hat{\Delta}_t^T\phi_t)^2| \\ &= I_1 + I_2 + I_3. \end{aligned} \tag{23}$$

We get for I_3:

$$|I_3| \leq E(\|\bar{\Delta}_t\|_1 + \|\hat{\Delta}_t\|_1)\|\bar{\Delta}_t - \hat{\Delta}_t\|_1\|\phi_t\phi_t^T\|_\infty \leq 2eL^2E\|\hat{\Delta}_t - \bar{\Delta}_t\|_1 \leq \frac{2e^2L^2s}{m}.$$

The same bound is valid for I_1. For I_2 we obtain due to Assumption 2

$$|I_2| = |E\hat{\Delta}_t^T E(\Phi - \phi_t\phi_t^T|\mathcal{F}_{t-s})\hat{\Delta}_t| \leq KL^2(1-\rho)^{s-1}E\|\hat{\Delta}_t\|_1^2 \leq e^2KL^2(1-\rho)^{s-1}.$$

Finally

$$|E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - E(\bar{\Delta}_t^T\phi_t)^2| \leq \frac{4e^2L^2}{m}\left(s + \frac{Km}{4}(1-\rho)^{s-1}\right)$$

When choosing $s = [-\ln(K'+1)/\ln(1-\rho)]$, $K' = Km/4$ we obtain

$$\begin{aligned}|E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - E(\bar{\Delta}_t^T\phi_t)^2| &\leq \frac{4e^2L^2}{m}\left(s + K'(1-\rho)^{-\ln(K'+1)/\ln(1-\rho)-2}\right) \\ &= 4\frac{e^2L^2}{m}\left(s + \frac{K'}{K'+1}(1-\rho)^{-2}\right) \\ &\leq 4\frac{e^2L^2\ln(K'+1)}{m}\left(\rho^{-1} + (1-\rho)^{-2}\right) \\ &\leq 4\frac{e^2L^2\ln(Km/4+1)}{m}\rho^{-1}(1-\rho)^{-2}\end{aligned}$$

Step 5. We gather now the results of the previous steps. Due to the convexity of $\Delta^T\Phi\Delta$ we have

$$\begin{aligned}E(\bar{\Delta}_t^T\phi_t)^2 &\leq |E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - E(\bar{\Delta}_t^T\phi_t)^2| + E\bar{\Delta}_t^T\Phi\bar{\Delta}_t \\ &\leq |E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - E(\bar{\Delta}_t^T\phi_t)^2| + \frac{1}{m}\sum_{i=t-m+1}^{t} E\Delta_{i-1}^T\Phi\Delta_{i-1} \\ &\leq |E\bar{\Delta}_t^T\Phi\bar{\Delta}_t - E(\bar{\Delta}_t^T\phi_t)^2| + \frac{1}{m}\sum_{i=t-m+1}^{t}\left[E\Delta_{i-1}^T(\Phi - \phi_i\phi_i^T)\Delta_{i-1} + E(\Delta_{i-1}^T\phi_i)^2\right] \\ &\leq \frac{2e^2L^2\ln(Kt+1)}{t}\rho^{-1}(1-\rho)^{-2} + \frac{6eL^2K}{\rho t}\left(e + 3\ln(M)(eL+\sigma_e+\epsilon)^2\gamma_1\sqrt{t}\right) \\ &\quad + \frac{6}{\gamma_1\sqrt{t}} + 2\ln(M)L^2(eL+\sigma_e+\epsilon)^2\frac{\gamma_1}{\sqrt{t}} + e\epsilon L \\ &\leq \frac{2e^2L^2\ln(Kt+1)}{t}\rho^{-1}(1-\rho)^{-2} + e\epsilon L + \frac{6e^2L^2K}{\rho t} \\ &\quad + \frac{6}{\gamma_1\sqrt{t}} + 2L^2e\ln(M)(eL+\sigma_e+\epsilon)^2\frac{\gamma_1}{\sqrt{t}}\left(\frac{9K}{\rho} + 1\right).\end{aligned}$$

Now note that the two last terms only depend on $\gamma_t = \gamma_1/\sqrt{t}$. When choosing γ_1 to balance these terms we get

$$\gamma_1 = \left(\sqrt{3e\ln(M)(9K\rho^{-1}+1)}L(L+\sigma_e+\epsilon)\right)^{-1},$$

and the bound (14) for $E(\bar{\Delta}_t^T\phi_t)^2$. ■

4.2 Proof of Theorem 2

Step 1. According to (17) and (16) we have for all large enough N

$$M \leq (1+2\epsilon^{-1}\rho(\eta))^d \leq (C\epsilon^{-1}\rho(\eta))^d \leq \left(C' N^{\nu} L^* \sqrt{d}(L^*+\sigma_e)\eta^{-2}\right)^d \tag{24}$$

Let μ_t be the distribution of Y_t, $t=1,\ldots$.

Step 2. Let us verify that for every $f \in \mathcal{F}_N^d(L^*,\nu)$ there exists $\lambda^* \in \Lambda$ such that the function

$$\tilde{f}(x) = \phi(x)^T \lambda^*$$

with $\omega_k \in \Omega$ satisfies

$$\|\tilde{f} - f\|_{2,\mu_t} \leq 3\eta.$$

Indeed, by (16) we have

$$\int_{|\omega|>\rho(\eta)} |\widehat{F}(d\omega)| \leq N^{\nu}\rho^{-1}(\eta) = \eta.$$

This implies that if we define the measure $\widehat{G}$ as $\widehat{G}(A) = \widehat{F}(A \cap W_{\rho(\eta)})$ and define g as the Fourier transform of $\widehat{G}$, then

$$\|f-g\|_{2,\mu_t} \leq \|f-g\|_\infty \leq \eta. \tag{25}$$

On the other hand, it follows from Barron's proof of (8) (see [1]) that one can find a function of the form

$$h = \sum_{k=1}^m \delta_k \exp(i\zeta_k^T x)$$

with $\zeta_k \in W_{\rho(\eta)}$, $m = \rfloor 1/\eta^2 \lfloor$ and $\|\delta\|_1 \leq L^*$ such that

$$\|h-g\|_{2,\mu_t} \leq \eta. \tag{26}$$

Since $\|f\|_\infty \leq L^*$, we conclude from (1) that

$$E_{f,y}|Y_t|^2 \leq d((L^*)^2 + \sigma_e^2),$$

so that for any $\omega, \omega' \in \boldsymbol{R}^d$

$$\int |e^{ix^T\omega} - e^{ix^T\omega'}|^2 \mu_t(dx) \leq 4|\omega-\omega'|^2 \int |x|^2 \mu_t(dx) = 4d((L^*)^2+\sigma_e^2)|\omega-\omega'|^2. \tag{27}$$

Let ω_k be the element of Ω closest to ζ_k. Then for

$$r(x) = \sum_{k=1}^m \delta_k \exp(i\omega_k^T x),$$

we obtain due to (27):

$$\|h-r\|_{2,\mu} \leq L^* \max_k \left(\int |e^{ix^T\omega_k} - e^{ix^T\zeta_k}|^2 \mu_t(dx)\right)^{1/2} \leq 2L^* \max_k |\omega_k - \zeta_k| \sqrt{d((L^*)^2+\sigma_e^2)} = \eta$$

(see (16)). Along with (25) and (26) this estimate yields $\|f-r\|_{2,\mu_t} \leq 3\eta$ for any $t \geq 1$. Now we can set $\tilde{f}(x) = \mathrm{Re}\{r(x)\}$.

Now, let ϕ_t be as in (15), $\lambda^*_{2k-1} = \mathrm{Re}(\delta_k)$ and $\lambda^*_{2k} = \mathrm{Im}(\delta_k)$, $k=1,\ldots,M$. Then for $y_t = f(Y_{t-1}) + e_t$ Assumption 2 holds with $\epsilon = 3\eta$ and σ_e unchanged. Since

$$|\mathrm{Re}(\delta_k)| + |\mathrm{Im}(\delta_k)| \leq \sqrt{2}|\delta_k|,$$

we have $\|\lambda^*\| \leq 1$ if we take $L = \sqrt{2}L^*$ in Assumption 2.

Furthermore, the Markov chain Y_t satisfies the Doeblin condition (cf. Case(b) p. 197 of [2]): there exists a conditional density of the d-step transition probability, which is uniformly (with respect to $x \in \boldsymbol{R}^d$) bounded from below on some subset of $\boldsymbol{R}^d$. For proving this, note first that, since $\|f\|_\infty \leq L^*$, the transition probability from $x = Y_t$ to $y = y_{t+1}$ has a density $p(x, y)$ which satisfies for $y \in [-\sigma_e/2, \sigma_e/2]$ and $x \in \boldsymbol{R}^d$

$$p(x, y) \geq \delta = \frac{1}{\sqrt{2\pi}\sigma_e} \exp\left(-\frac{(L^* + \sigma_e/2)}{2\sigma_e^2} \right).$$

Thus, for any function g

$$\begin{aligned} E[g(Y_{t+d})|Y_t] &= \int g(y_{t+d}, \dots, y_{t+1}) p(Y_t, y_{t+1}) \dots p(Y_{t+d-1}, y_{t+d}) dy_{t+1} \dots dy_{t+d} \\ &\geq \delta^d \int_{[-\sigma_e/2, \sigma_e/2]^d} g(y_{t+d}, \dots, y_{t+1}) dy_{t+1} \dots dy_{t+d}. \end{aligned}$$

This means that the d-step transition probability of the chain Y_t has an absolutely continuous part with a density larger than δ^d on $[-\sigma_e/2, \sigma_e/2]^d$. Then the inequality (5.6) p. 197 of [2] holds with $\nu = d$, i.e., the transition probability of Y_t satisfies

$$|P(x, A) - \pi(A)| \leq (1 - \rho_0)^{n/d - 1}$$

($\pi(\cdot)$ stands for the invariant distribution of Y_t) with

$$\rho_0 = \delta^d \sigma_e^d = (\sqrt{2\pi})^{-d/2} \exp\left(-\frac{d(L^* + \sigma_e/2)^2}{2\sigma_e^2} \right).$$

In other words, the process (Y_t) is exponentially ϕ-mixing. Then using Theorem A.6 of [3] we conclude that for any continuous function $g : \boldsymbol{R}^d \to \boldsymbol{R}$:

$$|Eg(Y_m) - \pi g| \leq 2\|g\|_\infty (1 - \rho_0)^{m/d - 1},$$

what gives the mixing inequality of Assumption 3, with $K = 2/(1 - \rho_0)$ and $\rho = \rho_0/d$.

Step 3. When applying the bound (14) of Theorem 1 we get for all large enough values of N (note that here ϵ depends on N and tends to zero)

$$E\|\widehat{f}_N - f\|^2 \leq C' \rho^{-1/2} \left(\frac{\ln M}{N} \right)^{1/2} L^*(\sigma_e + L^*),$$

The latter quantity, as it is immediately seen from (24) and (16), is bounded from above by

$$\kappa \rho^{-1/2} L^*(\sigma_e + L^*) \sqrt{d\nu \ln N / N}$$

with properly chosen absolute constant κ. ∎

4.3 Proof of Theorem 3

Let φ_k, $k = 1, ..., N$, be

$$\varphi_k(x) = L \cos(\frac{2\pi}{\sigma_e} kx).$$

Given a positive integer p, let us denote by $\mathcal{F}_p$ the set of all convex combinations of the functions $\varphi_1, ..., \varphi_N$ with the coefficients as follows: $2p$ out of the N coefficients are equal to $(2p)^{-1}$, and other coefficients vanish.

Step 1. It is easily seen that if $p \leq \sqrt{N}$, then $\mathcal{F}_p$ contains a subset $\mathcal{F}_p^*$ with the following properties:

1. Every two functions f, g from $\mathcal{F}_p^*$ have at most p common nonzero Fourier coefficients, so that

$$\frac{L^2\sigma_e}{4p} \leq \int_{-\sigma_e/2}^{\sigma_e/2} |f(x) - g(x)|^2 dx. \tag{28}$$

2. The cardinality K of $\mathcal{F}_p^*$ satisfies the relation

$$K \geq M^{\kappa_1 p}. \tag{29}$$

Step 2. Consider $f \in \mathcal{F}_p^*$ and let $\pi(\cdot) = \pi_f(\cdot)$ denote the invariant distribution of the Markov chain (y_t) defined by (1). Let us prove that for p large enough and any function $g \in \mathcal{F}_p^*$

$$\frac{L^2}{16p} \leq \pi[(f-g)^2], \text{ and } \pi(f^2) \leq \frac{2L^2}{p} \tag{30}$$

Note first that for any a and $|b| \leq b_0$, due to the convexity of $\exp(x) - 1$ (used on the interval $[0, b_0^2]$)

$$\mathrm{e}^{-(a+b)^2} \leq \mathrm{e}^{-a^2/2+b^2} \leq \mathrm{e}^{-a^2/2}\left(1 + (\mathrm{e}^{b_0^2} - 1)\frac{b^2}{b_0^2}\right).$$

Since $|f(x)| \leq L$ by construction, we obtain for the invariant density $p(\cdot)$:

$$\begin{aligned} p(y) &= \int p(x)\frac{1}{\sqrt{2\pi}\sigma_e}\mathrm{e}^{-\frac{(y-f(x))^2}{2\sigma_e^2}}dx \\ &\leq \int \frac{p(x)}{\sqrt{2\pi}\sigma_e}\mathrm{e}^{-\frac{y^2}{4\sigma^2}}\left(1 + (\mathrm{e}^{\frac{L^2}{2\sigma_e^2}} - 1)\frac{f^2(x)}{L^2}\right)dx \\ &= \frac{1}{\sqrt{2\pi}\sigma_e}\mathrm{e}^{-\frac{y^2}{4\sigma^2}}(1 + C(L,\sigma_e)\pi(f^2)). \end{aligned}$$

Then, due to (28)

$$\pi(f^2) = \int f^2(x)p(x)dx \leq \frac{1 + C(L,\sigma_e)\pi(f^2)}{\sqrt{2\pi}\sigma_e}\int f^2(x)\mathrm{e}^{-\frac{x^2}{4\sigma^2}}dx \leq (1 + C(L,\sigma_e)\pi(f^2))\frac{L^2}{p},$$

and for $p \geq 2L^2C(L,\sigma_e)$, one has $\pi(f^2) \leq 2L^2/p$.
Now consider the first inequality of (30). Note that

$$\mathrm{e}^{-(a+b)^2} \geq \mathrm{e}^{-2a^2-2b^2} \geq \mathrm{e}^{-2a^2}(1 - 2b^2).$$

Therefore, using the bound for $\pi(f^2)$ we conclude that the invariant density $p(x)$ satisfies

$$p(y) = \int p(x)\frac{1}{\sqrt{2\pi}\sigma_e}\mathrm{e}^{-\frac{(y-f(x))^2}{2\sigma_e^2}}dx \geq \frac{1}{\sqrt{2\pi}\sigma_e}\mathrm{e}^{-\frac{y^2}{\sigma_e^2}}(1 - \frac{\pi(f^2)}{\sigma_e^2}) \geq \frac{1}{3\sigma_e}\mathrm{e}^{-\frac{y^2}{\sigma_e^2}}$$

for p large enough. Thus

$$\begin{aligned} \int |f(y) - g(y)|^2\pi(dy) &\geq \frac{1}{3\sigma_e}\int_{-\sigma_e/2}^{\sigma_e/2}|f(y) - g(y)|^2\mathrm{e}^{-\frac{y^2}{\sigma_e^2}}dy \\ &\geq \frac{1}{3\sigma_e}\mathrm{e}^{-1/4}\int_{-\sigma_e/2}^{\sigma_e/2}|f(x) - g(x)|^2dx \geq \frac{L^2}{16p}. \end{aligned}$$

Step 3. Now let $\epsilon(p) = \max_{f\in\mathcal{F}_p^*} E[\pi((\widehat{f}-f)^2)]$ (recall that the measures E and π depend on the true function f). We claim that for any $p \leq \sqrt{N}$ large enough, the inequality $\epsilon(p) < \frac{L^2}{256p}$ would imply

$$N \geq \kappa L^{-2}\sigma_e^2 p^2 \ln N. \tag{31}$$

Indeed, let us associate with $\widehat{f}$ the method $\mathcal{B}$ for distinguishing between K hypotheses, the k-th of them stating that N observations (1) come from some element $f \in \mathcal{F}_p^*$. $\mathcal{B}$ is as follows: Given observations, we use $\widehat{f}$ to estimate f; then we find the closest (in the sense of $\int(f(x)-\widehat{f}(x))^2\pi(dx)$) to $\widehat{f}$ element (any one of them in the non-uniqueness case) in $\mathcal{F}_p^*$ and claim that this is the function underlying our observations.

It is immediately seen if any one of our K hypotheses is true, the probability that $\mathcal{B}$ fails to recognize it properly is at most $1/4$. Indeed, assume that the true hypothesis is associated with $f \in \mathcal{F}_p^*$. If $\mathcal{B}$ fails to say that it is the case, then the estimate $\widehat{f}$ is at least at the same distance from f as from some $g \in \mathcal{F}_p^*$ distinct from f. Taking into account the first inequality in (30), we conclude that then

$$\pi[(f-\widehat{f})^2]^{1/2} \geq 1/2\sqrt{\frac{L^2}{16p}}.$$

Then we get from the definition of $\epsilon(p)$ and the Chebyshev inequality:

$$P(\mathcal{B}\neq f) \leq P\left(\pi[(f(x)-\widehat{f})^2] \geq \frac{L^2}{64p}\right) \leq \frac{64p\epsilon(p)}{L^2} \leq \frac{1}{4}.$$

Now note that the Shannon information $I(Y^N, f)$ of the distribution of N-observation samples (1) coming from an element of $\mathcal{F}_p^*$, in view of the second inequality in (30), can be bounded as follows:

$$I(Y^N,f) \leq \frac{C}{\sigma_e^2}\sum_{t=1}^N Ef^2(y_t) \leq \frac{C' N\sigma_e^{-2}L^2}{p}.$$

Then the Fano inequality [5] implies that the above K hypotheses can be distinguished only if

$$\frac{N\sigma_e^{-2}L^2}{p} \geq C\ln K = \kappa p \ln N$$

(we have used (29)), as required in the conclusion of (31). We now take $p = 2L\sqrt{N}/(\sigma_e\sqrt{\kappa\ln(N)})$ so that (31) is violated and the conclusion of Theorem 3 is nothing but $\epsilon(p) > \frac{L^2}{256p}$. ∎

References

[1] A. Barron “Universal approximation bounds for superpositions of a sigmoidal function”, *IEEE Trans. Inf. Theory*, v. 39, N. 3, pp. 930-945, 1993.

[2] J.L. Doob *Stochastic Processes*, J.Wiley & Sons, N.Y, 1953.

[3] P. Hall and C.C. Heyde *Martingale Theory.*

[4] W. Härdle, *Applied Nonparametric Regression*, ES Monograph Series 19, Cambridge, U.K., Cambridge University Press, 1990.

[5] Ibragimov and R. Khas’minski, *Estimation Theory*, Springer, 1981.

[6] A. Juditsky and A. Nemirovski,*Functional Aggregation for Nonparametric Regression*, IRISA Internal Report PI-993, March 1996.

[7] A. Nemirovski and D. Yudin, *Problem complexity and method efficiency in optimization* – J. Wiley & Sons, 1983.