XXVII Saint-Flour Summer School on Probability and Statistics, Saint-Flour, France, 1998

Topics in Non-Parametric Statistics

Arkadi Nemirovski[1]

Preface

The subject of Nonparametric statistics is statistical inference applied to noisy observations of infinite-dimensional "parameters" like images and time-dependent signals. This is a mathematical area on the border between Statistics and Functional Analysis, the latter name taken in its "literal" meaning – as geometry of spaces of functions. What follows is the 8-lecture course given by the author at The XXVIII Saint-Flour Summer School on Probability Theory. It would be impossible to outline in a short course the contents of rich and highly developed area of Non-parametric Statistics; we restrict ourselves with a number of selected topics related to estimating nonparametric regression functions and functionals of these functions. The presentation is self-contained, modulo a few facts from the theory of functional spaces.

[1] Faculty of Industrial Engineering and Management, Technion – Israel Institute of Technology, Technion City, Haifa 32000, Israel; e-mail: `nemirovs@ie.technion.ac.il`

Contents

Chapter 1

Estimating regression functions from Hölder balls

1.1 Introduction

We start with brief outline of the problems we are interested in and the goals we are aimed at.

Statistical problems and estimates. A typical problem of Statistics is as follows:

> (*) We are given a Polish (i.e., metric, separable and complete) "space of observations" Y along with a family of Borel probability distributions $\{\Phi_f(\cdot)\}_{f\in\mathcal{F}}$ on Y; the family is parameterized by a "parameter" f varying in a metric space $\mathcal{F}$.
>
> The goal is, given an "observation " – a realization
>
> $$y \sim \Phi_f$$
>
> of random variable associated with an *unknown* $f \in \mathcal{F}$, to make conclusions about f, e.g.
>
> I. $\underline{\text{[Identification]}}$ To estimate f,
>
> F. $\underline{\text{[Evaluation of a functional]}}$ To estimate the value $F(f)$ at f of a given functional $F : \mathcal{F} \to \mathbf{R}$,
>
> H. $\underline{\text{[Hypotheses testing]}}$ Given a partition $\mathcal{F} = \bigcup\limits_{i=1}^{N} \mathcal{F}_i$ of $\mathcal{F}$, to decide to which element $\mathcal{F}_i$ of the partition f belongs,

In all these problems, a "candidate solution" is an *estimate* – a Borel function $\widehat{f}(y)$ on the "space of observations" Y taking values in an appropriately chosen Polish "space of answers" Z:

• In the case of Identification problem, $Z = \mathcal{F}$, and $\widehat{f}(y)$ is the estimated value of f;

• In the case of problem of evaluating a functional, $Z = \mathbf{R}$, and $\widehat{f}(y)$ is the estimated value of $F(f)$

• In the case of Hypotheses testing, $Z = \{1, ..., N\}$, and $\widehat{f}(y)$ is the (index of the) accepted hypothesis.

Risk of an estimate. Normally it is impossible to recover the "true answer" $f_*(f)$ *exactly*, and we should be satisfied with estimates $\widehat{f}(\cdot)$ which with "high" probability are "close" to true answers.

A natural way to quantify the quality of an estimate is to look at its (mean squared) *risk*

$$\mathcal{R}(\widehat{f}, f) = \left(\mathcal{E}_{\Phi_f}\left\{\mathrm{dist}_Z^2(\widehat{f}(y), f_*(f))\right\}\right)^{1/2}, \tag{1.1}$$

where

- $\mathcal{E}_{\Phi_f}\{\cdot\}$ is the expectation w.r.t. $y \sim \Phi_f$;
- $\mathrm{dist}_Z(\cdot,\cdot)$ is the metric on the "space of answers" Z;
- $f_*(f) \in Z$ is the true answer.

For example

- In the Identification problem, $Z = \mathcal{F}$, $\mathrm{dist}_Z(\cdot,\cdot)$ is the metric $\mathcal{F}$ is equipped with, and $f_*(f) = f$;
- In the Functional Evaluation problem, $Z = \mathbf{R}$, $\mathrm{dist}_Z(p,q) = |p-q|$ and $f_*(f) = F(f)$;
- In the Hypotheses testing problem,

$$Z = \{1, ..., N\}, \ \ \mathrm{dist}_Z(i,j) = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}, \ \ f_*(f) = i \text{ for } f \in \mathcal{F}_i.$$

In the latter case (1.1) is the square root of the probability to misclassify the parameter f of the distribution we observe.

Remark 1.1.1 Of course, (1.1) is not the only meaningful way to measure risk; a general scheme requires to choose a "loss function" – a nondecreasing function $\Psi(t)$ on the nonnegative ray such that $\Psi(0) = 0$ – and to associate with this loss function the risk

$$\mathcal{R}(\widehat{f}, f) = \Psi^{-1}\left(\mathcal{E}_{\Phi_f}\left\{\Psi\left(\mathrm{dist}_Z(\widehat{f}(y), f_*(f))\right)\right\}\right).$$

To order to simplify our considerations and notation (in our course, we shall have enough of other "parameters of situation" to trouble about), in what follows we focus on the mean square risk (1.1), i.e., on the simplest loss functions $\Psi(t) = t^2$.

Risk (1.1) depends on the "true parameter" f, and thus cannot be used "as it is" to quantify the quality of an estimate. There are two standard ways to eliminate the dependence on f and to get a quantitative characterization of an estimate:

- [Bayesian approach] To take average of $\mathcal{R}(\widehat{f}, f)$ over a given a priori distribution of $f \in \mathcal{F}$
- [Minimax approach] To take the supremum of $\mathcal{R}(\widehat{f}, f)$ over $f \in \mathcal{F}$, thus coming to the *worst-case* risk

$$\mathcal{R}(\widehat{f}; \mathcal{F}) = \sup_{f \in \mathcal{F}} \mathcal{R}(\widehat{f}, f) = \sup_{f \in \mathcal{F}} \left(\mathcal{E}_{\Phi_f}\left\{\mathrm{dist}_Z^2(\widehat{f}(y), f_*(f))\right\}\right)^{1/2} \tag{1.2}$$

of an estimate $\widehat{f}$ on the "parameter set" $\mathcal{F}$. In our course, we always use the minimax approach. The major reason for this choice is that we intend to work with infinite-dimensional parameter sets, and these sets usually do not admit "natural" a priori distributions.

With the minimax approach, the quality of "ideal" estimation becomes the *minimax risk*

$$\mathcal{R}^*(\mathcal{F}) = \inf_{\widehat{f}(\cdot)} \mathcal{R}(\widehat{f}; \mathcal{F}) \tag{1.3}$$

– the minimal, over all estimates, worst-case risk of an estimate.

Nonparametric regression problems. In the "parametric" Statistics, the parameter set $\mathcal{F}$ is finite-dimensional: $\mathcal{F} \subset \mathbf{R}^k$ ("the distribution is known up to finitely many parameters"). In the Nonparametric Statistics, the parameter set $\mathcal{F}$ is infinite-dimensional – typically, it is a compact subset of certain functional space, like the space $C([0,1]^d)$ of continuous functions on the unit d-dimensional cube. Typical generic examples are as follows:

- *Nonparametric regression estimation problem:*

  (R) *Recover a function $f : [0,1]^d \to \mathbf{R}$ known to belong to a given set $\mathcal{F} \subset C([0,1]^d)$ via n noisy observations*

  $$y = y^f = \{y_i = f(x_i) + \sigma\xi_i, i = 1, ..., n\} \tag{1.4}$$

  of the values of the function along n given points $x_i \in [0,1]^d$; here $\{\xi_i\}_{i=1}^n$ is the observation noise.

- *Nonparametric density estimation problem:*

  (D) *Recover a probability density f on $[0,1]^d$ known to belong to a given set $\mathcal{F} \subset C([0,1]^d)$ via n-element sample of independent realizations $\{x_i \sim f\}_{i=1}^n$.*

In our course, *we will focus on the Nonparametric regression estimation problem* and related problems of estimating functionals of a "signal" f via observations (1.4).

In order to get a particular "instance" of generic setting (R), we should specify the following "data elements":

1. The grid $\{x_i\}_{i=1}^n$

   Options:

   • n-point equidistant grid;

   • sample of n independent realizations of random vector with known/unknown distribution;

   • ...

2. Type of noises $\{\xi_i\}_{i=1}^n$

   Options:

   • independent $\mathcal{N}(0,1)$-random variables;

   • independent identically distributed random variables with known/unknown distribution;

   • dependent, in a prescribed fashion, random variables;

   • ...

3. The set $\mathcal{F}$

   Options:

   - a subset of $C([0,1]^d)$ comprised of functions satisfying certain smoothness conditions;
   - ...

4. The metric used to measure risk

   In our course, we measure recovering errors in the standard $\| \cdot \|_q$*-norms*

$$\| g \|_q = \begin{cases} \left(\int\limits_{[0,1]^d} |g(x)|^q dx \right)^{1/q}, & 1 \le q < \infty \\ \max\limits_{x \in [0,1]^d} |g(x)|, & q = \infty \end{cases}$$

   The risks associated with these norms are called q-risks.

It would be too ambitious for a single course to be aimed at achieving "maximal generality" with respect to all these "data elements". Our decision will be in favour of "generality in the classes of signals $\mathcal{F}$" rather than "generality with respect to the schemes of observations". Indeed, what makes the major difference between the parametric and the nonparametric statistics, is exactly the "nontrivial" infinite-dimensional geometry of the parameter set, and it is natural to focus first of all on the role of this geometry, not complicating things by considering "difficult" observation schemes. Specifically, the main part of the results to be presented deals with the simplest observation scheme, where the observations are taken along an equidistant grid, and the observation noises are independent $\mathcal{N}(0,1)$[1].

The asymptotic approach. After all "data elements" of the Regression estimation problem (recall that this is the problem we focus on) are specified, our "ideal goal" becomes to find the optimal, for a given volume n of observations (1.4), estimate – the one yielding the minimax risk. *As a matter of fact, this goal never is achievable* – we do not know what is the optimal in the minimax sense estimate even in the simplest – parametric! – problem

> "Recover a real $f \in \mathcal{F} = [0,1]$ via n independent observations $y_i = f + \xi_i$, $\xi_i \sim \mathcal{N}(0,1)$".

Thus, we are enforced to simplify our goal, and the standard "simplification" is to fix all data elements *except the volume of observations* n, to treat n as a varying "large parameter" and to speak about *asymptotically optimal in order/ asymptotically efficient* estimation methods defined as follows.

When n is treated as a varying parameter,

- The minimax risk becomes a function of n:

$$\mathcal{R}^*(n; \mathcal{F}) = \inf_{\widehat{f}_n(\cdot)} \left(\mathcal{E}_{n,f} \left\{ \mathrm{dist}^2(\widehat{f}_n(y^f, \cdot) - f(\cdot)) \right\} \right)^{1/2},$$

[1] Most of these results can be more or less straightforwardly extended to the case of more general schemes of observations, but all these extensions are beyond the scope of the course.

where

• inf is taken over the set of all possible estimates of f via n observations (1.4), i.e., all Borel functions $\widehat{f}(x;y) : [0,1]^d \times \mathbf{R}^n \to \mathbf{R}$

• $\mathcal{E}_{n,f}$ is the expectation over y^f.

• A candidate solution to the Regression estimation problem becomes an *estimation method* – a sequence of estimates

$$\left\{\widehat{f}_n(\cdot,\cdot) : [0,1]^d \times \mathbf{R}^n \to \mathbf{R}\right\}_{n=1}^{\infty}$$

indexed by volumes of observations used by the estimates;

• Our goal becomes either to find an *asymptotically efficient* estimation method:

$$\mathcal{R}(\widehat{f}_n;\mathcal{F}) = (1+o(1))\mathcal{R}^*(n;\mathcal{F}),\ n \to \infty,$$

or, which is more realistic, to find an *optimal in order* estimation method:

$$\mathcal{R}(\widehat{f}_n;\mathcal{F}) \le O(1)\mathcal{R}^*(n;\mathcal{F}),\ n \to \infty.$$

In our course, we focus primarily on building *optimal in order estimation methods for the Regression estimation problem* and *asymptotically efficient estimation of functionals of a regression function*. The only situation we are interested in is when *consistent* estimation is possible – i.e., when the minimax risk itself converges to zero as $n \to \infty$. Note that the latter assumption is satisfied only when $\mathcal{F}$ possesses some compactness properties (see Corollary 1.2.1), and that the rate of convergence of the minimax risk to 0 as $n \to \infty$ heavily depends on the geometry of $\mathcal{F}$ (and sometimes – on the metric used to measure the estimation error). These phenomena are characteristic for Non-parametric Statistics and reflect its "combined" (Statistics + Geometry of functional spaces) nature.

Just to give an impression of a typical result on estimating a non-parametric regression function, we are about to consider the simplest problem of this type – the one of recovering functions from Hölder balls. We start with the situation where the main ideas of the constructions to follow are most transparent, namely, with estimating a univariate Lipschitz continuous function, and then pass to the case of a general Hölder ball.

1.2 Recovering a univariate Lipschitz continuous function

The problem. Assume we are given n noisy observations

$$y_i = f(i/n) + \sigma\xi_i,\ i = 1,...,n \tag{1.5}$$

of a function

$$f(x) : [0,1] \to \mathbf{R},$$

$\{\xi_i\}$ being independent $\mathcal{N}(0,1)$ noises. Our a priori information on f is that f is Lipschitz continuous with a given constant $L > 0$. How to recover the function?

The recovering routine. Our problem is very simple and admits several standard "good" solutions. We shall discuss just one of them, the so called *locally polynomial*, or *window* estimate. The construction is as follows. In order to recover the value of f at a given point $x \in [0,1]$, let us choose somehow a *window* – a segment $B \subset [0,1]$ containing x and including at least one of the observation points $x_i = i/n$. Let us estimate $f(x)$ via the observations from the window as if f were constant in it. The most natural estimate of this type is just the arithmetic mean of the observations from the window:

$$\widehat{f}_B(x;y) = \frac{1}{n(B)} \sum_{i:x_i \in B} y_i, \tag{1.6}$$

where $n(B)$ stands for the number of observation points in a segment B. Recalling the origin of y_i's and taking into account that

$$f(x) = \frac{1}{n(B)} \sum_{i:x_i \in B} f(x),$$

we get

$$\begin{array}{rcl} \mathrm{err}_B(x;y) & \equiv & \widehat{f}_B(x;y) - f(x) \\ & = & d_B(x) + s_B, \\ d_B(x) & = & \frac{1}{n(B)} \sum\limits_{i:x_i \in B} [f(x_i) - f(x)], \\ s_B & = & \frac{1}{n(B)} \sum\limits_{i:x_i \in B} \sigma\xi_i. \end{array} \tag{1.7}$$

We have decomposed the estimation error in two components:

- deterministic *dynamic error* (bias) coming from the fact that f is not constant in the window,

- *stochastic error* s_B coming from observation noises and depending on window, not on f,

and this decomposition allows us to bound the estimation error from above. Indeed, the deterministic error clearly can be bounded as

$$\begin{array}{rcl} |d_B(x)| & \leq & \frac{1}{n(B)} \sum\limits_{i:x_i \in B} |f(x_i) - f(x)| \\ & \leq & \frac{1}{n(B)} \sum\limits_{i:x_i \in B} L|x_i - x| \\ & \leq & L|B|, \end{array} \tag{1.8}$$

where $|B|$ is the length of the segment B.

Now, the stochastic error is a Gaussian random variable with the standard deviation

$$\frac{\sigma}{\sqrt{n(B)}} \leq \sigma_n(|B|) \equiv \frac{\sigma}{\sqrt{n|B|/2}} \tag{1.9}$$

(we have taken into account that the number of observation points in B is at least $n|B|/2$), and we can therefore bound from above all moments of the stochastic error:

$$(\mathcal{E}\{|s_B|^q\})^{1/q} \leq O(1)\sigma_n(|B|)\sqrt{q}, \quad q \geq 1$$

(from now on, $\mathcal{E}$ is the expectation over the observation noise, and all $O(1)$'s are absolute constants). It follows that the moments of the estimation error $\mathrm{err}_B(x; y)$ can be bounded as follows:

$$\begin{array}{rcl} \left(\mathcal{E}\{|\mathrm{err}_B|^q(x;y)\}\right)^{1/q} & \leq & O(1)\sqrt{q}\varepsilon_n(|B|), \\ \varepsilon_n(h) & = & Lh + \sigma\sqrt{\frac{2}{nh}}. \end{array} \tag{1.10}$$

The concluding step is to choose the window width $h = |B|$ which results in the smallest possible $\varepsilon_n(h)$. Since we do not bother much about absolute constant factors, we may just balance the "deterministic" and the "stochastic" components of $\varepsilon_n(h)$:

$$Lh = \sigma(nh)^{-1/2} \Rightarrow h = \left(\frac{\sigma}{L\sqrt{n}}\right)^{2/3}.$$

Thus, we come to the estimation routine as follows:

> (Lip) *Let number of observations n, noise intensity $\sigma > 0$ and a real $L > 0$ be given, and let*
>
> $$h = \left(\frac{\sigma}{L\sqrt{n}}\right)^{2/3}. \tag{1.11}$$
>
> *In order to estimate an unknown regression function f at a point $x \in [0, 1]$ via observations* (1.5), *we*
>
> *– cover x by a segment $B_x \subset [0, 1]$ of the length h*
> *(for the sake of definiteness, let this segment be centered at x, if the distance from x to both endpoints of $[0, 1]$ is $\geq h/2$, otherwise let B_x be either $[0, h]$, or $[1 - h, 1]$, depending on which of the points – 0 or 1 – is closer to x);*
>
> *– take, as an estimate of $f(x)$, the quantity*
>
> $$\widehat{f}_n(x; y) = \frac{1}{n(B_x)} \sum_{i: x_i \in B_x} y_i.$$

Note that the resulting estimate is *linear* in observations:

$$\widehat{f}_n(x; y) = \sum_{i=1}^{n} \phi_{i,n}(x) y_i$$

with piecewise constant "weight functions" $\phi_{i,n}(\cdot)$.

It should be stressed that the above estimation routine is well-defined only in certain restricted domain of values of the parameters L, n, σ. Indeed, the resulting h should not exceed 1 – otherwise the required window will be too large to be contained in $[0, 1]$. At the same time, h should be at least n^{-1}, since otherwise the window may be too small to contain even a single observation point. Thus, the above construction is well-defined only in the case when

$$1 \leq \left(\frac{L\sqrt{n}}{\sigma}\right)^{2/3} \leq n. \tag{1.12}$$

Note that for any fixed pair (L, σ), the relation (1.12) is satisfied for all large enough values of n.

Bounds for q-risks, $1 \le q < \infty$**.** The quality of our estimator is described by the following simple

Proposition 1.2.1 *Let n, L, σ satisfy the restriction* (1.12). *Whenever a regression function f underlying observations* (1.5) *is Lipschitz continuous on* $[0,1]$ *with constant L, the estimate $\widehat{f}_n(\cdot;\cdot)$ given by the estimation routine* (Lip) *satisfies the relations*

$$\left(\mathcal{E}\{\| f - \widehat{f}_n \|_q^2\}\right)^{1/2} \le O(1)\sqrt{q}L\left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2}{3}}, \ 1 \le q < \infty. \tag{1.13}$$

In particular, whenever $1 \le q < \infty$, the (worst-case) q-risk

$$\mathcal{R}_q(\widehat{f}_n; \mathcal{H}_1^1(L)) = \sup_{f\in\mathcal{H}_1^1(L)} \left(\mathcal{E}\{\| \widehat{f}_n - f \|_q^2\}\right)^{1/2}$$

of the estimate $\widehat{f}_n$ on the Lipschitz ball

$$\mathcal{H}_1^1(L) = \{f : [0,1] \to \mathbf{R} \mid |f(x) - f(x')| \le L|x - x'| \quad \forall x, x' \in [0,1]\}$$

can be bounded from above as

$$\mathcal{R}_q(\widehat{f}_n; \mathcal{H}_1^1(L)) \le O(1)\sqrt{q}L\left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2}{3}}. \tag{1.14}$$

Proof. Let $q \in [1,\infty)$. Relations (1.10) and (1.11) imply that for every Lipschitz continuous, with constant L, function f and for every $x \in [0,1]$ one has

$$\begin{array}{rcl}
\mathcal{E}\{|\widehat{f}_n(x;y) - f(x)|^{2q}\} & \le & \left[O(1)\sqrt{q}Lh\right]^{2q} \\
& & \Rightarrow \\
\mathcal{E}\left\{\left[\int\limits_0^1 |\widehat{f}_n(x;y) - f(x)|^q dx\right]^2\right\} & \le & \mathcal{E}\left\{\int\limits_0^1 |\widehat{f}_n(x;y) - f(x)|^{2q} dx\right\} \\
& \le & \left[O(1)\sqrt{q}Lh\right]^{2q} \\
& & \Rightarrow \\
\left(\mathcal{E}\{\| \widehat{f}_n - f \|_q^2\}\right)^{1/2} & \le & O(1)\sqrt{q}Lh \\
& = & O(1)\sqrt{q}L^{1/3}\sigma^{2/3}n^{-1/3} \\
& & \text{[see (1.11)]}
\end{array} \qquad \blacksquare$$

Bound for the ∞-risk. The bounds established in Proposition 1.2.1 relate to q-risks with $q < \infty$ only; as we shall see in a while, these bounds are optimal in order in the minimax sense. In order to get a similarly "good" bound for the ∞-risk, the above construction should be slightly modified. Namely, let us fix h, $1 \ge h \ge n^{-1}$, and consider an estimate of the same structure as (Lip):

$$\widehat{f}^h(x;y) = \frac{1}{n(B_x)} \sum_{i:x_i\in B_x} y_i, \tag{1.15}$$

with all windows B_x being of the same width h [2)]. In view of (1.7), the $\| \cdot \|_\infty$-error of the estimator $\widehat{f}_h$ can be bounded from above by the sum of the maximal, over $x \in [0,1]$, deterministic and stochastic errors:

$$\| \widehat{f}^h - f \|_\infty \leq \left\{ \sup_{x\in[0,1]} |d_{B_x}(x)| \right\}_1 + \left\{ \sup_{x\in[0,1]} |s_{B_x}| \right\}_2 .$$

According to (1.8), the right hand side term $\{\cdot\}_1$ does not exceed Lh. In order to evaluate the term $\{\cdot\}_2$, note that every s_{B_x} is a Gaussian random variable with the zero mean and the standard deviation not exceeding $\sigma_n(h) = \sigma\sqrt{\frac{2}{nh}}$, see (1.9). Besides this, the number of distinct random variables among s_{B_x} does not exceed $O(1)n^2$ (indeed, every stochastic error is the arithmetic mean of several "neighbouring" observation noises $\sigma\xi_i, \sigma\xi_{i+1}, ..., \sigma\xi_j$, and there are no more than $n(n+1)/2$ groups of this type). It follows that

$$\mathcal{E}\left\{\{\cdot\}_2^2\right\} \leq O(1)\sigma_n^2(h)\ln n,$$

whence

$$\left(\mathcal{E}\left\{\| \widehat{f}^h - f \|_\infty^2\right\}\right)^{1/2} \leq O(1)\left[Lh + \frac{\sigma\sqrt{\ln n}}{\sqrt{nh}}\right]. \qquad (1.16)$$

Choosing h which balances the "deterministic" and the "stochastic" terms Lh, $\frac{\sigma\sqrt{\ln n}}{\sqrt{nh}}$, respectively, we get

$$h = \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2}{3}}. \qquad (1.17)$$

Denoting by $\widehat{f}_n^\infty(\cdot)$ the estimate (1.15), (1.17) and applying (1.16), we get the following risk bound:

$$\mathcal{R}_\infty(\widehat{f}_n^\infty; \mathcal{H}_1^1(L)) \equiv \sup_{f\in\mathcal{H}_1^1(L)} \left(\mathcal{E}\left\{\| \widehat{f}_n^\infty(\cdot) - f(\cdot) \|_\infty^2\right\}\right)^{1/2} \leq O(1)L\left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2}{3}}. \qquad (1.18)$$

Note that the construction makes sense only when h given by (1.17) belongs to the segment $[n^{-1}, 1]$, i.e., when

$$1 \leq \left(\frac{L\sqrt{n}}{\sigma\sqrt{\ln n}}\right)^{2/3} \leq n, \qquad (1.19)$$

which for sure is the case for all large enough values of n.

Note that the q-risks of the estimate $\widehat{f}_n^\infty(\cdot)$, $1 \leq q < \infty$, are worse than those of the estimate $\widehat{f}_n$ by a logarithmic in n factor only; similarly, the ∞-risk of the estimate $\widehat{f}_n$ is only by a logarithmic in n factor worse than the ∞-risk of the estimate $\widehat{f}_n^\infty$.

Lower bounds. We have build two estimates $\widehat{f}_n$, $\widehat{f}_n^\infty$ for recovering a Lipschitz continuous, with a known constant, function from observations (1.5). It is time now to demonstrate that these estimates are optimal in order in the minimax sense:

[2)] Same as in (Lip), $B_x = \begin{cases} [0,h], & 0 \leq x \leq h/2 \\ [x-h/2, x+h/2], & h/2 \leq x \leq 1-h/2 \\ [1-h, 1 & 1-h/2 \leq x \leq 1 \end{cases}.$

Proposition 1.2.2 *For every triple L, σ, n satisfying* (1.12) *and every $q \in [1, \infty)$ the minimax q-risk of estimating functions from $\mathcal{H}_1^1(L)$ via observations* (1.5) *can be bounded from below as*

$$\mathcal{R}_q^*(n; \mathcal{H}_1^1(L)) \geq O(1) L \left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2}{3}}. \tag{1.20}$$

For every fixed $\kappa > 0$, for every triple L, σ, n satisfying the assumption

$$n^\kappa \leq \left(\frac{L\sqrt{n}}{\sigma\sqrt{\ln n}}\right)^{\frac{2}{3}} \leq n \tag{1.21}$$

(cf. (1.19)*), the minimax ∞-risk of estimating functions from $\mathcal{H}_1^1(L)$ via observations* (1.5) *can be bounded from below as*

$$\mathcal{R}_\infty^*(n; \mathcal{H}_1^1(L)) \geq C(\kappa) L \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2}{3}} \tag{1.22}$$

($C(\kappa) > 0$ depends on κ only).

Consequently, in the case of (1.12) *the estimate $\widehat{f}_n$ is minimax optimal, up to a factor depending on q only, with respect to q-risk, $1 \leq q < \infty$, on the set $\mathcal{H}_1^1(L)$. Similarly, in the case of* (1.21) *the estimate $\widehat{f}_n^\infty$ is minimax optimal, up to a factor depending on κ only, with respect to the ∞-risk on the same set.*

The proof of this Proposition, same as basically all other lower bounds in regression estimation, is based on information-type inequalities. It makes sense to summarize these arguments in the following statement:

Proposition 1.2.3 *Let*

- *$\mathcal{L}$ be a space of real-valued functions on a set $\mathcal{X}$, and $\rho(f,g)$ be a metric on the functional space $\mathcal{L}$;*
- *$\mathcal{F}$ be a subset of the space $\mathcal{L}$;*
- *X_n be an n-point subset of X;*
- *$\mathcal{F}_N = \{f_1, ..., f_N\}$ be an N-element subset of $\mathcal{F}$;*
- *σ be a positive real.*

Given the indicated data, let us set

$$\begin{array}{rcl}\mathrm{Resolution}(\mathcal{F}_N) & = & \min\left\{\rho(f_i, f_j) \mid 1 \leq i < j \leq N\right\}; \\ \mathrm{Diameter}(\mathcal{F}_N|X_n) & = & \frac{1}{2}\max\limits_{1\leq i\leq j\leq N} \sum\limits_{x\in X_n} |f_i(x) - f_j(x)|^2\end{array}$$

and assume that

$$\mathrm{Diameter}(\mathcal{F}_N|X_n) < \sigma^2\left[\frac{1}{2}\ln(N-1) - \ln 2\right]. \tag{1.23}$$

Now consider the problem of recovering a function $f \in \mathcal{F}$ from n observations

$$y_f = \{y_f(x) = f(x) + \sigma\xi_x\}_{x \in X_n},$$

$\xi = \{\xi_x\}_{x \in X_n}$ *being a collection of independent* $\mathcal{N}(0,1)$ *noises, and let* $\tilde{f}$ *be an arbitrary estimate*[3)]*. Then the worst-case ρ-risk*

$$\mathcal{R}^\rho(\tilde{f};\mathcal{F}) \equiv \sup_{f \in \mathcal{F}} \mathcal{E}\{\rho(f(\cdot), \tilde{f}(\cdot, y_f))\}$$

of the estimate $\tilde{f}$ on $\mathcal{F}$ can be bounded from below as

$$\mathcal{R}^\rho(\tilde{f};\mathcal{F}) \geq \frac{1}{4}\mathrm{Resolution}(\mathcal{F}_N). \tag{1.24}$$

Corollary 1.2.1 *Let $\mathcal{L}$ be a space of real-valued functions on a set $\mathcal{X}$, ρ be a metric on $\mathcal{L}$ and $\mathcal{F}$ be a subset of $\mathcal{L}$. Assume that functions from $\mathcal{F}$ are uniformly bounded and that $\mathcal{F}$ is not pre-compact with respect to ρ: there exists a sequence $\{f_i \in \mathcal{F}\}_{i=1}^\infty$ and $\varepsilon > 0$ such that $\rho(f_i, f_j) \geq \varepsilon$ for all $i \neq j$. Then $\mathcal{F}$ does not admit consistent estimation: for every sequence $\{X_n \subset \mathcal{X}\}_{n=1}^\infty$ of finite subsets of $\mathcal{X}$,* $\mathrm{Card}(X_n) = n$*, the minimax ρ-risk*

$$\mathcal{R}^*_\rho(n;\mathcal{F}) \equiv \inf_{\tilde{f}_n} \sup_{f \in \mathcal{F}} \mathcal{E}\{\rho(f(\cdot), \tilde{f}(\cdot, y_f))\}$$

of estimating $f \in \mathcal{F}$ via observations

$$y_f = \{y_f(x) = f(x) + \sigma\xi_x\}_{x \in X_n} \qquad [\xi_i \sim \mathcal{N}(0,1) \text{ are independent}]$$

remains bounded away from 0 as $n \to \infty$:

$$\mathcal{R}^*_\rho(n;\mathcal{F}) \geq \frac{1}{4}\varepsilon. \tag{1.25}$$

Proof. Under the premise of Corollary there exist subsets $\mathcal{F}_N \subset \mathcal{F}$ of arbitrary large cardinality N with $\mathrm{Resolution}(\mathcal{F}_N) \geq \varepsilon$ and bounded, by a constant depending on $\mathcal{F}$ only, $\mathrm{Diameter}(\mathcal{F}_N|X_n)$ (since all functions from $\mathcal{F}$ are uniformly bounded). It follows that for every n we can find $\mathcal{F}_N \subset \mathcal{F}$ satisfying (1.23) and such that the associated lower bound (1.24) implies (1.25).

Proof of Proposition 1.2.3. Consider N hypotheses H_i, $i = 1, ..., N$, on the distribution of a random vector $y \in \mathbf{R}^n$; according to i-th of them, the distribution is the one of the vector y_{f_i}, i.e., n-dimensional Gaussian distribution $F_i(\cdot)$ with the covariance matrix $\sigma^2 I$ and the mean $\bar{f}_i$, $\bar{f}_i$ being the restriction of f_i onto X_n. Assuming that there exists an estimate $\tilde{f}$ which does *not* satisfy (1.24), let us build a routine $\mathcal{S}$ for distinguishing between these hypotheses:

[3)] Here an estimate is a function

$$\tilde{f}(x, y) : X \times \mathbf{R}^n \to \mathbf{R}$$

such that $\tilde{f}(\cdot, y) \in \mathcal{L}$ for all $y \in \mathbf{R}^n$ and the function $\rho(f(\cdot), \tilde{f}(\cdot, y))$ is Borel in y for every $f \in \mathcal{F}$

Given observations y, we build the function $f^y(\cdot) = \tilde{f}(\cdot, y) \in \mathcal{L}$ and check whether there exists $i \leq N$ such that

$$\rho(f^y(\cdot), f_i(\cdot)) < \frac{1}{2}\text{Resolution}(\mathcal{F}_N).$$

If it is the case, then the associated i is uniquely defined by the observations (by definition of Resolution), and we accept the hypothesis H_i, otherwise we accept, say, the hypothesis H_1.

Note that since $\tilde{f}$ does not satisfy (1.24), then for every $i \leq N$ the probability to accept hypothesis H_i if it indeed is true is $\geq 1/2$ (recall that $\mathcal{F}_N \subset \mathcal{F}$ and use the Tschebyshev inequality). On the other hand, the *Kullback distance*

$$\mathcal{K}(F_i : F_j) \equiv \int\limits_{\mathbf{R}^n} \ln\left(\frac{dF_j(y)}{dF_i(y)}\right) dF_j(y)$$

between the distributions F_i and F_j is at most $\sigma^{-2}\text{Diameter}(\mathcal{F}_N|X_n)$:

$$\begin{array}{rcl}
\mathcal{K}(F_i : F_j) & = & \int \left(\frac{\| y - \bar{f}_i \|_2^2 - \| y - \bar{f}_j \|_2^2}{2\sigma^2}\right) (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{\| y - \bar{f}_j \|_2^2}{2\sigma^2}\right\} dy \\
& = & \int \left(\frac{\| z - [\bar{f}_i - \bar{f}_j] \|_2^2 - \| z \|_2^2}{2\sigma^2}\right) (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{\| z \|_2^2}{2\sigma^2}\right\} dz \\
& = & \int \left(\frac{\| \bar{f}_i - \bar{f}_j \|_2^2 - 2z^T[\bar{f}_i - \bar{f}_j]}{2\sigma^2}\right) (2\pi)^{-n/2}\sigma^{-n} \exp\left\{-\frac{\| z \|_2^2}{2\sigma^2}\right\} dz \\
& = & \frac{\|\bar{f}_i - \bar{f}_j\|_2^2}{2\sigma^2}.
\end{array}$$

It remains to make use of the following fundamental

Theorem 1.2.1 [Fano's inequality, Fano '61] *Let $(\Omega, \mathcal{F})$ be a Polish space with the Borel σ-algebra, let $F_1, ..., F_N$ be N mutually absolutely continuous probability distributions on $(\Omega, \mathcal{F})$. Let also*

$$\mathcal{K}(F_i : F_j) = \int\limits_{\Omega} \ln\left(\frac{dF_j(\omega)}{dF_i(\omega)}\right) dF_j(\omega)$$

be the Kullback distance from F_j to F_i, and let

$$\mathcal{K} = \max_{i,j} \mathcal{K}(F_i : F_j).$$

Given a positive integer m, consider N hypotheses on the distribution of a random point $\omega^m \in \Omega^m$, i-th of the hypotheses being that the distribution is F_i^m (i.e., that the entries $\omega_1, ..., \omega_m$ of ω^m are mutually independent and distributed according to F_i). Assume that for some reals $\delta_i \in (0, 1)$, $i = 1, ..., N$, there exists a decision rule – a Borel function

$$\mathcal{D} : \Omega^m \to [\overline{1, N}] = \{1, 2, ..., N\}$$

– such that the probability to accept i-th hypothesis if it indeed is true is at least δ_i:

$$F_i^m(\{\omega^m : \mathcal{D}(\omega^m) = i\}) \geq \delta_i, \; i = 1, ..., N.$$

Then for every probability distribution $\{p(i)\}_{i=1}^{N}$ *on* $[\overline{1,N}]$ *it holds*

$$\begin{array}{rcl} m\mathcal{K} & \geq & -\sum\limits_{i} p(i)\ln p(i) - (1-\sum\limits_{i} p(i)\delta(i))\ln(N-1) - \ln 2, \\ \theta & = & \sum\limits_{i} p(i)\delta(i). \end{array} \tag{1.26}$$

In particular,

$$\begin{array}{rcl} m\mathcal{K} & \geq & \theta_* \ln(N-1) - \ln 2, \\ \theta_* & = & \frac{1}{N}\sum\limits_{i} \delta(i). \end{array} \tag{1.27}$$

As we have seen, for the routine $\mathcal{S}$ we have built the probabilities to accept every one of the hypotheses H_i if it is true are at least 1/2. Besides this, we have seen that for the hypotheses in question $\mathcal{K} \leq \sigma^{-2}\text{Diameter}(\mathcal{F}_N|X_n)$. Applying (1.27) with $m = 1$, we get

$$\sigma^{-2}\text{Diameter}(\mathcal{F}_N|X_n) \geq \frac{1}{2}\ln(N-1) - \ln 2,$$

which is impossible by (1.23). ■

Proof of Proposition 1.2.2. A. In order to prove (1.20), let us fix $q \in [1,\infty)$ and specify the data of Proposition 1.2.3 as follows:

- $\mathcal{L} = L_q[0,1]$, $\rho(f,g) = \| f-g \|_q$;
- $\mathcal{F} = \mathcal{H}_1^1(L)$;
- $X_n = \{i/n, i = 1,...,n\}$.

It remains to define a candidate to the role of $\mathcal{F}_N$. To this end let us choose somehow a positive $h < 1$ (our choice will be specified later). Note that we can find a collection of

$$M = M(h) \geq \frac{1}{2h}$$

non-overlapping segments $B_l \in [0,1]$, $l = 1,...,M$, of the length h each. Now consider functions f as follows:

$$\begin{array}{c} f \text{ is zero outside } \bigcup\limits_{l=1}^{M} B_l, \text{ and in every segment } B_l = [x_l - h/2, x_l + h/2] \\ \text{the function is either } L[0.5h - |x - x_l|], \text{ or } -L[0.5h - |x - x_l|]. \end{array}$$

It is clear that there exist $2^{M(h)}$ distinct functions of this type, and all of them belong to $\mathcal{H}_1^1(L)$. Moreover, it is easily seen that one can find a collection $\mathcal{F}_{N(h)} = \{f_1,...,f_N\}$ comprised of

$$N(h) \geq 2^{O(1)M(h)} \tag{1.28}$$

functions of the indicated type in such a way that for distinct i,j the number $n(i,j)$ of those segments B_l where f_i differs from f_j is at least $O(1)M(h)$. It is immediately seen that the latter property implies that

$$i \neq j \Rightarrow \| f_i - f_j \|_q \geq O(1)Lh(O(1)M(h)h)^{1/q} \geq O(1)Lh,$$

so that

$$\text{Resolution}(\mathcal{F}_{N(h)}) \geq O(1)Lh. \tag{1.29}$$

Now let us specify h in a way which ensures (1.23) for $\mathcal{F}_N = \mathcal{F}_{N(h)}$. The uniform distance between any two functions f_i, f_j does not exceed Lh, hence

$$\mathrm{Diameter}(\mathcal{F}_{N(h)}|X_n) \le L^2h^2n. \tag{1.30}$$

In view of (1.28), for $N = N(h)$ the right hand side of (1.23) is at least $O(1)\sigma^2h^{-1}$, provided that h is less than a small enough absolute constant. On the other hand, by (1.30) the right hand side of (1.23) for $\mathcal{F}_N = \mathcal{F}_{N(h)}$ is at most nL^2h^2. We see that in order to ensure (1.23) for $\mathcal{F}_N = \mathcal{F}_{N(h)}$ it suffices to set

$$h = O(1)\min\left[1, n^{-1/3}L^{-2/3}\sigma^{2/3}\right] = O(1)n^{-1/3}L^{-2/3}\sigma^{2/3},$$

the concluding relation being given by (1.12). In view of (1.29), with this choice of h Proposition 1.2.3 yields (1.20).

B. In order to prove (1.22), one can use a construction similar to the one of **A.** Namely, let us set

- $\mathcal{L} = L_\infty[0,1]$, $\rho(f,g) = \| f - g \|_\infty$;
- $\mathcal{F} = \mathcal{H}_1^1(L)$;
- $X_n = \{i/n, i = 1, ..., n\}$,

choose $h \in [0,1)$ and build

$$M(h) \ge O(1)h^{-1}$$

non-overlapping segments $B_j = [x_j - h/2, x_j + h/2] \subset [0,1]$. Associating with j-th segment the function

$$f_j(x) = \begin{cases} 0, & x \notin B_j \\ L[0.5h - |x - x_j|], & x \in B_j \end{cases},$$

we get a collection $\mathcal{F}_{M(h)}$ of $M(h)$ functions such that

$$\mathrm{Resolution}(\mathcal{F}_{M(h)}) = 0.5Lh \tag{1.31}$$

and

$$\mathrm{Diameter}(\mathcal{F}_{M(h)}|X_n) \le O(1)L^2h^3n$$

(indeed, the difference of two functions from $\mathcal{F}_{M(h)}$ is of the uniform norm at most $0.5Lh$ and differs from zero at no more than $O(1)nh$ point of the grid X_n). We see that for $\mathcal{F}_N = \mathcal{F}_{M(h)}$ the left hand side in (1.23) is at most $O(1)L^2h^3n$, while the right hand side is at least $O(1)\sigma^2\ln M(h) = O(1)\sigma^2\ln h^{-1}$, provided that h is less than a small enough absolute constant. It follows that in order to ensure (1.23) it suffices to choose h less than an appropriate absolute constant and satisfying the relation

$$L^2h^3n \le O(1)\sigma^2\ln h^{-1}.$$

In the case of (1.21) the latter requirement, in turn, is satisfied by

$$h = d(\kappa)\left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2}{3}}$$

with properly chosen $d(\kappa) > 0$ (depending on κ only). With this h, Proposition 1.2.3 combined with (1.31) yields the bound (1.22). ■

1.3 Extension: recovering functions from Hölder balls

The constructions and results related to recovering univariate Lipschitz continuous functions can be straightforwardly extended to the case of general Hölder balls.

Hölder ball $\mathcal{H}^s_d(L)$ is specified by the parameters $s > 0$ (order of smoothness), $d \in \mathbf{Z}_+$ (dimension of the argument) and $L > 0$ (smoothness constant) and is as follows. A positive real s can be uniquely represented as

$$s = k + \alpha, \tag{1.32}$$

where k is a nonnegative integer and $0 < \alpha \leq 1$. By definition, $\mathcal{H}^s_d(L)$ is comprised of all k times continuously differentiable functions

$$f : [0,1]^d \to \mathbf{R}$$

with Hölder continuous, with exponent α and constant L, derivatives of order k:

$$|D^k f(x)[h,...,h] - D^k f(x')[h,...,h]| \leq L|x - x'|^\alpha |h|^k \quad \forall x, x' \in [0,1]^d \forall h \in \mathbf{R}^d.$$

Here $|\cdot|$ is the standard Euclidean norm on $\mathbf{R}^d$, and $D^k f(x)[h_1,...,h_k]$ is k-th differential of f taken at a point x along the directions $h_1,...,h_k$:

$$Df^k(x)[h_1,...,h_k] = \left.\frac{\partial^k}{\partial t_1 ... \partial t_k}\right|_{t_1=t_2=...=t_k=0} f(x + t_1h_1 + t_2h_2 + ... + t_kh_k).$$

Note that $\mathcal{H}^1_d(L)$ is just the set of all Lipschitz continuous, with constant L, functions on the unit d-dimensional cube $[0,1]^d$.

The problem we now are interested in is as follows. Assume we are given $n = m^d$ noisy observations

$$\begin{array}{c} y = y_f(\xi) = \left\{ y_\iota = f(x_\iota) + \sigma\xi_\iota | \iota = (i_1,...,i_d) \in \overline{[1,m]}^d \right\} \\ \left[x_{(i_1,...,i_d)} = (i_1/m, i_2/m, ..., i_d/m)^T \right] \end{array} \tag{1.33}$$

of unknown regression function f; here $\{\xi_\iota\}$ are independent $\mathcal{N}(0,1)$ noises. All we know in advance about f is that the function belongs to a given Hölder ball $\mathcal{H}^s_d(L)$, and our goal is to recover the function from the observations.

The recovering routine we are about to present is quite similar to the one of the previous Section. Namely, we fix a "window width" h such that

$$\frac{k+2}{m} \leq h \leq 1, \tag{1.34}$$

k being given by (1.32). In order to estimate the value $f(x)$ of f at a point $x \in [0,1]^d$, we choose somehow a "window" – a cube $B_x \subset [0,1]^d$ such that $x \in B_x$ and the edges of B_x are equal to h, and estimate $f(x)$ via observations from the window as if f was a polynomial of degree k. Let us explain the exact meaning of the latter sentence.

Estimating polynomials. Let $B_h = \{x \in \mathbf{R}^d \mid a_i \le x_i \le a_i + h, i = 1, ..., d\}$ be a d-dimensional cube with edges $h > 0$, and let $\Gamma_\delta = \delta \mathbf{Z}^d$ be the regular grid with resolution $\delta > 0$. Assume that

$$\frac{h}{(k+2)\delta} \ge 1, \tag{1.35}$$

and let B_h^δ be the intersection of the cube B_h and the grid Γ_δ. Let also $\mathcal{P}_d^k$ be the space of all polynomials of (full) degree k of d variables. Consider the following auxiliary problem:

> (*) *Given $x \in B_h$, find "interpolation weights" $\omega = \{\omega(u)\}_{u \in B_h^\delta}$ which reproduce the value at x of every polynomial of degree k via its restriction on B_h^δ:*
>
> $$p(x) = \sum_{u \in B_h^\delta} \omega(u) p(u) \quad \forall p \in \mathcal{P}_d^k \tag{1.36}$$
>
> *with the smallest possible variance*
>
> $$\| \omega \|_2^2 = \sum_{u \in B_h^\delta} \omega^2(u).$$

Lemma 1.3.1 *Problem (*) is solvable, and its optimal solution ω_x is unique and continuously depends on x. Moreover,*

$$\| \omega_x \|_2^2 \le \kappa_2(k,d) \left(\frac{\delta}{h}\right)^d \tag{1.37}$$

and

$$\| \omega_x \|_1 \equiv \sum_{u \in B_h^\delta} |\omega(u)| \le \kappa_1(k,d) \tag{1.38}$$

with factors $\kappa_{1,2}(k,d)$ depending on k, d only.

Proof. 1^0. Observe, first, that if G_i, $i = 1, ..., d$, are finite sets of reals, each of the sets being comprised of $l_i \ge k+1$ equidistantly placed points, and

$$G^d = G_1 \times G_2 \times ... \times G_d,$$

then the only polynomial from $\mathcal{P}_d^k$ vanishing at the grid G^d is the zero polynomial (this observation is given by a straightforward induction in d). In other words, if $p^1, ..., p^N$ is a basis in $\mathcal{P}_d^k$ and P is the matrix with columns being the restrictions of the basic polynomials $p^i_{G^d}$ on the grid G^d:

$$P = [p^1_{G^d}; ...; p^N_{G^d}],$$

then the kernel of the matrix P is trivial. Denoting by $\widehat{p}$ the vector of coefficients of a polynomial $p \in \mathcal{P}_d^k$ in the basis $p^1, ..., p^N$ and observing that

$$p_{G^d} = P\widehat{p} \quad \forall p \in \mathcal{P}_d^k,$$

we conclude that $\widehat{p}$ can be expressed, in a linear fashion, via p_{G^d}. Consequently, the value of $p \in \mathcal{P}_d^k$ at a given point u can also be expressed as a linear function of p_{G^d}:

$$\exists \lambda : \quad \lambda^T p_{G^d} = p(u) \quad \forall p \in \mathcal{P}_d^k.$$

The corresponding vectors of coefficients λ are exactly the solutions to the linear system

$$P^T\lambda = \begin{pmatrix} p^1(u) \\ \cdots \\ p^N(u) \end{pmatrix} \tag{1.39}$$

As we have seen, (1.39) is solvable, and the matrix P is with the trivial kernel; under these conditions Linear Algebra says that the matrix P^TP is non-singular and that the (unique) least norm solution to (1.39) is given by

$$\lambda_u = P(P^TP)^{-1} \begin{pmatrix} p^1(u) \\ \cdots \\ p^N(u) \end{pmatrix}.$$

In particular, λ_u is a continuous function of u.

2^0. In view of (1.35), the set B_h^δ is a grid of the type considered in 1^0; in view of the results of 1^0, the weight vector ω_x is well-defined and is continuous in x.

3^0. To prove (1.37), let us come back to the situation of 1^0 and assume for a moment that the cardinality of every "partial grid" G_i is exactly $k+1$, and the convex hull of the grid is the segment $[0,1]$. In this case the norms $\parallel \lambda_u \parallel_2$ of the weight vectors λ_u, being continuous functions of u, are bounded in the cube

$$-1 \le u_i \le 2, \;\; i = 1, ..., d$$

by certain constant $C_1(k,d)$ depending on k, d only. By evident similarity reasons we conclude that if the partial grids G_i are arbitrary equidistant grids of the cardinality $k+1$ each, the parallelotope $B(G^d)$ is the convex hull of G^d and $B^+(G^d)$ is the concentric to $B(G^d)$ three times larger parallelotope, then for the corresponding weight vectors it holds

$$\parallel \lambda_u \parallel_2 \le C_1(k,d) \quad \forall u \in B^+(G^d). \tag{1.40}$$

Let q be the largest integer such that $q(k+2)\delta \le h$; note that by (1.35) we have $q \ge 1$. As we just have mentioned, the grid B_h^δ is a direct product of d partial equidistant grids $\widehat{G}_i$, and the cardinality of every one of these grids is at least $q(k+1)$. For every i, let us partition the grid $\widehat{G}_i$ into q mutually disjoint equidistant sub-grids $G_{i,l}$, $l = 1, ..., q$ of cardinality $k+1$ each as follows: $G_{i,l}$ contains the l-th, the $(l+q)$-th,...,the $(l+kq)$-th points of the grid $\widehat{G}_i$. For every collection $\nu = (\nu_1, ..., \nu_d)$ of integers $\nu_i \in \overline{[1,q]}$, we can build the d-dimensional grid

$$G_\nu^d = G_{1,\nu_1} \times G_{2,\nu_2} \times ... \times G_{d,\nu_d}.$$

By construction, all q^d d-dimensional grids we can get in this way from q^d distinct collections ν are mutually disjoint and are contained in B_h^δ. Moreover, it is easily seen that every one of the parallelotopes $B^+(G_\nu^d)$ contains B_h. As we just have seen, for every ν there exists a representation

$$p(x) = \sum_{u \in G_\nu^d} \lambda_\nu(u) p(u) \quad \forall p \in \mathcal{P}_d^k$$

with

$$\sum_{u \in G_\nu^d} \lambda_\nu^2(u) \le C_1^2(k,d).$$

It follows that for every $p \in \mathcal{P}_d^k$ it holds

$$p(x) = \sum_{u \in B_h^\delta} \omega(u) p(u) \equiv \frac{1}{q^d} \sum_\nu \sum_{u \in G_\nu^d} \lambda_\nu(u) p(u).$$

The variance of the resulting interpolation weights clearly is

$$\frac{1}{q^{2d}} \sum_\nu \sum_{u \in G_\nu^d} \lambda_\nu^2(u) \le \frac{1}{q^d} C_1^2(k,d) \le C_2(k,d)(\delta/h)^d$$

(we have used (1.40) and the fact that $q \ge O(1) h\delta^{-1}(k+2)^{-1}$). Since the variance of the optimal interpolation weights (those coming from the optimal solution to (*)) cannot be worse than the variance of the weights we just have built, we come to (1.37). It remains to note that (1.38) follows from (1.37) in view of the Cauchy inequality. ■

Window estimates. The simple technique for estimating polynomials we have developed gives rise to a useful construction we shall use a lot – the one of a *window estimate* of $f(x)$ via observations (1.33). For a given volume of observations n, such an estimate is specified by its *order* k (which is a nonnegative integer) and a *window* B (recall that this is a cube containing x and contained in $[0,1]^d$ with the edges of a length h satisfying (1.34)) and is as follows. Let B^n be the intersection of B with the observation grid. In view of Lemma 1.3.1, problem (*) associated with the data $x, B_h = B, B_h^\delta = B^n, k$ is solvable; its optimal solution is certain collection of weights

$$\omega \equiv \omega_x^B = \left\{ \omega_\iota^B(x) \mid \iota : x_\iota \in B \right\}.$$

The *order* k *window estimate of* $f(x)$ *associated with the cube* B is

$$\widehat{f}(x;y) \equiv \widehat{f}_n^B(x;y) = \sum_{\iota : x_\iota \in B_x} \omega_\iota^B(x) y_\iota. \tag{1.41}$$

The following proposition summarizes some useful properties of window estimates.

Proposition 1.3.1 *Let $x \in [0,1]^d$, k, n be given, and let B be a window for x. Given a continuous function $f : [0,1]^d \mapsto \mathbf{R}$, let us define $\Phi_k(f,B)$ as the smallest uniform error of approximating f in B by a polynomial of degree $\le k$:*

$$\Phi_k(f,B) = \min_{p \in \mathcal{P}_d^k} \max_{u \in B} |f(u) - p(u)|.$$

Then the error of the order k window estimate of $f(x)$ associated with the window B can be bounded as follows:

$$|\widehat{f}_n^B(x; y_f(\xi)) - f(x)| \le O_{k,d}(1) \left[\Phi_k(f,B) + \frac{\sigma}{\sqrt{n}} D^{-d/2}(B) |\zeta_x^B(\xi)| \right], \tag{1.42}$$

where

- *$D(B)$ is the edge of the cube B;*
- *$\zeta_x^B(\xi)$ is a linear combination of the noises $\{\xi_\iota\}$ with variance 1.*

Here in what follows $O_{\dots}(1)$ denotes a positive quantity depending solely on the parameter(s) listed in the subscript.

Furthermore, let

$$\Theta_n = \Theta_n(\xi) = \sup\left\{|\zeta_x^B(\xi)| \mid x \in [0,1]^d,\ B \textit{ is a window for } x\right\}$$

Then the random variable Θ_n is "of order of $\sqrt{\ln n}$":

$$\forall w \geq 1: \qquad \mathrm{Prob}\left\{\Theta_n > O_{k,d}(1)w\sqrt{\ln n}\right\} \leq \exp\left\{-\frac{w^2 \ln n}{2}\right\}. \tag{1.43}$$

Proof. Let $n(B)$ be the number of observation points in the window B and $p(\cdot)$ be a polynomial of degree $\leq k$ such that

$$\max_{u\in B} |f(u) - p(u)| = \Phi_k(f, B).$$

We have

$$\begin{array}{rcl} & & |f(x) - \widehat{f}_n^B(x; y_f(\xi))| \\ & = & |f(x) - \sum\limits_{\iota: x_\iota \in B} \omega_\iota^B(x)\left[p(x_\iota) + [f(x_\iota) - p(x_\iota)]\right] + \sum\limits_{\iota: x_\iota \in B} \omega_\iota^B(x)\sigma\xi_\iota| \\ & = & |f(x) - p(x) + \sum\limits_{\iota: x_\iota \in B} \omega_\iota^B(x)\left[f(x_\iota) - p(x_\iota)\right] + \sum\limits_{\iota: x_\iota \in B} \omega_\iota^B(x)\sigma\xi_\iota| \\ & & [\text{by } (1.36)] \\ & \leq & |f(x) - p(x)| + \sum\limits_{\iota: x_\iota \in B} |\omega_\iota^B(x)||f(x_\iota) - p(x_\iota)| + |\sum\limits_{\iota: x_\iota \in B} \omega_\iota^B(x)\sigma\xi_\iota| \\ & \leq & \Phi_k(f, B)\left[1 + \| \omega_x^B \|_1\right] + \sigma \| \omega_x^B \|_2 |\zeta_x^B|, \\ \zeta_x^B & = & \frac{1}{\|\omega_x^B\|_2} \sum\limits_{\iota: x_\iota \in B} \omega_\iota^B(x)\xi_\iota. \end{array} \tag{1.44}$$

By Lemma 1.3.1 (applied with $\delta = n^{-1/d}$) one has

$$\| \omega_x^B \|_1 \leq \kappa_1(k, d),\ \| \omega_x^B \|_2 \leq \kappa_2(k, d) n^{-1/2} D^{-d/2}(B),$$

and (1.44) implies (1.42).

The proof of (1.43) is left to the reader; we just indicate that the key argument is that, as it is immediately seen from the proof of Lemma 1.3.1, for fixed B the weights $\omega_\iota^B(x)$ are polynomials of x of degree $\leq k$. ■

From estimating polynomials to estimating functions from Hölder balls. Let us estimate $f(\cdot)$ at every point by a window estimate, all windows being of the same size; the underlying length of window edges – the "window width" h – is the parameter of our construction. Let us specify somehow the correspondence $x \mapsto B_x$, B_x being the window used to estimate $f(x)$; we may, e.g., use the "direct product" of the rules used in the univariate case. Let $\widehat{f}^h(\cdot; y)$ denote the resulting estimate of the regression function f underlying observations y (see (1.33)).

Bounds for q-risks, $1 \le q < \infty$. Observe that for $f \in \mathcal{H}^s_d(L)$ and all cubes $B \subset [0,1]^d$ we clearly have

$$\Phi_k(f,B) \le O_{k,d}(1) L D^s(B) \tag{1.45}$$

(the right hand side is just the standard upper bound for the error, on B, of approximating f by its Taylor polynomial of the degree k taken at a point from B). From this observation and (1.44) it follows that for the window estimate $\widehat{f}^h(x)$ we have

$$\left(\mathcal{E}\{|f(x) - \widehat{f}^h(x;y)|^q\}\right)^{1/q} \le O_{k,d}(1)\sqrt{q}\left[Lh^s + \frac{\sigma}{\sqrt{nh^d}}\right], \tag{1.46}$$

provided that h satisfies (1.34).

Now let us choose the window width h which balances the terms Lh^s and $\frac{\sigma}{\sqrt{nh^d}}$ in the right hand side of (1.46):

$$h = \left(\frac{\sigma}{L\sqrt{n}}\right)^{2/(2s+d)}. \tag{1.47}$$

Assuming that the resulting h satisfies (1.34), i.e., that

$$1 < \left(\frac{L\sqrt{n}}{\sigma}\right)^{\frac{2d}{2s+d}} \le (k+2)^{-d} n, \tag{1.48}$$

(cf. (1.12); note that for every pair of (positive) L, σ this relation is satisfied by all large enough values of n), we come to certain estimate, let it be denoted by $\widehat{f}_n(x;y)$. In view of (1.46), the q-risk of this estimate on $\mathcal{H}^s_d(L)$ can be bounded as follows:

$$\mathcal{R}_q(\widehat{f}_n; \mathcal{H}^s_d(L)) \le O_{s,d}(1)\sqrt{q} L \left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2s}{2s+d}}. \tag{1.49}$$

Note that our estimate, same as in the univariate case, is linear in observations:

$$\widehat{f}_n(x;y) = \sum_\iota \phi_{\iota,n}(x) y_\iota.$$

Bound for ∞-risk. When interested in the estimate of the outlined type with ∞-risk being as small as possible, we should choose the window width h in a way slightly different from (1.47) (same as we did so in the previous Section). Indeed, for $f \in \mathcal{H}^s_d(L)$, the uniform risk of the estimate $\widehat{f}^h$, in view of (1.45), (1.44), can be bounded as

$$\| f(\cdot) - \widehat{f}(\cdot, y) \|_\infty \le O_{k,d}(1)\left[Lh^s + \frac{\sigma}{\sqrt{n}}\Theta_n\right]. \tag{1.50}$$

As we know from (1.43), the "typical values" of Θ_n are of order of $\sqrt{\ln n}$. Consequently, a reasonable choice of h should balance the "deterministic term" Lh^s and the "typical value" $\sigma n^{-1/2}\sqrt{\ln n}$ of the "stochastic term" in the right hand side of (1.50). We come to the choice

$$h = \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{2/(2s+d)}. \tag{1.51}$$

Assume that this choice fits (1.34), i.e., that

$$1 < \left(\frac{L\sqrt{n}}{\sigma\sqrt{\ln n}}\right)^{\frac{2d}{2s+d}} < (k+2)^{-d} n, \tag{1.52}$$

and let us denote the resulting estimate $\widehat{f}_n^\infty$. From (1.43) combined with (1.50) we get the following bound on the ∞-risk of the estimate on $\mathcal{H}_d^s(L)$:

$$\mathcal{R}_\infty(\widehat{f}_n^\infty; \mathcal{H}_d^s(L)) \le O_{s,d}(1) L \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2s}{2s+d}}. \tag{1.53}$$

Lower bounds on the minimax q-risks of recovering functions from a Hölder ball $\mathcal{H}_d^s(L)$ are given by essentially the same reasoning as in the particular case considered in the previous Section; they are particular cases of the bounds from Theorem 2.1.1 proved in Section 2.3. We come to the result as follows:

Theorem 1.3.1 *For every collection* $d, s, L, \sigma, n = m^d$ *satisfying* (1.48) *and every* $q \in [1, \infty)$ *the minimax q-risk of estimating functions from* $\mathcal{H}_d^s(L)$ *via observations* (1.33) *can be bounded from below as*

$$\mathcal{R}_q^*(n; \mathcal{H}_d^s(L)) \ge O_{s,d}(1) L \left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2s}{2s+d}}. \tag{1.54}$$

For every fixed $\kappa > 0$, *for every collection* $s, d, L, \sigma, n = m^d$ *satisfying the assumption*

$$n^\kappa \le \left(\frac{L\sqrt{n}}{\sigma\sqrt{\ln n}}\right)^{\frac{2d}{2s+d}} \le (k+2)^{-d} n \tag{1.55}$$

(cf. (1.52)*), the minimax* ∞*-risk of estimating functions from* $\mathcal{H}_d^s(L)$ *via observations* (1.33) *can be bounded from below as*

$$\mathcal{R}_\infty^*(n; \mathcal{H}_d^s(L)) \ge O_{\kappa,s,d}(1) L \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2s}{2s+d}}. \tag{1.56}$$

Consequently, in the case of (1.48) *the estimation method* $\{\widehat{f}_n\}_n$ *given by* (1.41), (1.47) *is minimax optimal on* $\mathcal{H}_d^s(L)$ *with respect to q-risk,* $1 \le q < \infty$, *up to a factor depending on* s, d, q *only. Similarly, in the case of* (1.55) *the estimate* $\widehat{f}_n^\infty$ *is minimax optimal on* $\mathcal{H}_d^s(L)$ *with respect to the* ∞*-risk up to a factor depending on* s, d, κ *only.*

As a corollary, we get the following expressions for the minimax risks of estimating functions from Hölder balls $\mathcal{H}_d^s(L)$ via observations (1.33):

For all large enough values of n (cf. (1.48), (1.55)), *one has*

$$\boxed{\begin{array}{rcl} \mathcal{R}_q^*(n; \mathcal{H}_d^s(L)) & = & O_{s,d,q}\left(L\left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2s}{2s+d}}\right), \\ & & 1 \le q < \infty; \\ \mathcal{R}_\infty^*(n; \mathcal{H}_d^s(L)) & = & O_{s,d,\kappa}\left(L\left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2s}{2s+d}}\right) \end{array}} \tag{1.57}$$

(From now on, we write $f(n) = O_\theta(g(n))$, θ being a collection of parameters, if both $f(n)/g(n)$ and $g(n)/f(n)$ admit upper bounds depending on Θ only.)

Note that the estimates underlying the upper bounds in (1.57) can be chosen to be linear in observations.

1.4 Appendix: proof of the Fano inequality

The proof of the Fano inequality is given by the following two lemmas.

Lemma 1.4.1 *Let* $\{\pi(i,j)\}_{i,j=1}^{N}$ *be a probability distribution on* $[\overline{1,N}]^2$, *let*

$$\begin{array}{rcl} p(i) & = & \sum\limits_{j=1}^{N} \pi(i,j),\ i=1,...,N \\ q(j) & = & \sum\limits_{i=1}^{N} \pi(i,j),\ j=1,...,N \end{array}$$

be the associated marginal distributions, and let

$$\theta = \sum_{i=1}^{N} \pi(i,i).$$

Then

$$\begin{array}{rcl} I[\pi] & \equiv & \sum\limits_{i,j} \pi(i,j) \ln\left(\frac{\pi(i,j)}{p(i)q(j)}\right) \\ & \geq & -\sum\limits_{i} p(i)\ln p(i) - (1-\theta)\ln(N-1) + [(1-\theta)\ln(1-\theta) + \theta\ln\theta] \\ & \geq & -\sum\limits_{i} p(i)\ln p(i) - (1-\theta)\ln(N-1) - \ln 2. \end{array} \tag{1.58}$$

Proof. We have

$$\begin{array}{rcl} I[\pi] & = & \sum\limits_{i,j} \left[-\pi(i,j)\ln p(i) + \pi(i,j)\ln\left(\frac{\pi(i,j)}{q(j)}\right)\right] \\ & = & -\sum\limits_{i} p(i)\ln p(i) + \sum\limits_{i,j} \pi(i,j)\ln\left(\frac{\pi(i,j)}{q(j)}\right) \\ & \geq & -\sum\limits_{i} p(i)\ln p(i) \\ & & + \min\limits_{\xi(\cdot,\cdot),\eta(\cdot)\in B} \sum\limits_{i,j} \xi(i,j)\ln\left(\dfrac{\xi(i,j)}{\eta(j)}\right), \\ & & B = \Big\{ \left(\{\xi(i,j)\geq 0\}_{i,j=1}^{N}, \{\eta(j)\geq 0\}_{j=1}^{N}\right) \mid \xi(i,i) = \pi(i,i), \\ & & \quad \sum\limits_{i} \xi(i,j) = \eta(j) \quad \forall j, \sum\limits_{j} \eta(j) = 1 \Big\}. \end{array} \tag{1.59}$$

The function $p\ln\frac{p}{q}$ [4] is convex and lower semicontinuous in $p,q\geq 0$, so that the function

$$f(\xi,\eta) = \sum_{i,j} \xi(i,j)\ln\left(\frac{\xi(i,j)}{\eta(j)}\right)$$

is convex on the convex set B. To compute its minimum on B, let us fix $\{\eta(j)\geq\pi(j,j)\}_{j=1}^{N}$ with $\sum\limits_{j}\eta(j)=1$ and minimize $f(\xi,\eta)$ over those ξ for which $(\xi,\eta)\in B$. Due to the separable structure of f, this minimization results in

$$\min_{\xi:(\xi,\eta)\in B} f(\xi,\eta) = \sum_{j} \min_{\{\xi(i)\geq 0\}_{i=1}^{N}:\xi(j)=\pi(j,j),\sum\limits_{i}\xi(i)=\eta(j)} \sum_{i}\xi(i)\ln\left\{\frac{\xi(i)}{\eta(j)}\right\}.$$

[4] By definition, $0\ln\frac{0}{q}=0$ for all $q\geq 0$ and $p\ln\frac{p}{0}=+\infty$ whenever $p>0$

For every j, a solution to the problem

$$\min_{\{\xi(i)\geq 0\}_{i=1}^N : \xi(j)=\pi(j,j), \sum\limits_i \xi(i)=\eta(j)} \sum_i \xi(i) \ln\left\{\frac{\xi(i)}{\eta(j)}\right\}$$

is given by

$$\xi(i) = \begin{cases} \pi(j,j), & i=j \\ \frac{\eta(j)-\pi(j,j)}{N-1}, & i \neq j \end{cases},$$ [5]

so that

$$g(\eta) \equiv \min_{\xi:(\xi,\eta)\in B} f(\xi,\eta) = \sum_j \left[[\eta(j)-\pi(j,j)] \ln\left(\frac{\eta(j)-\pi(j,j)}{(N-1)\eta(j)}\right) + \pi(j,j)\ln\left(\frac{\pi(j,j)}{\eta(j)}\right)\right].$$

It remains to minimize $g(\eta)$ over

$$\eta \in B' = \left\{ \{\eta(j)\}_{j=1}^N \mid \eta(j) \geq \pi(j,j), \sum_j \eta(j) = 1 \right\}.$$

We claim that the required minimizer η_* is given by

$$\eta_*(j) = \frac{1}{\theta}\pi(j,j),\ j=1,...,N.$$

Indeed, g is convex on the convex set B', so that in order to verify that the above η_* (which clearly belongs to B') minimizes g on B', it suffices to verify that the derivative of g at η_* is proportional to the vector of ones (i.e., to the normal to the hyperplane $\sum\limits_j \eta(j) = 1$ containing B'). We have

$$\begin{array}{rcl} \frac{\partial}{\partial \eta(j)} g(\eta_*) & = & \ln\left(\frac{\eta_*(j)-\pi(j,j)}{(N-1)\eta_*(j)}\right) + 1 - \frac{\eta_*(j)-\pi(j,j)}{\eta_*(j)} - \frac{\pi(j,j)}{\eta_*(j)} \\ & = & \frac{1-\theta}{N-1}, \end{array}$$

as required.

We conclude that

$$\begin{array}{rcl} \min\limits_{(\xi,\eta)\in B} f(\xi,\eta) & = & g(\eta_*) \\ & = & \sum\limits_j \pi(j,j)\left[\left(\frac{1}{\theta}-1\right)\ln\left(\frac{1-\theta}{N-1}\right) + \ln\theta\right] \\ & = & (1-\theta)\ln(1-\theta) - (1-\theta)\ln(N-1) + \theta\ln\theta, \end{array}$$

and (1.58) follows. ■

Now let us set $H_i = F_i^m$, so that H_i is a probability distribution on $(\Omega, \mathcal{F})^m$, and let Ω_j be the set of those $\omega^m \in \Omega^m$ at which $\mathcal{D}(\cdot)$ is equal to j, so that $\{\Omega_j\}_{j=1}^N$ is a partition of Ω into N non-overlapping Borel sets. Given a probability distribution $\{p(i)\}_{i=1}^N$ on $[\overline{1,N}]$, let us set

$$\begin{array}{rcl} \kappa(i,j) & = & \int\limits_{\Omega_j} dH_i(\omega^m), \\ \pi(i,j) & = & p(i)\kappa(i,j), \end{array}$$

[5] Indeed, the function we are minimizing is lower semicontinuous, and it is minimized on a compact set, so that the minimum is attained. Since the set is convex and the function is convex and symmetric in $\{\xi(i)\}_{i\neq j}$, it has a minimizer where all $\xi(i)$ with $i \neq j$ are equal to each other.

so that $\pi(\cdot,\cdot)$ is a probability distribution on $[\overline{1,N}]^2$. Note that by evident reasons

$$\mathcal{K}(H_i : H_j) = m\mathcal{K}(F_i : F_j),$$

so that

$$\mathcal{K}(H_i : H_j) \le m\mathcal{K}. \tag{1.60}$$

Lemma 1.4.2 *One has*

$$I[\pi] \le \mathcal{K}. \tag{1.61}$$

Proof. Denoting $H = \sum\limits_j H_j$ and $h_j(\omega^m) = \frac{dH_j(\omega^m)}{dH(\omega^m)}$, we get

$$\begin{array}{rcl}
\mathcal{K}(H_i : H_j) & = & \sum\limits_k \int\limits_{\Omega_k} h_j(\omega^m) \ln\left(\frac{h_j(\omega^m)}{h_i(\omega^m)}\right) dH(\omega^m) \\
& = & -\sum\limits_k \int\limits_{\Omega_k} h_j(\omega^m) \ln\left(\frac{h_i(\omega^m)}{h_j(\omega^m)}\right) dH(\omega^m) \\
& \ge & -\sum\limits_k \kappa(j,k) \ln\left(\int\limits_{\Omega_k} \frac{h_j(\omega^m)}{\kappa(j,k)} \frac{h_i(\omega^m)}{h_j(\omega^m)} dH(\omega^m)\right) \\
& & [\text{Jensen's inequality for the concave function } \ln(\cdot)] \\
& = & \sum\limits_k \kappa(j,k) \ln\left(\frac{\kappa(j,k)}{\kappa(i,k)}\right).
\end{array} \tag{1.62}$$

Thus, in view of (1.60)

$$m\mathcal{K} \ge \mathcal{K}(H_i : H_j) \ge \sum_k \kappa(j,k) \ln\left(\frac{\kappa(j,k)}{\kappa(i,k)}\right) \quad \forall i,j. \tag{1.63}$$

We now have

$$\begin{array}{rcl}
I[\pi] & = & \sum\limits_{j,k} p(j)\kappa(j,k) \ln\left(\frac{p(j)\kappa(j,k)}{\left(\sum\limits_i p(j)\kappa(j,i)\right)\left(\sum\limits_i p(i)\kappa(i,k)\right)}\right) \\
& = & \sum\limits_{j,k} p(j)\kappa(j,k) \ln\left(\frac{\kappa(j,k)}{\sum\limits_i p(i)\kappa(i,k)}\right) \\
& \le & \sum\limits_{i,j,k} p(i)p(j)\kappa(j,k) \ln\left(\frac{\kappa(j,k)}{\kappa(i,k)}\right) \\
& & [\text{Jensen's inequality for the convex function } f(t) = \ln\frac{a}{t}] \\
& \le & \sum\limits_{i,j} p(i)p(j)m\mathcal{K} \\
& & [\text{by (1.63)}] \\
& = & m\mathcal{K}.
\end{array} \tag{1.64}$$

Combining (1.64) and (1.58), we come to (1.26); setting in the latter inequality $p(i) = \frac{1}{N}$, $i = 1, ..., N$, we get (1.27). ■

Remark 1.4.1 In course of proving the Fano inequality (see (1.62), we have obtained a result which is important by itself:

> *Let F, G be two mutually absolutely continuous probability distributions on Ω, and let*
>
> $$\Omega = \bigcup_{i=1}^{I} \Omega_i$$

be a partitioning of Ω into $I < \infty$ mutually disjoint sets from the underlying σ-algebra. Let $\widehat{F}$, $\widehat{G}$ be the distributions of the "point index"

$$i(\omega) = \begin{cases} 1, & \omega \in \Omega_1 \\ 2, & \omega \in \Omega_2, \\ ... & \\ I, & \omega \in \Omega_I \end{cases}$$

induced by F, G, respectively. Then

$$\mathcal{K}(\widehat{F} : \widehat{G}) \le \mathcal{K}(F : G).$$

Chapter 2

Estimating regression functions from Sobolev balls

We have seen what are the possibilities to recover a "uniformly smooth" regression function – one from a given Hölder ball. What happens if the function f in question is smooth in a "non-uniform" sense, i.e., bounds on somehow averaged magnitudes of the derivatives of f are imposed? In this case, the function is allowed to have "nearly singularities" – it may vary rapidly in small neighbourhoods of certain points. The most convenient form of a "non-uniform" smoothness assumption is that the observed function $f : [0,1]^d \to \mathbf{R}$ belongs to a given *Sobolev ball* $\mathcal{S}_d^{k,p}(L)$.

A Sobolev ball $\mathcal{S}_d^{k,p}(L)$ is given by four parameters:

- positive integer k – order of smoothness,
- positive integer d – dimensionality,
- $p \in (d, \infty]$,
- $L > 0$,

and is comprised of all continuous functions $f : [0,1]^d \to \mathbf{R}$ such that the partial derivatives of order k of f (understood in the sense of distributions) form a usual vector-function $D^k f(\cdot)$ with

$$\| D^k f(\cdot) \|_p \le L.$$

It is known [2] that functions $f \in \mathcal{S}_d^{k,p}(L)$ are $(k-1)$ times continuously differentiable (this is ensured by the restriction $p > d$), and we denote by $D^s f(\cdot)$ the vector-function comprised of partial derivatives of order $s < k$ of a function $f \in \mathcal{S}_d^{k,p}(L)$.

Note that Hölder balls $\mathcal{H}_d^s(L)$ with *integer* s are essentially the same as Sobolev balls $\mathcal{S}_d^{s,\infty}(L)$.

The problem we are interested in is as follows. Given $n = m^d$ observations

$$\begin{gathered} y \equiv y_f(\xi) = \left\{ y_\iota = f(x_\iota) + \sigma\xi_\iota | \iota = (i_1, ..., i_d) \in \overline{[1,m]}^d \right\} \\ \left[\begin{array}{c} x_{(i_1,...,i_d)} = (i_1/m, i_2/m, ..., i_d/m)^T, \\ \xi = \{\xi_\iota\} : \ \xi_\iota \text{ are independent } \mathcal{N}(0,1) \end{array} \right] \end{gathered} \tag{2.1}$$

of an unknown regression function $f : [0,1]^d \to \mathbf{R}$ belonging to a given Sobolev ball $\mathcal{S} = \mathcal{S}_d^{k,p}(L)$, we want to recover f along with its partial derivatives

$$D^{(\alpha)} f = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \ldots \partial x_d^{\alpha_d}} f$$

of orders $|\alpha| \equiv \sum\limits_{i=1}^{d} \alpha_i \le k-1$.

Notation and conventions. In what follows, for an estimate $\hat{f}^{n,(\alpha)}(x;y)$ of $D^{(\alpha)} f$ via observations (2.1), we denote by

$$\mathcal{R}_{q,(\alpha)}(\hat{f}^{n,(\alpha)};\mathcal{S}) = \sup_{f\in\mathcal{S}} \left(\mathcal{E}\left\{ \| \hat{f}^{n,(\alpha)}(\cdot;y) - D^{(\alpha)} f(\cdot) \|_q^2 \right\} \right)^{1/2}$$

the q-risk of the estimate on the Sobolev ball in question; here $1 \le q \le \infty$. The associated minimax risk is defined as

$$\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}) = \inf_{\hat{f}^{n,(\alpha)}} \mathcal{R}_{q,(\alpha)}(\hat{f}^{n,(\alpha)};\mathcal{S})$$

Below we deal a lot with the parameters $p, q \in [1,\infty]$ (coming from the description of the Sobolev ball and the risk we are interested in, respectively); let us make the convention to denote

$$\pi = \frac{1}{p},\ \theta = \frac{1}{q}.$$

We call a *cube* a subset $B \subset [0,1]^d$ of the form $\{x \mid [0 \le]\ a_i \le x_i \le a_i + h\ [\le 1],\ i \in \overline{[1,d]}\}$ and denote by $D(B) = h$ the edge length, and by $|B| = h^d$ the d-dimensional volume of such a cube B.

For a collection $k, d \in \mathbf{N}; p \in (d,\infty]; q \in [1,\infty]$ and $l \in \overline{[0,k-1]}$ let

$$\beta_l(p,k,d,q) = \begin{cases} \frac{k-l}{2k+d}, & \theta \ge \pi\frac{2l+d}{2k+d} \\ \frac{k-l+d\theta-d\pi}{2k-2d\pi+d}, & \theta \le \pi\frac{2l+d}{2k+d} \end{cases}; \tag{2.2}$$

when the parameters p, k, d, q are clear from the context, we shorten $\beta_l(p,k,d,q)$ to β_l.

We denote by $\mathcal{A}$ the set of the admissible for us values of the parameters p, k, d, i.e.,

$$\mathcal{A} = \{(p,k,d) \mid k, d \in \mathbf{N}, p \in (d,\infty]\}.$$

In what follows we denote by C (perhaps with sub- or superscripts) positive quantities depending on k, d only, and by P (perhaps with sub- or superscripts) – quantities ≥ 1 depending solely on $(p,k,d) \in \mathcal{A}$ and nonincreasing in p.

Finally, $|\cdot|$ stands both for the absolute value of a real and the Euclidean norm of a vector.

2.1 Lower bounds for the minimax risk

The lower bounds for the minimax risk are given by the following

Theorem 2.1.1 *Let $\sigma, L > 0$, $(p,k,d) \in \mathcal{A}$, $q \in [1,\infty]$, $l \in \overline{[0,k-1]}$ and (α), $|\alpha| = l$, be given. Assume that the volume of observations n is large enough, namely,*

$$1 \leq \frac{L\sqrt{n}}{\sigma}. \tag{2.3}$$

Then the minimax q-risk of estimating $D^{(\alpha)}f$ for functions f from the Sobolev ball $\mathcal{S}_d^{k,p}(L)$ via observations (2.1) can be bounded from below as

$$\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}) \geq O_{k,d}(1)L\left(\frac{\sigma}{L\sqrt{n}}\right)^{2\beta_l(p,k,d,q)}. \tag{2.4}$$

If the volume of observations n is so large that

$$n^{\varepsilon} \leq \frac{L\sqrt{n}}{\sigma} \tag{2.5}$$

for some positive ε, then in the case of "large" ratios q/p, namely, $\frac{q}{p} \geq \frac{2k+d}{2l+d}$, the lower bound can be strengthened to

$$\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}) \geq O_{k,d,\varepsilon}(1)L\left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{2\beta_l(p,k,d,q)}. \tag{2.6}$$

The proof (completely similar to the proof of the lower bounds from Section 1.2) is placed in Section 2.3.

Comments, I. The lower bounds for the minimax risk (2.4), (2.6) (which, as we shall see in a while, are sharp in order) demonstrate the following behaviour of the minimax risk as a function of the volume of observations n:

1. For given k, d and $l = |\alpha| < k$ there exists the "standard" asymptotics of the risk $\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S})$ which is
$$O(n^{-(k-l)/(2k+d)});$$
this is the behaviour of the risk for "small" ratios q/p, namely, when
$$q/p = \pi/\theta < \frac{2k+d}{2l+d}.$$
Note that the standard asymptotics is independent of p, q – i.e., of the particular norms in which we measure the magnitude of $D^k f$ and the estimation error. Note also that in the case of $l = |\alpha| = 0$, i.e., when speaking about recovering the regression function itself rather than its derivatives, the standard asymptotics of risk is $O(n^{-k/(2k+d)})$ – the result already known to us in the particular case of $p = \infty$, $q < \infty$, i.e., when the Sobolev ball in question is in fact the Hölder ball $\mathcal{H}_d^k(L)$.

2. When the ratio q/p is greater than or equal to the "critical level" $\frac{2k+d}{2l+d}$, the asymptotics of the minimax risk becomes
$$O\left(\left(\frac{\ln n}{n}\right)^{\frac{k-l+d\theta-d\pi}{2k-2d\pi+d}}\right)$$

and starts to depend on p, q. As q grows, p being fixed, it becomes worse and worse, and the worst asymptotics corresponds to $q = \infty$ and is

$$O\left(\left(\frac{\ln n}{n}\right)^{\frac{k-l-d\pi}{2k-2d\pi+d}}\right).$$

Comments, II. We have seen that when recovering "uniformly smooth" regression functions – those from Hölder balls – an optimal in order estimate can be chosen to be linear in observations. In the case of "non-uniform" smoothness linear estimates work well in a restricted range of values of q only – essentially, when $q \le p < \infty$. The exact claims are as follows:

(i) The lower bounds from Theorem 2.1.1 in the case of $q \le p < \infty$ can be achieved (up to independent of n factors) by properly chosen linear estimates;

(ii) If $\infty > q > p$ and $q \ge 2$, no linear estimation method can achieve the rates of convergence indicated in Theorem 2.1.1.

We shall check (i) in the case when our target is to recover the regression function, and not its derivatives; namely, we shall demonstrate that the order $k-1$ window estimate $\widehat{f}_n$ (see Section 1.3) recovers functions $f \in \mathcal{S} = \mathcal{S}_d^{k,p}(L)$ with the desired order of convergence of q-risk to 0 as $n \to \infty$ (provided that $q \le p$) [1]. Recall that $\widehat{f}_n$ uses windows of the same width h to recover $f(x)$ at all points x. Let us specify this width as (cf. (1.47))

$$h = \left(\frac{\sigma}{L\sqrt{n}}\right)^{2/(2k+d)} \tag{2.7}$$

and assume that n is large, namely, that

$$1 \le \left(\frac{L\sqrt{n}}{\sigma}\right)^{\frac{2d}{2k+d}} \le (k+2)^{-d} n \tag{2.8}$$

(cf. (1.48)). Under this assumption

$$h \ge \frac{k+2}{n^{1/d}}, \tag{2.9}$$

so that our estimate $\widehat{f}_n$ is well-defined.

To bound the risk of the resulting estimate, we need the following fact from Analysis (see [2]):

Lemma 2.1.1 *Let $B \subset [0,1]^d$ be a cube, let $p \in (d, \infty]$, and let $g \in \mathcal{S}_p^{1,d}(\cdot)$. Then the function g is Hölder continuous in B with Hölder exponent $1 - d\pi \equiv 1 - d/p$; namely,*

$$\forall x, x' \in B: \qquad |g(x) - g(x')| \le O_{p,d}(1)|x - x'|^{1-d\pi} \left(\int\limits_B |Dg(u)du|^p\right)^{1/p}. \tag{2.10}$$

with $O_{p,d}(1)$ nonincreasing in $p > d$.

[1] One can easily build optimal in order, in the case of $q \le p$, window estimates of the derivatives as well.

An immediate consequence of Lemma 2.1.1 is the following useful relation:

$$f \in \mathcal{S}_d^{k,p}(L) \Rightarrow \Phi_{k-1}(f,B) \le O_{p,k,d}(1) D^{k-d\pi}(B) \left(\int\limits_B |D^k f(u)|^p du \right)^{1/p} \tag{2.11}$$

($B \in [0,1]^d$ is a cube, $D(B)$ is the edge length of B) with $O_{p,d}(1)$ nonincreasing in $p > d$; here $\Phi_{k-1}(f,B)$ is the quality of the best uniform, on B, approximation of f by a polynomial of degree $\le k-1$, see Proposition 1.3.1. The right hand side in the inequality in (2.11) is nothing but an upper bound (given by (2.10) as applied to $g = D^{k-1}f$) on the error of approximating f in B by its Taylor polynomial of the degree $k-1$, the polynomial being taken at a point from B.

Now we are ready to evaluate the q-risks, $q \le p$, of the window estimate $\widehat{f}_n$ on a Sobolev ball $\mathcal{S} = \mathcal{S}_d^{k,p}(L)$. Let us start with the case of $q = p$. Assuming $f \in \mathcal{S}$ and combining the bound (1.42) from Proposition 1.3.1 and (2.11), we get

$$\begin{array}{rcl} |\widehat{f}_n(x; y_f(\xi)) - f(x)| & \le & d(x) + s(x,\xi), \\ d(x) & = & O_{k,p,d}(1) D^{k-d\pi}(B(x)) \left(\int\limits_{B(x)} |D^k f(u)|^p du \right)^{1/p}, \\ s(x,\xi) & = & O_{k,d}(1) \frac{\sigma}{\sqrt{nh^d}} |\zeta_x^{B(x)}|; \end{array} \tag{2.12}$$

here $B(x)$ is the window used by $\widehat{f}_n$ to recover $f(x)$.

Now, the function $d(x)$ is non-random, and its p-norm can be evaluated as follows. Let us extend the function $\ell(u) = |D^k f(u)|$ from $[0,1]^d$ to the entire $\mathbf{R}^d$ as 0 outside the unit cube. Then

$$\begin{array}{rcl} \int d^p(x) dx & = & O^p_{k,p,d}(1) h^{kp-d} \int\limits_{[0,1]^d} \left[\int\limits_{B(x)} \ell^p(u) du \right] dx \\ & \le & O^p_{k,p,d}(1) h^{kp-d} \int \left[\int\limits_{-h}^{h} \dots \int\limits_{-h}^{h} \ell^p(x-u) du \right] dx \\ & & [\text{since } B(x) \subset \{u \mid x_i - h \le u_i \le x_i + h, i \le d\}] \\ & = & O^p_{k,p,d}(1) h^{kp-d} \int\limits_{-h}^{h} \dots \int\limits_{-h}^{h} \left[\int \ell^p(x-u) dx \right] du \\ & = & O^p_{k,p,d}(1) h^{kp-d} (2h)^d \int \ell^p(x) dx \\ & \le & O^p_{k,p,d}(1) h^{kp} L^p. \end{array}$$

Thus,

$$\| d(x) \|_p \le O_{k,p,d}(1) h^k L. \tag{2.13}$$

Furthermore, $\zeta_x^{B(x)}$ is $\mathcal{N}(0,1)$-random variable, so that

$$\left(\mathcal{E}_\xi \left\{ \| s(\cdot,\xi) \|_p^2 \right\} \right)^{1/2} \le O_{k,p,d}(1) \frac{\sigma}{\sqrt{nh^d}}. \tag{2.14}$$

Relations (2.12), (2.13), (2.14) imply that

$$\mathcal{R}_p(\widehat{f}_n; \mathcal{S}) \le O_{k,p,d}(1) \left[h^k L + \frac{\sigma}{\sqrt{nh^d}} \right];$$

substituting the expression for h from (2.7), we see that the risk bound

$$\mathcal{R}_q(\widehat{f}_n;\mathcal{S}) \le O_{k,p,d}(1) L \left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2k}{2k+d}} \tag{2.15}$$

is valid in the case of $q = p$. Since the left hand side in (2.15) clearly is nondecreasing in q, it follows that the bound is valid for $1 \le q \le p$ as well. It remains to note that the right hand side of our upper bound is, up to a factor depending on k, d, p only, the same as the lower bound on the minimax risk (2.4) (look what is β_l in the case of $l = 0, q/p \le 1$).

Now let us verify our second claim – that in the case of $\infty > q > p$, $q \ge 2$, the q-risk of a linear estimate on a Sobolev ball $\mathcal{S}_d^{k,p}(L)$ never is optimal in order. Let Lin_n be the set of all *linear* in observations (2.1) estimates

$$\hat{f}^{n,(\alpha)}(x;y) = \sum_\iota \phi_{\iota,n}(x) y_\iota$$

of $D^{(\alpha)} f(\cdot)$, and let

$$\mathcal{R}^{\mathrm{Lin}}_{q,(\alpha)}(n;\mathcal{S}) = \inf_{\hat{f}^{n,(\alpha)} \in \mathrm{Lin}} \sup_{f \in \mathcal{S}_d^{k,p}(L)} \left(\mathcal{E}\left\{ \| D^{(\alpha)} f(\cdot) - \hat{f}^{n,(\alpha)}(\cdot;y) \|_q^2 \right\} \right)^{1/2}$$

be the associated minimax risk.

Theorem 2.1.2 *Let us fix $\sigma > 0$, $(p,k,d) \in \mathcal{A}$, $l \in \overline{[0,k-1]}$ and (α), $|\alpha| = l$. For every $q \in [2,\infty)$ such that $q > p$ and for all large enough volumes n of observations one has*

$$\begin{array}{rcl} \mathcal{R}^{\mathrm{Lin}}_{q,(\alpha)}(n;\mathcal{S}) & \ge & O_{p,k,d,q}(1) L \left(\frac{\sigma}{L\sqrt{n}}\right)^{2\mu_l}, \\ \mu_l \equiv \mu_l(p,k,d,q) & = & \frac{k-l-d\pi+d\theta}{2k-2\pi d+2d\theta+d} < \beta_l(p,k,d,q). \end{array} \tag{2.16}$$

The proof is placed in Section 2.3.

As we just have mentioned, the lower bounds on the minimax risk $\mathcal{R}^*_{q,(\alpha)}(n,\mathcal{S})$ given in Theorem 2.1.1 are sharp in order, so that (2.16) implies that

$$\infty > q \ge 2, q > p \Rightarrow \frac{\mathcal{R}^{\mathrm{Lin}}_{q,(\alpha)}(n;\mathcal{S})}{\mathcal{R}^*_{q,(\alpha)}(n,\mathcal{S})} \to \infty \text{ as } n \to \infty;$$

thus, for “large” q/p linear estimators cannot be optimal in order on $\mathcal{S}_d^{k,p}(L)$, independently of whether we are interested in recovering the regression function or its derivatives. Note also that the lower bound (2.16) is valid for an arbitrary n-point observation grid, not necessary the equidistant one.

2.2 Upper bounds on the minimax risk

In order to bound the minimax risk from above, we are about to build a particular recovering routine and to investigate its risks. In what follows, Γ_n is the equidistant observation grid from (2.1).

The recovering routine is as follows. Let $\mathcal{B}_n$ be the system of all distinct cubes with vertices from Γ_n, and let $n(B)$ be the number of observation points in a cube B. Let us associate with every cube $B \in \mathcal{B}_n$ the linear functional (the "B-average")

$$\phi_B(g) = n^{-1/2}(B) \sum_{\iota : x_\iota \in B} g(\iota)$$

on the space of real-valued functions defined on the grid Γ_n.

Let us call a system $\mathcal{B} \subset \mathcal{B}_n$ *normal*, if it meets the following requirement:

> (*) *For every cube $B \subset [0,1]^d$ such that $|B| > 6^d n^{-1}$, there exists a cube $B' \in \mathcal{B}$ such that*
>
> $$B' \subset B \text{ and } |B'| \geq 6^{-d}|B|.$$

Note that normal systems clearly exist (e.g., $\mathcal{B} = \mathcal{B}_n$; in fact one can build a normal system with $O(n)$ cubes).

Given observations (2.1) (which together form a function on Γ_n), we may compute all the averages $\phi_B(y)$, $B \in \mathcal{B}$. Consider the following optimization problem:

$$\Phi_{\mathcal{B}}(g, y) \equiv \max_{B \in \mathcal{B}} |\phi_B(g) - \phi_B(y)| \to \min \mid g \in \mathcal{S}_d^{k,p}(L). \tag{2.17}$$

It can be easily verified that the problem is solvable and that its optimal solution can be chosen to be a Borel function $\hat{f}^n(x; y)$ of x, y. $\hat{f}^n(\cdot; y)$ is exactly the estimate of f we are interested in, and we estimate the derivative $D^{(\alpha)} f(\cdot)$ of f just by the corresponding derivative

$$\hat{f}^{n,(\alpha)}(x; y) = \frac{\partial^l}{\partial_{x_1}^{\alpha_1} \ldots \partial_{x_d}^{\alpha_d}} \hat{f}^n(x; y) \quad [l \equiv |\alpha| \leq k-1]$$

of $\hat{f}(\cdot; y)$.

Risks of the estimates $\hat{f}^{n,(\alpha)}(\cdot; \cdot)$ on the Sobolev ball $\mathcal{S}_d^{k,p}(L)$ are given by the following

Theorem 2.2.1 *For every $\sigma > 0$, $(p, k, d) \in \mathcal{A}$, $q \in [1, \infty]$, $l \in \overline{[0, k-1]}$ and (α), $|\alpha| = l$, for all large enough volumes n of observations, namely, such that $n > P$ and*

$$1 \leq \frac{L\sqrt{n}}{\sigma\sqrt{\ln n}} \leq n^{\frac{2k - 2d\pi + d}{2d}} \tag{2.18}$$

one has

$$\begin{array}{c} \mathcal{R}_{q,(\alpha)}(\hat{f}^{n,(\alpha)}; \mathcal{S}_d^{k,p}(L)) \leq PL \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{2\beta_l}, \\ \beta_l \equiv \beta_l(p, k, d, q); \end{array} \tag{2.19}$$

here $P \geq 1$ depends on k, p, d only and is nonincreasing in $p > d$.

In particular (cf. Theorem 2.1.1), in the case of "large" ratios q/p:

$$q/p \geq \frac{2k + d}{2l + d}$$

the estimate $\hat{f}^{n,(\alpha)}$ is asymptotically optimal in order in the minimax sense:

$$\frac{\mathcal{R}_{q,(\alpha)}(\hat{f}^{n,(\alpha)};\mathcal{S}_d^{k,p}(L))}{\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}_d^{k,p}(L))} \leq P$$

for all large enough values of n.

In the case of "small" ratios q/p:

$$q/p < \frac{2k+d}{2l+d}$$

the estimate is optimal in order up to a logarithmic in n factor: for all large enough values of n,

$$\frac{\mathcal{R}_{q,(\alpha)}(\hat{f}^{n,(\alpha)};\mathcal{S}_d^{k,p}(L))}{\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}_d^{k,p}(L))} \leq P(\ln n)^{\beta_l}.$$

Proof. In what follows we fix p, k, d, α, l, L, q satisfying the premise of the theorem. We write $\mathcal{S}$ instead of $\mathcal{S}_d^{k,p}(L)$ and denote

$$\mathbf{S} = \mathbf{S}_d^{k,p} = \bigcup_{L>0} \mathcal{S}_d^{k,p}(L).$$

Let us set

$$\| g \|_{\mathcal{B}} = \max_{B\in\mathcal{B}} |\phi_B(g)|,$$

and let

$$\Theta(\xi) = \| \sigma\xi \|_{\mathcal{B}} \equiv \sigma \max_{B\in\mathcal{B}} \frac{1}{\sqrt{n(B)}} \sum_{\iota:x_\iota\in B} \xi_\iota. \tag{2.20}$$

Our central auxiliary result is as follows:

Lemma 2.2.1 *There exists $P \geq 1$ depending on p, k, d only and nonincreasing in $p > d$ such that whenever $n \geq P$ one has*

$$\begin{array}{l} \forall f \in \mathcal{S} : \\ \| \hat{f}^n(\cdot; y_f(\xi)) - f \|_{\mathcal{B}} \leq 2\Theta(\xi) \end{array} \tag{2.21}$$

and

$$\begin{array}{l} \forall g \in \mathbf{S},\ \forall(\alpha), l \equiv |\alpha| < k : \\ \| D^{(\alpha)} g \|_q \leq P_0 \max\left\{ \left(\frac{\|g\|^2_{\mathcal{B}}}{\|D^k g\|^2_p n}\right)^{\beta} \| D^k g \|_p; \left(\frac{\|g\|^2_{\mathcal{B}}}{n}\right)^{1/2}; n^{-\lambda} \| D^k g \|_p \right\}, \\ \textit{where} \\ \beta = \beta_l(p,k,d,q), \quad \lambda = \lambda_l(p,k,d,q) = \frac{2k+d-2d\pi}{d}\beta_l(p,k,d,q). \end{array} \tag{2.22}$$

From Lemma 2.2.1 to Theorem 2.2.1. Assuming $f \in \mathcal{S}$, denoting

$$g(x) = f(x) - \hat{f}^n(x; y_f(\xi))$$

and taking into account (2.21), we get

$$\| g \|_{\mathcal{B}} \leq 2\Theta(\xi),$$

and by construction $\hat{f}^n(\cdot, y_f(\xi)) \in \mathcal{S}$, so that

$$\| D^k g(\cdot) \|_p \leq 2L.$$

In view of these observations, (2.22) says that

$$\| D^{(\alpha)} g \|_q^2 \leq P_1 \max\left\{ L^2 \left(\frac{\Theta^2(\xi)}{L^2 n}\right)^{2\beta_l}; \left(\frac{\Theta^2(\xi)}{n}\right); L n^{-2\lambda_l} \right\}. \tag{2.23}$$

Since Θ is the maximum of no more than $\mathrm{Card}(\mathcal{B}) \leq n^2$ $\mathcal{N}(0, \sigma^2)$ random variables (see (2.20)), we get

$$\left(\mathcal{E}\left\{\| D^{(\alpha)} g \|_q^2\right\}\right)^{1/2} \leq P_2 \max\left\{ L \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{2\beta_l}; \left(\frac{\sigma\sqrt{\ln n}}{\sqrt{n}}\right); L n^{-\lambda_l} \right\}. \tag{2.24}$$

It is immediately seen that in the case of (2.18) the maximum in the right hand side of this bound equals to $L\left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{2\beta_l}$, so that (2.24) is the required bound (2.19).

Proof of Lemma 2.2.1. 1^0. Relation (2.21) is evident: since f is a feasible solution to the optimization problem (2.17) and the value of the objective of the problem at this feasible solution is $\Theta(\xi)$, the optimal value of the problem does not exceed $\Theta(\xi)$; consequently, by the triangle inequality

$$\| f(\cdot) - \hat{f}^n(\cdot; y) \|_{\mathcal{B}} \leq \Phi_{\mathcal{B}}(f, y) + \Phi_{\mathcal{B}}(\hat{f}(\cdot; y), y) \leq 2\Theta(\xi) \qquad [y = y_f(\xi)],$$

as claimed in (2.17).

2^0. In order to prove (2.21), note first of all that a function $g \in \mathbf{S}_d^{k,p}$ can be approximated by a sequence of C^∞ functions g_t in the sense that

$$\| D^k g_t \|_p \to \| D^k g \|, \; \| g_t \|_{\mathcal{B}} \to \| g \|_{\mathcal{B}}, \; \| D^{(\alpha)} g_t \|_q \to \| D^{(\alpha)} g \|_q$$

as $t \to \infty$; consequently, it suffices to prove (2.22) for a C^∞ function g.

3^0. We shall use the following well-known fact (given by embedding theorems for Sobolev spaces, see [2]):

Lemma 2.2.2 *For properly chosen P_3, P_4 and for every $r, q \in [1, \infty]$, $l \in \overline{[0, k-1]}$, for every C^∞ function $g : [0,1]^d \to \mathbf{R}$ one has:*

- *either*

$$\| D^k g \|_p \leq P_3 \| g \|_1, \tag{2.25}$$

and then

$$\| D^l g \|_\infty \leq P_4 \| g \|_1, \tag{2.26}$$

- *or*

$$\begin{array}{rcl} \| D^l g \|_q & \leq & P_5 \| g \|_r^{\psi} \| D^k g \|_p^{1-\psi}, \\ \textit{where} & & \\ \psi & = & \begin{cases} \frac{k-l}{k}, & \theta \geq \frac{\pi l + (k-l)/r}{k}, \\ \frac{k-l-d\pi+d\theta}{k-d\pi+d/r}, & \theta \leq \frac{\pi l + (k-l)/r}{k} \end{cases} \end{array} \tag{2.27}$$

Recall that by Lemma 2.1.1 for smooth functions $g : [0,1]^d \to \mathbf{R}$ and for every cube $B \subset [0,1]^d$ one has

$$\begin{array}{rcl} \left|D^{k-1}g(x) - D^{k-1}g(y)\right| & \leq & P_4|x-y|^{1-d\pi}\Omega(g,B) \quad \forall x,y \in B, \\ \Omega(g,B) & = & \left(\int\limits_B |D^k g(u)|^p du\right)^{1/p}. \end{array} \tag{2.28}$$

4^0. From (2.28) it follows that whenever $B \subset [0,1]^d$, $x \in B$ and $g_s(y)$ is the Taylor polynomial, taken at x, of degree $k-1$ of g, then

$$\begin{array}{rcl} \max_{y\in B} |g(y) - g_x(y)| & \leq & P_5[D(B)]^{k+\delta-1}\Omega(g,B), \\ \delta & \equiv & 1 - d\pi. \end{array} \tag{2.29}$$

5^0. Let us call a cube $B \subset [0,1]^d$ *regular*, if

$$g(B) \equiv \max_{x\in B} |g(x)| \geq 4P_5[D(B)]^{k+\delta-1}\Omega(g,B). \tag{2.30}$$

Note that for a regular cube B, in view of (2.29), one has

$$\forall x \in B : \quad \max_{y\in B} |g(y) - g_x(y)| \leq \frac{1}{4}g(B). \tag{2.31}$$

It is clearly seen that if

$$U = \{x \in (0,1)^d \mid g(x) \neq 0\},$$

then every point $x \in U$ is an interior point of a regular cube B; among these cubes, there clearly exists a maximal one (i.e., a one which is not a proper subset of any other regular cube). For every $x \in U$, let us denote by B_x a maximal regular cube containing x as an interior point, and let

$$U' = \bigcup_{x\in U} B_x^0,$$

B_x^0 being the interior of the cube B_x. By the standard separability arguments,

$$U' = \bigcup_{i=1}^{\infty} B_{x_i}^0$$

for properly chosen sequence $x_1, x_2, ...$

In what follows we consider separately two cases

A. The cube $[0,1]^d$ is not regular;

B. The cube $[0,1]^d$ is regular.

6^0. For the time being, let **A** be the case. Since $[0,1]^d$ is not regular, every maximal regular cube B must satisfy (2.30) as an equality. In particular,

$$g(B_{x_i}) = 4P_5[D(B_{x_i}]^{k+\delta-1}\Omega(g,B_{x_i}),\ i = 1,2,... \tag{2.32}$$

6^0.a) We start with the following Lemma (which essentially originates from Banach):

Lemma 2.2.3 *One can extract from the system of cubes $\mathcal{A}_0 = \{B_{x_i}\}_{i=1}^{\infty}$ a sub-system $\mathcal{A}$ with the following properties:*

- *Cubes from $\mathcal{A}$ are mutually disjoint;*
- *For every cube $B \in \mathcal{A}_0$ there exists a cube $B' \in \mathcal{A}$ such that B intersects with B' and $D(B) \leq 2D(B')$.*

Proof of Lemma 2.2.3: Let us choose as the first cube of $\mathcal{A}$ a cube $B^1 \in \mathcal{A}_0$ with

$$D(B^1) \geq \frac{1}{2} \sup_{B\in\mathcal{A}_0} D(B).$$

After B^1 is chosen, we set $\mathcal{A}_1 = \{B \in \mathcal{A}_0 \mid B \cap B^1 = \emptyset\}$. If $\mathcal{A}_1$ is empty, we terminate; otherwise, we choose a cube B^2 from the collection $\mathcal{A}_1$ exactly in the same manner as B^1 was chosen from $\mathcal{A}_0$ and set $\mathcal{A}_2 = \{B \in \mathcal{A}_1 \mid B \cap B^2 = \emptyset\}$. If $\mathcal{A}_2$ is empty, we terminate, otherwise choose in the outlined fashion a cube $B^3 \in \mathcal{A}_2$ and replace $\mathcal{A}_2$ by $\mathcal{A}_3$, and so on.

As a result of this construction, we get a finite or a countable collection $\mathcal{A}$ of cubes B^1, B^2,...; it is immediately seen that this collection satisfies the requirements of Lemma. ∎

6^0.2) For $B \in \mathcal{A}$, let $U(B)$ be the union of all those cubes from $\mathcal{A}_0$ which intersect B and have edges not exceeding $2D(B)$. In view of Lemma 2.2.3, we have

$$U \subset U' \subset \bigcup_{B\in\mathcal{A}} U(B).$$

Let us choose

$$r \in [\frac{d+2k}{d}, \infty). \tag{2.33}$$

We have

$$\begin{array}{rcl} \| g \|_r^r & = & \int\limits_U |g(x)|^r dx \\ & \leq & \sum\limits_{B\in\mathcal{A}} \int\limits_{U(B)} |g(x)|^r dx \\ & \leq & 5^d \sum\limits_{B\in\mathcal{A}} |B| \hat{g}^r(B), \\ \hat{g}(B) & = & \sup_{x\in U(B)} |g(x)|. \end{array} \tag{2.34}$$

We claim that for every $B \in \mathcal{A}$ it holds

$$\hat{g}(B) \leq P_6 g(B). \tag{2.35}$$

Indeed, let $y \in U(B)$; then there exists $B' \in \mathcal{A}_0$ such that $y \in B'$, $B' \cap B \neq \emptyset$ and $D(B') \leq 2D(B)$. Choosing a point $x \in B \cap B'$ and applying (2.31) to the regular cubes B, B', we get

$$\max_{u\in D} |g(u) - g_x(u)| \leq \frac{1}{4} \max_{u\in D} |g(u)|$$

both for $D = B$ and $D = B'$. It follows that

$$\max_{u\in D} |g(u)| \leq \frac{4}{3} \max_{u\in D} |g_x(u)| \tag{2.36}$$

for both $D = B$ and $D = B'$. Since $g_x(\cdot)$ is a polynomial of degree $k-1$ and B' is contained in 5 times larger than the cube B concentric to B cube, we have

$$\max_{u\in B'} |g_x(u)| \leq P_7 \max_{u\in B} |g_x(u)|,$$

whence, in view of (2.36),

$$\max_{u\in B\cup B'} |g(u)| \leq \frac{4}{3} P_7 \max_{u\in B} |g_x(u)|.$$

Recalling that $\max_{u\in B} |g_x(u)| \leq \frac{5}{4} g(B)$ by (2.31), we come to

$$|g(y)| \leq \max_{u\in B'} |g_x(u)| \leq \frac{5}{3} P_7 g(B),$$

so that the choice $P_6 = \frac{5}{3}P_7$ ensures (2.35).

Combining (2.34) and (2.35), we get

$$\parallel g \parallel_r^r \leq P_8^r \sum_{B\in\mathcal{A}} |B| g^r(B). \tag{2.37}$$

6^0.c) Since $\mathcal{A} \subset \mathcal{A}_0$, (2.32) says that for every $B \in \mathcal{A}$ it holds

$$g(B) = 4P_5[D(B)]^{k+\delta-1}\Omega(g,B),$$

so that (2.37) yields the inequality

$$\parallel g \parallel_r^r \leq P_9^r \sum_{B\in\mathcal{A}} |B|^{1+\frac{r(k+\delta-1)}{d}} \Omega^r(B,g). \tag{2.38}$$

6^0.d) Let us set

$$A = \sup_{B\in\mathcal{A}} g(B)|B|^{1/2};$$

note that by (2.32) we have

$$A \geq P_{10} \sup_{B\in\mathcal{A}} |B|^{\frac{1}{2}+\frac{k+\delta-1}{d}} \Omega(B,g). \tag{2.39}$$

Let

$$\zeta = \frac{1 + \frac{r(k+\delta-1)}{d}}{\frac{1}{2} + \frac{k+\delta-1}{d}}.$$

Then

$$\begin{array}{rcl}
B & \in & \mathcal{A} \\
\Rightarrow \quad |B|^{1+\frac{r(k+\delta-1)}{d}} & \leq & P_{11}^{\zeta} A^{\zeta} \Omega^{-\zeta}(B,g) \\
 & & \text{[see (2.39)]} \\
\Rightarrow \quad \parallel g \parallel_r^r & \leq & P_9^r P_{11}^{\zeta} A^{\zeta} \sum\limits_{B\in\mathcal{A}} \Omega^{r-\zeta}(B,g) \\
 & & \text{[see (2.38)]} \\
 & \leq & P_9^r P_{11}^{\zeta} A^{\zeta} \left(\sum\limits_{B\in\mathcal{A}} \Omega^p(B,g) \right)^{\frac{r-\zeta}{p}} \\
 & & \text{[since } r - \zeta \geq p \text{ in view of (2.33)]} \\
 & \leq & P_9^r P_{11}^{\zeta} A^{\zeta} \parallel D^k g \parallel_p^{r-\zeta} \\
 & & \text{[since the cubes } B \in \mathcal{A} \text{ are mutually disjoint]} \\
\Rightarrow \quad \parallel g \parallel_r & \leq & P_{12} A^{\gamma} \parallel D^k g \parallel_p^{1-\gamma}, \\
\gamma & = & \frac{\zeta}{r}
\end{array}$$

The resulting estimate was established in the case of (2.33); passing to limit as $r \to \infty$, we see that it is valid for $r = \infty$ as well, so that

$$\infty \geq r \geq \frac{2k+d}{d}p \Rightarrow \parallel g \parallel_r \leq P_{12} A^{\gamma} \parallel D^k g \parallel^{1-\gamma}, \quad \gamma = \frac{2(k - d\pi + d/r)}{2k - 2d\pi + d}. \tag{2.40}$$

6^0.d) By definition of A, there exists a regular cube B such that

$$g(B)|B|^{1/2} \geq \frac{1}{2}A. \tag{2.41}$$

Let $x_0 \in B$; since B is regular, we have

$$\begin{array}{rcl} \sup_{x\in B}|g(x) - g_{x_0}(x)| & \leq & \frac{1}{4}g(B) \\ & & \text{[see (2.31)]} \\ \Rightarrow \ \frac{3}{4}g(B) \leq \max_{x\in B}|g_{x_0}(x)| & \leq & \frac{5}{4}g(B). \end{array} \tag{2.42}$$

In view of the latter inequalities and since $g_{x_0}(\cdot)$ is a polynomial of degree $k-1$, there exists a cube $B^* \subset B$ such that $|B^*| \geq P_{13}|B|$ and $|g_{x_0}(x)| \geq \frac{1}{2}g(B)$ for all $x \in B^*$, whence, in view of the first inequality in (2.42), $|g(x)| \geq \frac{1}{4}g(B)$ whenever $x \in B^*$. Combining these observations and (2.41), we conclude that

$$A \leq P_{14}|B^*|^{1/2} \min_{x\in B^*}|g(x)|,$$

so that by (2.40)

$$\begin{array}{l} \exists B^* : \\ \infty \geq r \geq \frac{2k+d}{d}p \Rightarrow \parallel g \parallel_r \leq P_{14}\left[|B^*|^{1/2}\min_{x\in B^*}|g(x)|\right]^{\gamma} \parallel D^k g \parallel_p^{1-\gamma}, \\ \gamma = \frac{2(k-d\pi+d/r)}{2k-2d\pi+d}. \end{array} \tag{2.43}$$

Consider two possible cases:
(I): $|B^*| \geq 6^d n^{-1}$;
(II): $|B^*| < 6^d n^{-1}$.

In the case of (I), since $\mathcal{B}$ is a normal system, there exists a cube $\hat{B} \in \mathcal{B}$ such that $\hat{B} \geq 6^{-d}|B^*|$ and $\hat{B} \subset B^*$, and we get

$$\parallel g \parallel_{\mathcal{B}} \geq n^{1/2}(\hat{B}) \min_{x\in\hat{B}} \geq 6^{-d/2} n \min_{x\in\hat{B}} |g(x)||\hat{B}|^{1/2}.$$

Thus, in the case of (I) relation (2.43) implies that

$$\infty \geq r \geq \frac{2k+d}{d}p \Rightarrow \parallel g \parallel_r \leq P_{15}\left(\frac{\parallel g \parallel_{\mathcal{B}}^2}{n \parallel D^k g \parallel_p^2}\right)^{\gamma/2} \parallel D^k g \parallel_p . \tag{2.44}$$

In the case of (II) relation (2.43) applied with $r = \infty$ yields

$$\begin{array}{rcl} \parallel g \parallel_\infty & \leq & P_{14}\left[|B^*|^{1/2}\min_{x\in B^*}|g(x)|\right]^{\gamma^*} \parallel D^k g \parallel_p^{1-\gamma^*} \\ & & [\gamma^* = \frac{2(k-d\pi)}{2k-2d\pi+d}] \\ & \leq & \left[|B^*|^{1/2} \parallel g \parallel_\infty\right]^{\gamma^*} \parallel D^k g \parallel_p^{1-\gamma^*} \\ \Rightarrow \ \parallel g \parallel_\infty & \leq & P_{16}|B^*|^{\frac{\gamma^*}{2(1-\gamma^*)}} \\ \Rightarrow \ \parallel g \parallel_r & \leq & P_{17}|B^*|^{\frac{\gamma}{2(1-\gamma^*)}} \parallel D^k g \parallel_p \\ & & \text{[in view of (2.43)]} \\ \Rightarrow \ \parallel g \parallel_r & \leq & P_{18} n^{-\frac{k-d\pi+d/r}{d}} \\ & & \text{[since (II) is the case].} \end{array}$$

Combining the concluding inequality with (2.44), we see that for $r \in [\frac{2k+d}{d}p, \infty]$ it holds

$$\| g \|_r \leq P_{19} \max \left\{ \left(\frac{\| g \|_{\mathcal{B}}^2}{n \| D^k g \|_p^2} \right)^{\beta_0(p,k,d,r)} \| D^k g \|_p ; n^{-\lambda_0(p,k,n,r)} \| D^k g \|_p \right\} \tag{2.45}$$

(we have used the fact that for the values of r in question one has $\gamma/2 = \beta_0(p,k,d,r)$, $\frac{k-d\pi+d/r}{d} = \lambda_0(p,k,d,r)$).

Since the values of $\beta_0(p,k,d,r), \lambda_0(p,k,d,r)$ for $r < \frac{2k+d}{d}p$ are the same as for $r = \frac{2k+d}{d}p$, relation (2.45) is in fact valid for all $r \in [1, \infty]$.

6^0.e) Since we are in the case of **A** – i.e., $[0,1]^d$ is not a regular cube – we have $\| D^k g \|_p \geq P_{20} \| g \|_\infty$. Tracing the origin of P_{20}, one can easily see that we can ensure $P_{20} > P_3$, P_3 being defined in Lemma 2.2.2. Thus, in the case under consideration the first of the alternatives stated by Lemma 2.2.2 does not take place, and therefore (2.26) is valid. Assuming that $q \geq \frac{2k+d}{2l+d}p$, let us set $r = \frac{2k+d}{d}p$, thus getting $\theta \leq \frac{(k-l)/r+\pi l}{k}$. Applying (2.26) with the indicated r and (2.45), we get for $q \geq \frac{2k+d}{2l+d}p$:

$$\begin{array}{l} l \equiv |\alpha| < k \Rightarrow \\ \| D^{(\alpha)} g \|_q \leq P_{21} \left\{ \left(\frac{\|g\|_{\mathcal{B}}^2}{n \|D^k g\|_p^2} \right)^{\beta_l(p,k,d,q)} \| D^k g \|_p ; n^{-\lambda_l(p,k,d,q)} \| D^k g \|_p \right\}. \end{array} \tag{2.46}$$

Since $\beta_l(p,k,d,q), \lambda_l(p,k,d,q)$ are independent of q in the segment $[1, \frac{2k+d}{2l+d}p]$ of values of the parameter, relation (2.46) in fact is valid for all q. Thus, we have proved (2.22) in the case of **A**.

7^0. It remains to prove (2.22) in the case of **B**, i.e., when $[0,1]^d$ is a regular cube, whence

$$\| D^k g \| \leq P_{22} \| g \|_\infty . \tag{2.47}$$

In this case we can apply (2.31) to the cube $B = [0,1]^n$ to get the inequality

$$\max_{x \in [0,1]^n} |g(x) - g_0(x)| \leq \frac{1}{4} \| g \|_\infty,$$

whence, same as in 6^0.d), there exists a cube B such that $|B| \geq P_{23}$ and $|g(x)| \geq \| g \|_\infty$ for $x \in B$. Since $\mathcal{B}$ is a normal system, there exists $B^* \in \mathcal{B}$ such that $B^* \in B$ and $|B^*| \geq P_{24}$, provided that n is large enough, and we get

$$\| g \|_{\mathcal{B}} \geq P_{25} n^{1/2} \| g \|_\infty,$$

whence

$$\| g \|_\infty \leq P_{26} \frac{\| g \|_{\mathcal{B}}}{n^{\frac{1}{2}}}. \tag{2.48}$$

Combining (2.47), (2.48) and Lemma 2.2.2, we come to (2.22). The proof of Lemma 2.2.1 is completed. ■

2.3 Appendix: Proofs of Theorems 2.1.1, 2.1.2

Proof of Theorem 2.1.1. Let us fix a C$^\infty$ function $h(\cdot) \not\equiv 0$ such that

$$\text{supp}(h) = [0,1]^d; \quad \| D^k h \|_\infty \leq 1; \quad \| h \|_\infty \leq 1. \tag{2.49}$$

Let also

$$C_1 = \min\{\| D^{(\alpha)}h \|_1 | \ 0 \le |\alpha| < k\};$$

(recall that C_i stand for positive quantities depending on k, d only). Let us fix the volume of observations $n = m^d$ and a $\Delta \in (\frac{1}{m}, \frac{1}{8})$, and let $B_1, ..., B_N$ be a maximal in cardinality system of mutually disjoint cubes with the edges Δ, all cubes of the system belonging to $[0,1]^d$. Note that the number of points from the observation grid Γ_n in every one of the cubes B_i does not exceed

$$n_\Delta = (2\Delta)^d n.$$

As it is immediately seen,

$$N \ge \max\{8; C_2\Delta^{-d}\}. \tag{2.50}$$

Let

$$h^\Delta(x) = L\Delta^{k-d\pi}h(x/\Delta),$$

and let h_j be the translation of h^δ with the support B_j, $j = 1, ..., N$; it is immediately seen that $h_j \in \mathcal{S} \equiv \mathcal{S}_d^{k,p}(L)$. Now consider N hypotheses on the distribution of observations y, j-th of the hypotheses being that the distribution is the one of the vector $y_{h_j}(\xi)$, see (2.1).

Let us fix α such that

$$l \equiv |\alpha| < k,$$

and let

$$\varepsilon(\Delta) = \frac{1}{4}C_1\Delta^{k-l-d\pi+d\theta}L.$$

We have

$$\begin{array}{rcl} i & \neq & j \Rightarrow \\ \| D^{(\alpha)}h_i - D^{(\alpha)}h_j \|_q & \ge & \| D^{(\alpha)}h_i \|_q \\ & = & L\Delta^{k-l-d\pi+d\theta} \| D^{(\alpha)}h \|_q \\ & \ge & L\Delta^{k-l-d\pi+d\theta}C_1 \\ & = & 4\varepsilon(\Delta). \end{array}$$

Consequently (cf. the proof of Proposition 1.2.3), under the assumption that the minimax q-risk of estimating $D^{(\alpha)}f$, $f \in \mathcal{S}$ is $\le \varepsilon(\Delta)$:

$$\mathcal{R}^*_{q,(\alpha)}(n; \mathcal{S}) < \varepsilon(\Delta), \tag{2.51}$$

there exists a routine for distinguishing our N hypotheses with probability to reject a hypothesis when it is true at most 1/4. On the other hand, the Kullback distance between pairs of distributions associated with our hypotheses is at most

$$\sigma^{-2}\text{Diameter}(\{h_j(\cdot)\}_{j=1}^N|\Gamma_n) \le 2\sigma^{-2}n_\Delta \| h^\Delta \|_\infty^2 \le C_3 n\sigma^{-2}L^2\Delta^{d+2(k-d\pi)}.$$

Applying the Fano inequality (1.27), we see that the assumption (2.51) implies the relation

$$(L/\sigma)^2 n\Delta^{d+2(k-d\pi)} \ge C_4 \ln N \ge C_5 \ln\frac{1}{\Delta}, \tag{2.52}$$

the concluding inequality being given by (2.50). Now let us set

$$\Delta_1 = C_6\left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2}{2k-2d\pi+d}};$$

it is clearly seen that if C_6 is a properly chosen function of k, d and (2.3) takes place, then (2.52) fails to be true when $\Delta = \Delta_1$. Consequently, for $\Delta = \Delta_1$ (2.51) cannot be valid, and we come to

$$\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}) \geq \varepsilon(\Delta_1) \geq C_7 L \left(\frac{\sigma}{L\sqrt{n}}\right)^{2\frac{k-l-d\pi+d\theta}{2k-2d\pi+d}}; \tag{2.53}$$

this is exactly the bound (2.4) for the case of large ratios q/p (i.e., $\frac{q}{p} \geq \frac{2k+d}{2l+d}$).

Now assume that (2.5) takes place, and let us set

$$\Delta_2 = F\left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{\frac{2}{2k-2d\pi+d}};$$

it is immediately seen that for properly chosen $F > 0$ (depending on k, d, ε only) relation (2.52) fails to be true when $\Delta = \Delta_2$. Consequently, for $\Delta = \Delta_2$ (2.51) cannot be valid, and we come to

$$\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}) \geq \varepsilon(\Delta_2) \geq C(\varepsilon) L \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{2\frac{k-l-d\pi+d\theta}{2k-2d\pi+d}}; \tag{2.54}$$

this is exactly the bound (2.5) for the case of large ratios q/p.

We have established the desired lower bounds for the case of large ratios q/p. The lower bound (2.4) in the case of small ratios q/p: $\frac{q}{p} < \frac{2k+d}{2l+d}$ is given by exactly the same construction as in the case of Hölder balls. Namely, let us redefine h^Δ as follows:

$$h^\Delta(x) = L\Delta^k h(x/\Delta),$$

let h_j be the translation of h^Δ with the support B_j, and let $\mathcal{F}^*_N$ be the set of 2^N functions $\sum\limits_{j=1}^{m} \varepsilon_j h_j(x)$, where $\varepsilon_j = \pm 1$. The set $\mathcal{F}^*_N$ clearly is contained in $\mathcal{S}^{k,\infty}_d(L)$ and possesses a subset $\mathcal{F}_M$ comprised of

$$M \geq 2^{N/8}$$

functions with the following property: if f, g are two distinct functions from $\mathcal{F}_M$, then f differs from g on at least $N/8$ of the cubes $B_1, ..., B_N$. Now let us fix α with $l = |\alpha| < k$; for two distinct functions $f, g \in \mathcal{F}_M$ one clearly has

$$\parallel D^{(\alpha)} f - D^{(\alpha)} g \parallel_1 \geq C_8 L \Delta^{k-l} \Delta^d N \geq C_9 L \Delta^{k-l}.$$

Setting

$$\varepsilon(\Delta) = \frac{1}{4} C_9 L \Delta^{k-l},$$

we, same as above, conclude that under the assumption that

$$\mathcal{R}^*_{1,(\alpha)}(n; \mathcal{S}^{k,\infty}_d(L)) < \varepsilon(\Delta) \tag{2.55}$$

one can "reliably" distinguish between M hypotheses on the distribution of observations (2.1), the Kullback distances between pairs of the distributions not exceeding

$$\sigma^{-2} n \max_{f,g\in\mathcal{F}_M} \parallel f - g \parallel^2_\infty \leq C_{10}(L/\sigma)^2 \Delta^{2k}.$$

Applying the Fano inequality and taking into account that $M \geq 2^{N/8} \geq \exp\{C_{11}\Delta^{-d}\}$, we see that (2.55) implies the relation

$$n(L/\sigma)^2\Delta^{2k} \geq C_{12}\Delta^{-d}. \tag{2.56}$$

Now let us set

$$\Delta = C_{13}\left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2}{2k+d}};$$

for properly chosen C_{13} and all n satisfying (2.3) the relation (2.56) (and therefore (2.55) as well) fails to be true. Thus, for the indicated values of n one has

$$\mathcal{R}^*_{1,(\alpha)}(n;\mathcal{S}^{k,\infty}_d(L)) \geq \varepsilon(\Delta) \geq C_{14}L\left(\frac{\sigma}{L\sqrt{n}}\right)^{\frac{2(k-l)}{2k+d}}.$$

Since the risk $\mathcal{R}^*_{q,(\alpha)}(n;\mathcal{S}^{k,p}_d(L))$ is nondecreasing in q and nonincreasing in p, the left hand side of this inequality is $\leq$ the one in (2.4), while the right hand side is exactly as required in (2.4) in the case of small ratios q/p. ∎

Proof of Theorem 2.1.2. Same as in the proof of Theorem 2.1.1, below C_i are positive quantities depending on k, d only.

Let $h(\cdot)$ and C_1 be the same as in the proof of Theorem 2.1.1, and let C_2 be such that

$$\text{mes}\{x \in [0,1]^d \mid |D^{(\alpha)}h| > C_2\} > C_2 \quad \forall \alpha, |\alpha| < k.$$

Let us fix L, σ, d, p, k, q satisfying the premise of Theorem 2.1.2, the volume n of observations (2.1) and α, $|\alpha| \equiv l < k$. Consider a linear estimate of $D^{(\alpha)}f$, $f \in \mathcal{S} \equiv \mathcal{S}^{k,p}_d(L)$, based on observations (2.1), let it be

$$\hat{f}_n(x;y) = \sum_\iota \phi_\iota(x)y_\iota,$$

and let ε be the worst-case, with respect to $\mathcal{S}$, q-risk of this estimate:

$$\varepsilon^2 = \sup_{f\in\mathcal{S}} \mathcal{E}\left\{\| D^{(\alpha)}f - \hat{f}_n \|_q^2\right\}.$$

We have

$$\begin{array}{rcl} \varepsilon^2 & \geq & \mathcal{E}\left\{\| \sum\limits_\iota \phi_\iota(\cdot)\sigma\xi_\iota \|_q^2\right\} \\ & \geq & \sigma^2\mathcal{E}\left\{\| \sum\limits_\iota \phi_\iota(\cdot)\xi_\iota \|_2^2\right\} \\ & & \text{[since } q \geq 2 \text{ by the premise of Theorem 2.1.2]} \\ & = & \sigma^2 \sum\limits_\iota \| \phi_\iota(\cdot) \|_2^2 \end{array} \tag{2.57}$$

and

$$\varepsilon^2 \geq \| D^{(\alpha)}f(\cdot) - \sum_\iota \phi_\iota(\cdot)f(x_\iota) \|_q^2 \quad \forall f \in \mathcal{S}. \tag{2.58}$$

Now assume that

$$\varepsilon < (0.5C_2)^{1+\theta}L, \tag{2.59}$$

and let τ be the largest integer less than the quantity

$$\left(\frac{(0.5C_2)^{1+\theta}L}{\varepsilon}\right)^{\frac{1}{k-l-d\pi+d\theta}}; \tag{2.60}$$

note that $\tau \geq 1$ by (2.59). Setting

$$\Delta = 1/\tau$$

and taking into account (2.57), we observe that there exists a cube $B \subset [0,1]^d$ such that the number $n(B)$ of observation points in (2.1) in B does not exceed $2n\Delta^d$, while

$$\sigma^2 \int\limits_B \sum_\iota \phi_\iota^2(x)dx \leq 2\Delta^d\varepsilon^2. \tag{2.61}$$

Now let $h_\Delta(\cdot)$ be the translation of the function $L\Delta^{k-d\pi}h(x/\Delta)$ such that the support of h_Δ is B. Setting

$$g_\Delta(x) = \sum_\iota \phi_\iota(x)h_\Delta(x),$$

and applying the Cauchy inequality, we get

$$\begin{array}{rcl} |g_\Delta(x)| & \leq & \| h_\Delta \|_\infty n^{1/2}(B)\left(\sum\limits_\iota \phi_\iota^2(x)\right)^{1/2} \\ & \leq & C_3L\Delta^{k-d\pi+d/2}n^{1/2}\left(\sum\limits_\iota \phi_\iota^2(x)\right)^{1/2} \\ & & [\text{since by construction } n(B) \leq 2n\Delta^d] \end{array} \tag{2.62}$$

The resulting inequality combined with (2.61) implies that there exist $B^* \subset B$ and C_4 such that

$$\begin{array}{lrcl} (a) & \mathrm{mes}\, B^* & \geq & (1-0.5C_2)\Delta^d; \\ (b) & x \in B^* \Rightarrow |g_\Delta(x)| & \leq & C_4L\sigma^{-1}\Delta^{k-d\pi+d/2}n^{1/2}\varepsilon. \end{array} \tag{2.63}$$

Now note that by construction τ is less than the quantity (2.60), so that

$$0.5C_2L\Delta^{k-d\pi-l}(0.5C_2\Delta^d)^\theta > \varepsilon. \tag{2.64}$$

We claim that

$$0.5C_2L\Delta^{k-d\pi-l} \leq C_4L\sigma^{-1}\Delta^{k-d\pi+d/2}n^{1/2}\varepsilon. \tag{2.65}$$

Indeed, assuming that the opposite inequality holds:

$$0.5C_2L\Delta^{k-d\pi-l} > C_4L\sigma^{-1}\Delta^{k-d\pi+d/2}n^{1/2}\varepsilon$$

and combining this inequality with (2.63), we would get

$$x \in B^* \Rightarrow |g_\Delta(x)| < 0.5C_2L\Delta^{k-d\pi-l} \quad [B^* \subset B, \mathrm{mes}\, B^* \geq (1-0.5C_2)\,\mathrm{mes}\, B].$$

Recalling the origin of C_2, we would further conclude that there exists $B^{**} \subset B$ such that

$$\begin{array}{c} \mathrm{mes}\, B^{**} \geq 0.5C_2\Delta^d; \\ x \in B^{**} \Rightarrow \left\{|g_\Delta(x)| \leq 0.5C_2L\Delta^{k-d\pi-l}\right\} \;\&\; \left\{|D^{(\alpha)}h_\Delta(x)| \geq C_2\Delta^{k-d\pi-l}L\right\}. \end{array}$$

Combining these observations and (2.58), we would get

$$\varepsilon \geq \| D^{(\alpha)}h_\Delta - g_\Delta \|_q \geq 0.5C_2L\Delta^{k-d\pi-l}(0.5C_2\Delta^d)^\theta,$$

which is impossible in view of (2.64).

In view of (2.65)

$$\begin{array}{rcl}\varepsilon & \geq & C_5\left(\frac{\sigma^2}{n}\right)^{1/2}\Delta^{-l-d/2}\\ & \geq & G_1\left(\frac{\sigma^2}{n}\right)^{1/2}\left(\frac{L}{\varepsilon}\right)^{\frac{d+2l}{2(k-l-d\pi+d\theta)}}\\ & & \text{[see the origin of } \Delta\text{]}\end{array}$$

with $G_1 > 0$ depending on k, p, d, q only. From the resulting inequality it follows that

$$\varepsilon > G_2 L\left(\frac{\sigma}{L\sqrt{n}}\right)^{2\mu_l(p,k,d,q)}. \tag{2.66}$$

with G_2 of the same type as G_1.

We have established the implication (2.59) $\Rightarrow$ (2.66); in view of this implication, (2.16) is valid for all large enough values of n, as stated in Theorem 2.1.2. ■

Chapter 3

Spatial adaptive estimation on Sobolev balls

3.1 Spatial adaptive estimation: the goal

We have seen what are the minimax risks of recovering functions f from Sobolev balls $\mathcal{S}_d^{k,p}(L)$ via their $n = m^d$ noisy observations

$$\begin{array}{c} y \equiv y_f(\xi) = \left\{ y_\iota = f(x_\iota) + \sigma\xi_\iota | \iota = (i_1, ..., i_d) \in \overline{[1,m]}^d \right\} \\ \left[\begin{array}{c} x_{(i_1,...,i_d)} = (i_1/m, i_2/m, ..., i_d/m)^T, \\ \xi = \{\xi_\iota\} : \xi_\iota \text{ are independent } \mathcal{N}(0,1) \end{array} \right] \end{array} \tag{3.1}$$

and have developed the associated, optimal in order up to logarithmic in n factors, estimates. These estimates, however, suffer two serious drawbacks:

- The estimates *are not adaptive to the parameters of smoothness* p, k, L of the regression function f to be recovered. An estimate depends on a particular a priori choice of these parameters and guarantees certain quality of recovering only in the case when f belongs to the corresponding Sobolev ball.

  In reality we hardly can know in advance the precise values of the parameters of smoothness of f and should therefore use certain guesses for them. If our guesses "underestimate" the smoothness of f, then the associated estimate does ensure the risk bounds corresponding to the guessed smoothness; these bounds, however, may be much worse than if we were capable to fit the estimate to the actual smoothness of f. And if our guesses for the parameters of smoothness of f "overestimate" the actual smoothness, we simply cannot guarantee anything.

- The estimates *are not spatial adaptive*: assume, e.g., that we know that the function $f : [0,1] \to \mathbf{R}$ to be recovered is continuously differentiable with, say, $\| f' \|_2 = L$, and that we know the value of L, so that there seemingly is no difficulty with tuning the recovering routine to the actual smoothness of f. Note, however, that $\| f' \|_2$ may come from a "local singularity" of f – a relatively small part of our "universe" $[0,1]$ where f varies rapidly, and there still may be large segments $B' \subset [0,1]$ where f is much more smooth than it is said by the inclusion $f \in \mathcal{S}_1^{1,2}(L)$. If we knew these "segments of high smoothness of f", along with the corresponding smoothness parameters, in advance, we could recover the function on these segments much better than it is possible on the entire $[0,1]$. However, the recovering routines we know to the moment are "too stupid" to adapt themselves to favourable local behaviour of the regression function in question.

For estimates aimed at recovering smooth regression functions, the "adaptive abilities" of an estimate can be quantified as follows.

For a cube

$$B = \{x \mid |x_i - c_i| \le h \;\, i = 1, ..., d\}$$

contained in $[0,1]^d$, let $\mathcal{S}_d^{k,p}(B;L)$ be the set of functions $f : [0,1]^d \to \mathbf{R}$ satisfying the following assumptions:

- f is continuous on $[0,1]^d$;
- f is k times differentiable on B, and $\| D^k f \|_{p,B} \le L$.

  Here $\| \cdot \|_{p,B}$ is the standard L_p-norm on B.

In this definition, similar to the definition of a Sobolev ball in Chapter 2,
- k is a positive integer – order of smoothness;
- d is a positive integer – dimensionality;
- $p \in (d, \infty]$;
- $L > 0$.

From now on, we fix the dimension d of the regression functions in question. In the sequel, we use for $\mathcal{S}_d^{k,p}(B;L)$ also the shortened notation $\mathcal{S}[\psi]$, where ψ stands for the collection of "parameters" (k, p, B, L), and call the set $\mathcal{S}[w]$ a *local* Sobolev ball.

Let us once for ever fix a "margin" – a real $\gamma \in (0,1)$ – and let B_γ, B being a cube, be the γ times smaller concentric cube:

$$\begin{array}{c} B = \{x \mid |x_i - c_i| \le h, \;\, i = 1, ..., d\} \subset [0,1]^d \\ \Downarrow \\ B_\gamma = \{x \mid |x_i - c_i| \le \gamma h, \;\, i = 1, ..., d\} \subset B \end{array}$$

Given an estimate $\widehat{f}_n$ based on observations (3.1), (i.e., a Borel real-valued function of $x \in [0,1]^d$ and $y \in \mathbf{R}^n$), let us characterize its quality on a set $\mathcal{S}[\psi]$ by the worst-case risks

$$\widehat{\mathcal{R}}_q\left(\widehat{f}_n; \mathcal{S}[\psi]\right) = \sup_{f \in \mathcal{S}[\psi]} \left(\mathcal{E}\left\{\| \widehat{f}_n(\cdot; y_f(\xi)) - f(\cdot) \|_{q,B_\gamma}^2\right\}\right)^{1/2},$$

and let

$$\widehat{\mathcal{R}}_q^*\left(n; \mathcal{S}[\psi]\right) = \inf_{\widehat{f}_n} \sup_{f \in \mathcal{S}[\psi]} \left(\mathcal{E}\left\{\| \widehat{f}_n(\cdot; y_f(\xi)) - f(\cdot) \|_{q,B_\gamma}^2\right\}\right)^{1/2}, \tag{3.2}$$

be the corresponding minimax risks [1)].

For a particular estimate $\widehat{f}_n$, the ratio

$$\frac{\widehat{\mathcal{R}}_q\left(\widehat{f}_n; \mathcal{S}[\psi]\right)}{\widehat{\mathcal{R}}_q^*\left(n; \mathcal{S}[\psi]\right)} \tag{$*$}$$

[1)] Note that we prefer to measure the estimation errors in the integral norms associated with a little bit smaller than B cube B_γ; this allows to avoid in the sequel boring analysis of "boundary effects".

measures the level of non-optimality, with respect to the q-risk, of the estimate $\widehat{f}$ on the set $\mathcal{S}[\psi]$. It is natural to measure adaptive abilities of an estimate $\widehat{f}_n$ by looking at "how wide" is the spectrum of local Sobolev balls for which the ratio (*) is "moderately large". The formal definition is as follows.

Definition 3.1.1 *Let*

1. $\mathbf{S} = \{\mathbf{S}_n\}_{n\geq 1}$ *be a "nested family" of local Sobolev balls on* $\mathbf{R}^d$*, i.e.,*

$$\mathbf{S}_n = \{\mathcal{S}[\psi] \mid \psi \in \Psi_n\}$$

and

$$\mathbf{S}_{n+1} \supset \mathbf{S}_n$$

for every n;

2. $\{\widehat{f}_n\}_{n\geq 1}$ *be an estimation method – a collection of estimates indexed by volumes* n *of observations* (3.1) *used by the estimates;*

3. $\Phi(n)$ *be a real-valued function.*

We say that the **S**-*nonoptimality index of the estimation method* $\{\widehat{f}_n\}_{n=1}^{\infty}$ *is* $\Phi(\cdot)$*, if, for every* $q \in [1,\infty]$ *and all large enough values of* n*, one has*

$$\sup_{\psi=(k,p,B,L)\in\Psi_n} \frac{\widehat{\mathcal{R}}_q\left(\widehat{f}_n; \mathcal{S}[\psi]\right)}{\widehat{\mathcal{R}}_q^*\left(n; \mathcal{S}[\psi]\right)} \leq O(\Phi(n)).$$

An "ideal" adaptive routine for recovering smooth regression functions would have a constant nonoptimality index with respect to the widest possible nested family of local Sobolev balls – the one for which $\mathbf{S}_n$, for every n, contains all local Sobolev balls. As we shall see in the mean time, such an ideal routine simply does not exist. Recently, several adaptive routines of nearly the same "adaptive power" were proposed (the wavelet-based estimators of Donoho et al. [5, 7], and Juditsky [15], adaptive kernel estimates of Lepskii, Mammen and Spokoiny [20])[2)]. What we are about to do is to build an extremely simple recovering routine with "nearly ideal" adaptive abilities – one for which the nonoptimality index with respect to certain "rapidly extending" nested family $\{\mathbf{S}_n\}$ grows with n "very slowly" – logarithmically. We shall also see that "logarithmic growth" of the nonoptimality index is an unavoidable price for ability of a routine to adapt itself to rapidly extending nested families of local Sobolev balls.

3.2 The estimate

The recovering routine we are about to build is aimed at estimating functions with order of smoothness not exceeding a given upper bound $\mu + 1$; μ (which should, of course, be a nonnegative integer) is the only parameter our construction depends upon.

[2)] In the cited papers, the smoothness of the signal is specified as membership in the Besov or Triebel spaces – extensions of the Sobolev spaces we deal with.

The idea of our construction is very simple. Given $n = m^d$ observations (3.1), we, same as in Chapter 1, use point-wise window estimator of f. Namely, to estimate f at a given point $x \in \mathrm{int}\,[0,1]^d$, we choose somehow an *admissible* window – a cube

$$B_h(x) = \{u \mid |u_i - x_i| \le h/2,\ i = 1, ..., d\} \subset [0,1]^d$$

centered at x and containing at least $(\mu+3)^d$ observation points:

$$h \ge \frac{\mu+3}{m}. \tag{3.3}$$

Note that since the window should be centered at x and be contained in $[0,1]^d$, the point x should be not too close to the boundary of $[0,1]^d$:

$$\frac{\mu+3}{2m} \le x_i \le 1 - \frac{\mu+3}{2m},\ i = 1, ..., d, \tag{3.4}$$

which we assume from now on.

The estimate $\widehat{f}_n(x; y)$ will be just the order μ window estimate (Chapter 1, Section 1.3 [3)]) *with the window width depending on x and chosen on the basis of observations.* Thus, the difference of the estimate we are about to build with the estimator from Chapter 1 is that now we choose its own window width for every point rather than to serve all points with the same window width.

The central issue is, of course, how to choose the window width for a given x, and the underlying idea (which goes back to Lepskii [19]) is as follows.

Let, as in Chapter 1,

$$\Phi_\mu(f, B_h(x)) = \min_{p \in \mathcal{P}_\mu} \max_{u \in B_h(x)} |f(u) - p(u)|,$$

$\mathcal{P}_\mu$ being the space of polynomials on $\mathbf{R}^d$ of total degree $\le \mu$. Applying Proposition 1.3.1, we come to the following upper bound on the error of estimating $f(x)$ by the estimate $\widehat{f}^h_n(x; \cdot)$ – the window estimate associated with the centered at x window of width h:

$$\mathrm{err}_h(f, x) \equiv |f(x) - \widehat{f}^h_n(x; y_f(\xi))| \le C_1 \left[\Phi_\mu(f, B_h(x)) + \frac{\sigma}{\sqrt{nh^d}}\Theta_n\right], \tag{3.5}$$

$\Theta_n = \Theta_n(\xi)$ being a deterministic function of the observation noises; from now on, C (perhaps with sub- or superscripts) are positive quantities depending on d, μ, γ only.

As we remember from (1.43), one has

$$\forall w \ge 1: \quad \mathrm{Prob}\left\{\Theta_n > O_{\mu,d}(1) w \sqrt{\ln n}\right\} \le \exp\left\{-\frac{w^2 \ln n}{2}\right\}, \tag{3.6}$$

Note that (3.5) implies that

$$\mathrm{err}_h(f, x) \le C_1 \left[\Phi_\mu(f, B_h(x)) + \frac{\sigma}{\sqrt{nh^d}}\Theta_n\right]. \tag{3.7}$$

[3)] In this chapter we assume that the window estimate associated with a window B does *not* use the observations at boundary points of the cube B; this is why we write $\mu+3$ instead of $\mu+2$ in (3.4).

Observe that the random variable Θ_n "is not too large" – (3.6) says that "typical values" of this variable do not exceed $O(\sqrt{\ln n})$. Let us fix a "safety factor" ω in such a way that the event $\Theta_n > \omega\sqrt{\ln n}$ is "highly un-probable", namely,

$$\mathrm{Prob}\left\{\Theta_n > \omega\sqrt{\ln n}\right\} \leq n^{-4(\mu+1)}; \tag{3.8}$$

by (3.6), the required ω may be chosen as a function of μ, d only.

Let us set

$$\Xi_n = \{\xi \mid \Theta_n \leq \omega\sqrt{\ln n}\}. \tag{3.9}$$

Note that (3.7) implies the "conditional" error bound

$$\begin{array}{rcl} \xi & \in & \Xi_n \Rightarrow \\ \mathrm{err}_h(f,x) & \leq & C_1\left[\Phi_\mu(f,B_h(x)) + S_n(h)\right], \\ S_n(h) & = & \frac{\sigma}{\sqrt{nh^d}}\omega\sqrt{\ln n}. \end{array} \tag{3.10}$$

The two terms in the right hand side of the resulting error bound – the *deterministic term* $\Phi_\mu(f,B_h(x))$ and the *stochastic term* $S_n(h)$ possess opposite monotonicity properties with respect to h: as h grows (i.e., as the window extends), the deterministic term does not decrease, while the stochastic term does not increase. It follows that if we were clever enough to find the "ideal window" – the one for which the deterministic term is equal to the stochastic one – we would get the best possible, up to factor 2, error bound (3.10). Of course, we never can be clever enough to specify the "ideal window", since we do not know the deterministic term. It turns out, however, that we can act nearly as if we knew everything.

Let us define the "ideal window" $B_*(x)$ as the largest admissible window for which the stochastic term dominates the deterministic one:

$$\begin{array}{rcl} B_*(x) & = & B_{h_*(x)}(x), \\ h_*(x) & = & \max\{h \mid h \geq \frac{\mu+3}{m}, B_h(x) \subset [0,1]^d, \Phi_\mu(f,B_h(x)) \leq S_n(h)\}. \end{array} \tag{3.11}$$

Note that such a window not necessarily exists: it may happen that f varies in a neighbourhood of x too rapidly, so that already for the smallest possible admissible window the deterministic term majorates the stochastic one. In this case we define $B_*(x)$ as the smallest possible window which is admissible for x. Thus, the ideal window $B_*(x)$ is well-defined for every x possessing admissible windows; we call it good if it is given by (3.11) and bad in the opposite case.

It is immediately seen that whenever $\xi \in \Xi_n$, the error bound (3.10) associated with the ideal window is, up to factor 2, better than the bound associated with any other (admissible) window, which motivates the term "ideal window".

To explain the idea of the estimate of $f(x)$ we are about to build, assume that the ideal window for x is a good one, and let $\xi \in \Xi_n$. Then the errors of all estimates $\widehat{f}_n^h(x;y)$ associated with admissible windows smaller than the ideal one are dominated by the corresponding stochastic terms:

$$\xi \in \Xi_n, h \in \left[\frac{\mu+3}{m}, h_*(x)\right] \Rightarrow \mathrm{err}_h(f,x) \leq 2C_1 S_n(h); \tag{3.12}$$

indeed, for the (good) ideal window $B_*(x)$ the deterministic term is equal to the stochastic one, so that for smaller windows the deterministic term is not greater than the stochastic one.

Now let us fix $\xi \in \Xi_n$ and call an admissible for x window $B_h(x)$ *normal*, if the associated estimate $\widehat{f}_n^h(x;y)$ differs from every estimate associated with a smaller admissible window by no more than $4C_1$ times the stochastic term of the latter estimate:

$$\begin{array}{c}\text{Window } B_h(x) \text{ is normal}\\ \Updownarrow \\ \left\{\begin{array}{l} B_h(x) \text{ is admissible} \\ \forall h' \in \left[\frac{\mu+3}{m}, h\right]: \quad |\widehat{f}_n^{h'}(x;y) - \widehat{f}_n^h(x;y)| \le 4C_1 S_n(h') \quad [y = y_f(\xi)] \end{array}\right.\end{array} \tag{3.13}$$

Note that if x possesses an admissible window, then it possesses a normal one as well (e.g., the smallest admissible for x window clearly is normal). Note also that (3.12) says that

> (!) *If $\xi \in \Xi_n$ (i.e., if Θ_n is not "pathologically large"), then the ideal window $B_*(x)$ is normal.*
>
> Indeed, for a good ideal window the claim follows from (3.12), while a bad ideal window is just the smallest window admissible for x and is therefore normal.

Now observe that the property of an admissible window to be normal is "observable" – given observations y, we can say whether a given window is or is not normal. Besides this, it is clear that among all normal windows there exists the largest one $B^+(x) = B_{h^+(x)}(x)$ (to ensure the latter property, we have redefined window estimates as ones using observations from the interior of the underlying windows rather than from entire windows). From (!) it follows that

> (!!) *If $\xi \in \Xi_n$ (i.e., if Θ_n is not "pathologically large"), then the largest normal window $B^+(x)$ contains the ideal window $B_*(x)$.*

By definition of a normal window, under the premise of (!!) we have

$$|\widehat{f}_n^{h^+(x)}(x;y) - \widehat{f}_n^{h_*(x)}(x;y)| \le 4C_1 S_n(h_*(x)),$$

and we come to the conclusion as follows:

> (*) *If $\xi \in \Xi_n$ (i.e., if Θ_n is not "pathologically large"), then the error of the estimate*
> $$\widehat{f}_n(x;y) \equiv \widehat{f}_n^{h^+(x)}(x;y)$$
> *is dominated by the error bound (3.10) associated with the ideal window:*

$$\begin{array}{rcl} \xi & \in & \Xi_n \Rightarrow \\ |\widehat{f}_n(x;y) - f(x)| & \le & 5C_1 \left[\Phi_\mu(f, B_{h_*(x)}(x)) + S_n(h_*(x))\right]. \end{array} \tag{3.14}$$

Thus, the estimate $\widehat{f}_n(\cdot;\cdot)$ – which is based solely on observations and does not require any a priori knowledge of smoothness of f – possesses basically the same accuracy as the "ideal" estimate associated with the ideal window (provided, of course, that the realization of noises is not pathological: $\xi \in \Xi_n$).

Note that the estimate $\widehat{f}_n(x;y)$ we have built – let us call it the *adaptive estimate* – depends on a single "design parameter" μ (and, of course, on σ, the volume of observations n and the dimensionality d).

3.3 Quality of estimation

Our main result is as follows:

Theorem 3.3.1 *Let $\gamma \in (0,1)$, $\mu \geq 0$ be an integer, let $\mathcal{S} = \mathcal{S}_d^{k,p}(B;L)$ be a local Sobolev ball with order of smoothness k not exceeding $\mu+1$ and with $p > d$. For properly chosen $P \geq 1$ depending solely on μ, d, p, γ and nonincreasing in $p > d$ the following statement takes place:*

If the volume $n = m^d$ of observations (3.1) *is large enough, namely,*

$$\begin{array}{c} P^{-1} n^{\frac{2k-2d\pi+d}{2d}} \geq \frac{L}{\widehat{\sigma}_n} \geq P D^{-\frac{2k-2d\pi+d}{2}}(B) \\ \left[\widehat{\sigma}_n = \sigma\sqrt{\frac{\ln n}{n}}, \quad \pi = \frac{1}{p}\right] \end{array} \tag{3.15}$$

($D(B)$ is the edge of the cube B), then for every $q \in [1,\infty]$ the worst case, with respect to $\mathcal{S}$, q-risk of the adaptive estimate $\widehat{f}_n(\cdot,\cdot)$ associated with the parameter μ can be bounded as follows (cf. (2.2)*):*

$$\begin{array}{rcl} \widehat{\mathcal{R}}_q\left(\widehat{f}_n;\mathcal{S}\right) & \equiv & \sup_{f\in\mathcal{S}}\left(\mathcal{E}\left\{\parallel \widehat{f}_n(\cdot;y_f(\xi)) - f(\cdot) \parallel_{q,B_\gamma}^2\right\}\right)^{1/2} \\ & \leq & PL\left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,d,q)} D^{d\lambda(p,k,d,q)}(B), \\ \beta(p,k,d,q) & = & \left\{\begin{array}{ll} \frac{k}{2k+d}, & \theta \geq \pi\frac{d}{2k+d} \\ \frac{k+d\theta-d\pi}{2k-2d\pi+d}, & \theta \leq \pi\frac{d}{2k+d} \end{array}\right., \\ \theta & = & \frac{1}{q}, \\ \lambda(p,k,d,q) & = & \left\{\begin{array}{ll} \theta - \frac{d\pi}{2k+d}, & \theta \geq \pi\frac{d}{2k+d} \\ 0, & \theta \leq \pi\frac{d}{2k+d} \end{array}\right.; \end{array} \tag{3.16}$$

here B_γ is the concentric to B γ times smaller in linear sizes cube.

Proof. 1^0. In the main body of the proof, we focus on the case $p, q < \infty$; the case of infinite p and/or q will be considered at the concluding step 4^0.

Let us fix a local Sobolev ball $\mathcal{S}_d^{k,p}(B;L)$ with the parameters satisfying the premise of Theorem 3.3.1 and a function f from this class.

Recall that by (2.11)

$$\begin{array}{l} \forall\,(x \in \mathrm{int}\ B)\ \ \forall\,(h, B_h(x) \subset B): \\ \Phi_\mu(f, B_h(x)) \leq P_1 h^{k-d\pi}\Omega(f, B_h(x)), \qquad \Omega(f,B') = \left(\int\limits_{B'} |D^k f(u)|^p du\right)^{1/p}; \end{array} \tag{3.17}$$

from now on, P (perhaps with sub- or superscripts) are quantities ≥ 1 depending on μ, d, γ, p only and nonincreasing in $p > d$, and $|\cdot|$ stands both for the absolute value of a real and for the Euclidean norm of a vector from $\mathbf{R}^k$.

2^0. Our central auxiliary result is as follows:

Lemma 3.3.1 *Assume that*

$$\begin{array}{lrcl} (a) & n & \geq & \left(\frac{2(\mu+3)}{(1-\gamma)D(B)}\right)^d, \\ (b) & n^{\frac{k-d\pi}{d}}\sqrt{\ln n} & \geq & P_1(\mu+3)^{k-d\pi+d/2}\frac{L}{\sigma\omega}. \end{array} \tag{3.18}$$

Given a point $x \in B_\gamma$, let us choose the largest $h = h(x)$ such that

$$\begin{array}{lrcl} (a) & h & \leq & (1-\gamma)D(B), \\ (b) & P_1 h^{k-d\pi}\Omega(f, B_h(x)) & \leq & S_n(h). \end{array} \tag{3.19}$$

Then

$$h(x) \geq \frac{\mu+3}{m}, \tag{3.20}$$

and the error at x of the adaptive estimate $\widehat{f}_n$ as applied to f can be bounded as follows:

$$\begin{array}{lrcl} (a) & \underline{\textit{in the case of } \xi \in \Xi_n} : & & \\ & |\widehat{f}_n(x;y) - f(x)| & \leq & C_2 S_n(h(x)); \\ (b) & \underline{\textit{in the case of } \ \xi \notin \Xi_n} : & & \\ & |\widehat{f}_n(x;y) - f(x)| & \leq & P_2 D^{k-d\pi}(B)L + C_2\sigma\Theta_n. \end{array} \tag{3.21}$$

Proof of Lemma. a^0. Let $h_- = \frac{\mu+3}{m}$. From (3.18) it follows that h_- satisfies (3.19.a), so that $B_{h_-}(x) \subset B$. Moreover, (3.18.b) implies that

$$P_1 h_-^{k-d\pi} L \leq S_n(h_-);$$

the latter inequality, in view of $\Omega(f, B_{h_-}(x)) \leq L$, says that h_- satisfies (3.19.b) as well. Thus, $h(x) \geq h_-$, as claimed in (3.20).

b^0. Consider the window $B_{h(x)}(x)$. By (3.19.a) it is admissible for x, while from (3.19.b) combined with (3.17) we get

$$\Phi_\mu(f, B_{h(x)}(x)) \leq S_n(h).$$

It follows that the ideal window $B_*(x)$ of x is not smaller than $B_{h(x)}(x)$ and is good.

c^0. Assume that $\xi \in \Xi_n$. Then, according to (3.14), we have

$$|\widehat{f}_n(x;y) - f(x)| \leq 5C_1\left[\Phi_\mu(f, B_{h_*(x)}(x)) + S_n(h_*(x))\right]. \tag{3.22}$$

Now, by the definition of a good ideal window,

either

$\underline{\text{case (a):}}$ $\Phi_\mu(f, B_{h_*(x)}(x)) = S_n(h_*(x))$,

or

$\underline{\text{case (b):}}$ $\Phi_\mu(f, B_{h_*(x)}(x)) \leq S_n(h_*(x))$ and $B_*(x)$ is the largest cube centered at x and contained in $[0,1]^d$.

If both cases, the right hand side in (3.22) does not exceed

$$10C_1 S_n(h_*(x)) \leq 10C_1 S_n(h(x))$$

(recall that, as we have seen, $h_*(x) \geq h(x)$), as required in (3.21.a).

d^0. Now let $\xi \notin \Xi_n$. Note that $\widehat{f}_n(x;y)$ is certain estimate $\widehat{f}^h(x;y)$ associated with a centered at x and admissible for x cube $B_h(x)$. There are two possible cases:

$\underline{\text{case (c):}}$ $B_h(x) \subset B$;

$\underline{\text{case (d):}}$ $B_h(x) \not\subset B$.

If (c) is the case, then

$$\begin{array}{rcl} |\widehat{f}_n(x;y)-f(x)| & \leq & C_1\left[\Phi_\mu(f,B_h(x))+\frac{\sigma}{\sqrt{nh^d}}\Theta_n\right] \\ & \leq & P'D^{k-d\pi}(B)L+C'\sigma\Theta_n, \end{array} \tag{3.23}$$

the concluding inequality being given by (3.17) as applied to the cube $B_h(x)\subset B$ combined with the fact that this cube is admissible for x and therefore $nh^d\geq 1$.

If (d) is the case, then the window $B_h(x)$ contains the cube $B_{h(x)}(x)$. For the estimate associated with the latter window we have (by the same reasons as in (3.5))

$$|\widehat{f}_n^{h(x)}(x;y)-f(x)|\leq P'D^{k-d\pi}(B)L+C'\sigma\Theta_n,$$

and since the estimate $\widehat{f}_n(x;y)$ is associated with a normal cube containing $B_{h(x)}(x)$, we have

$$|\widehat{f}_n^{h(x)}(x;y)-\widehat{f}_n(x;y)|\leq 4C_1S_n(h(x))\leq C''\sigma\Theta_n,$$

the concluding inequality being given by the definition of $S_n(\cdot)$ and the fact that $\omega\sqrt{\ln n}\leq\Theta_n$ due to $\xi\notin\Xi_n$. Combining our observations, we see that in both cases (c), (d) we have

$$|\widehat{f}_n(x;y)-f(x)|\leq P_2D^{k-d\pi}(B)L+C_2\sigma\Theta_n,$$

as required in (3.21.b). □

3^0. Now we are ready to complete the proof. Assume that (3.18) takes place, and let us fix q, $\frac{2k+d}{d}p\leq q<\infty$.

3^0.a) Note that for every $x\in B_\gamma$
– either

$$h(x)=(1-\gamma)D(B),$$

– or

$$\begin{array}{c} P_1h^{k-d\pi}(x)\Omega(f,B_{h(x)}(x))=S_n(h(x)) \\ \Updownarrow \\ h(x)=\left(\frac{\widehat{\sigma}_n}{P_1\Omega(f,B_{h(x)}(x))}\right)^{\frac{2}{2k+d-2d\pi}}. \end{array} \tag{3.24}$$

Let U,V be the sets of those $x\in B_\gamma$ for which the first, respectively, the second of this possibilities takes place.

If V is nonempty, let us partition it as follows.

1) Since $h(x)$ is bounded away from zero in B_γ by (3.20), we can choose $x_1\in V$ such that

$$h(x)\geq\frac{1}{2}h(x_1)\quad\forall x\in V.$$

After x_1 is chosen, we set

$$V_1=\{x\in V\mid B_{h(x)}(x)\cap B_{h(x_1)}(x_1)\neq\emptyset\}.$$

2) If the set $V\backslash V_1$ is nonempty, we apply the construction from 1) to this set, thus getting $x_2\in V\backslash V_1$ such that

$$h(x)\geq\frac{1}{2}h(x_2)\quad\forall x\in V\backslash V_1,$$

and set

$$V_2 = \{x \in V \backslash V_1 \mid B_{h(x)}(x) \cap B_{h(x_2)}(x_2) \neq \emptyset\}.$$

If the set $V \backslash (V_1 \cup V_2)$ still is nonempty, we apply the same construction to this set, thus getting x_3 and V_3, and so on.

The outlined process clearly terminates after certain step; indeed, by construction the cubes $B_{h(x_1)}(x_1), B_{h(x_2)}(x_2), ...$ are mutually disjoint and are contained in B_γ, while the sizes of these cubes are bounded away from 0. On termination, we get a collection of M points $x_1, ..., x_M \in V$ and a partition

$$V = V_1 \cup V_2 \cup ... \cup V_M$$

with the following properties:

(i) The cubes $B_{h(x_1)}(x_1), ..., B_{h(x_M)}(x_M)$ are mutually disjoint;

(ii) For every $\ell \leq M$ and every $x \in V_\ell$ we have

$$h(x) \geq \frac{1}{2} h(x_\ell) \text{ and } B_{h(x)}(x) \cap B_{h(x_\ell)}(x_\ell) \neq \emptyset.$$

We claim that also

(iii) For every $\ell \leq M$ and every $x \in V_\ell$:

$$h(x) \geq \frac{1}{2} \max \left[h(x_\ell); \| x - x_\ell \|_\infty \right]. \tag{3.25}$$

Indeed, $h(x) \geq \frac{1}{2} h(x_\ell)$ by (ii), so that it suffices to verify (3.25) in the case when $\| x - x_\ell \|_\infty \geq h(x_\ell)$. Since $B_{h(x)}(x)$ intersects $B_{h(x_\ell)}(x_\ell)$, we have $\| x - x_\ell \|_\infty \leq \frac{1}{2}(h(x) + h(x_\ell))$, whence

$$h(x) \geq 2 \| x - x_\ell \|_\infty - h(x_\ell) \geq \| x - x_\ell \|_\infty,$$

which is even more than we need.

3^0.b) Assume that $\xi \in \Xi_n$. Then

$$\begin{array}{rl}
& \| \widehat{f}_n(\cdot;y) - f(\cdot) \|_{q,B_\gamma}^q \\
\leq & C_2^q \int\limits_{B_\gamma} S_n^q(h(x))dx \quad \text{[by (3.21.}a\text{)]} \\
= & C_2^q \int\limits_{U} S_n^q(h(x))dx + C_2^q \sum\limits_{\ell=1}^{M} \int\limits_{V_\ell} S_n^q(h(x))dx \\
= & C_2^q \int\limits_{U} \left[\frac{\widehat{\sigma}_n}{((1-\gamma)D(B))^{d/2}}\right]^q dx + C_2^q \sum\limits_{\ell=1}^{M} \int\limits_{V_\ell} S_n^q(h(x))dx \\
& \text{[since } h(x) = (1-\gamma)D(B) \text{ for } x \in U\text{]} \\
\leq & C_3^q \widehat{\sigma}_n^q D^{d-dq/2}(B) + C_3^q \widehat{\sigma}_n^q \sum\limits_{\ell=1}^{M} \int\limits_{V_\ell} \left(\max\left[h(x_\ell), \| x - x_\ell \|_\infty\right]\right)^{-dq/2} dx \\
\leq & C_3^q \widehat{\sigma}_n^q D^{d-dq/2}(B) + C_4^q \widehat{\sigma}_n^q \sum\limits_{\ell=1}^{M} \int\limits_0^\infty r^{d-1} \left(\max\left[h(x_\ell), r\right]\right)^{-dq/2} dr \\
\leq & C_3^q \widehat{\sigma}_n^q D^{d-dq/2}(B) + C_5^q \widehat{\sigma}_n^q \sum\limits_{\ell=1}^{M} \left[h(x_\ell)\right]^{d-dq/2} \\
& \text{[note that } dq/2 - d + 1 \geq \frac{2k+d}{2}p - d + 1 \geq d^2/2 + 1 \\
& \text{in view of } q \geq \frac{2k+d}{d}p,\ k \geq 1 \text{ and } p > d\text{]} \\
= & C_3^q \widehat{\sigma}_n^q D^{d-dq/2}(B) + C_5^q \widehat{\sigma}_n^q \sum\limits_{\ell=1}^{M} \left[\frac{\widehat{\sigma}_n}{P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))}\right]^{\frac{2d-dq}{2k-2d\pi+d}} \qquad \text{[by (3.24)]} \\
= & C_3^q \widehat{\sigma}_n^q D^{d-dq/2}(B) + C_5^q \widehat{\sigma}_n^{2\beta(p,k,d,q)q} \sum\limits_{\ell=1}^{M} \left[P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))\right]^{\frac{dq-2d}{2k-2d\pi+d}} \\
& \text{[see the definition of } \beta(p,k,d,q)\text{]}
\end{array} \tag{3.26}$$

Now note that $\frac{dq-2d}{2k-2d\pi+d} \geq p$ in view of $q \geq \frac{2k+d}{d}p$, so that

$$\begin{array}{rl}
& \sum\limits_{\ell=1}^{M} \left[P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))\right]^{\frac{dq-2d}{2k-2d\pi+d}} \\
\leq & \left[\sum\limits_{\ell=1}^{M} \left(P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))\right)^p\right]^{\frac{dq-2d}{p(2k-2d\pi+d)}} \\
\leq & \left[P_1^p L^p\right]^{\frac{dq-2d}{p(2k-2d\pi+d)}}
\end{array}$$

(see (3.17) and take into account that the cubes $B_{h(x_\ell)}(x_\ell)$, $\ell = 1, ..., M$, are mutually disjoint by (i)). Thus, (3.26) results in

$$\begin{array}{l}
\xi \in \Xi_n \Rightarrow \\
\| \widehat{f}_n(\cdot; y_f(\xi)) - f(\cdot) \|_{q,B_\gamma} \leq C_6 \widehat{\sigma}_n D^{d\theta - d/2}(B) + P_2 \widehat{\sigma}_n^{2\beta(p,k,d,q)} L^{\frac{d-2\theta d}{2k-2d\pi+d}} \\
= C_6 \widehat{\sigma}_n D^{d\theta - d/2}(B) + P_2 L \left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,d,q)}
\end{array} \tag{3.27}$$

3^0.c) Now assume that $\xi \notin \Xi_n$. In this case, by (3.21),

$$|\widehat{f}_n(x;y) - f(x)| \leq P_2 D^{k-d\pi}(B) L + C_2 \sigma \Theta_n. \quad \forall x \in B_\gamma,$$

whence

$$\| \widehat{f}_n(\cdot;y) - f(\cdot) \|_{q,B_\gamma} \leq \left[P_2 D^{k-d\pi}(B) L + C_2 \sigma \Theta_n\right] D^{d/q}(B). \tag{3.28}$$

3^0.d) Combining (3.27) and (3.28), we get

$$\begin{array}{rcl}
 & & \left(\mathcal{E}\left\{\| \widehat{f}_n(\cdot;y) - f(\cdot) \|^2_{q,B_\gamma}\right\}\right)^{1/2} \\
 & \leq & C_7 \max\left[\widehat{\sigma}_n D^{-\frac{d-2d\theta}{2}}(B); P_4 L \left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,d,q)}; \mathcal{J}(f)\right], \\
\mathcal{J}(f) & = & \left(\mathcal{E}\left\{\chi_{\xi\not\in\Xi_n}\left[P_2 D^{2k-2d\pi}(B)L^2 + C_2\sigma^2\Theta_n^2\right]\right\}\right)^{1/2} \\
 & \leq & P_2 D^{k-d\pi}(B) L \mathrm{Prob}^{1/2}\{\xi\not\in\Xi_n\} + C_2\sigma\left(\mathrm{Prob}^{1/2}\{\xi\not\in\Xi_n\}\left(\mathcal{E}\left\{\Theta_n^4\right\}\right)^{1/2}\right)^{1/2} \\
 & \leq & P_2 D^{k-d\pi}(B) L \mathrm{Prob}^{1/2}\{\xi\not\in\Xi_n\} + C_2\sigma \mathrm{Prob}^{1/4}\{\xi\not\in\Xi_n\}\left(\mathcal{E}\left\{\Theta_n^4\right\}\right)^{1/4} \\
 & \leq & P_2 D^{k-d\pi}(B) L n^{-2(\mu+1)} + C_2\sigma n^{-(\mu+1)}\sqrt{\ln n} \\
 & & \text{[we have used (3.6) and (3.8)]}
\end{array} \tag{3.29}$$

Thus, under assumptions (3.18) for all $d < p < \infty$ and all q, $\frac{2k+d}{d}p \leq q < \infty$ we have

$$\begin{array}{rl}
 & \left(\mathcal{E}\left\{\| \widehat{f}_n(\cdot;y) - f(\cdot) \|^2_{q,B_\gamma}\right\}\right)^{1/2} \\
\leq & C_7 \max\left[\widehat{\sigma}_n D^{-\frac{d-2d\theta}{2}}(B); P_4 L \left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,d,q)}; \right. \\
 & \left. P_5 D^{k-d\pi}(B) L n^{-2(\mu+1)}; C_8 \sigma n^{-(\mu+1)}\sqrt{\ln n}\right].
\end{array} \tag{3.30}$$

Now, it is easily seen that if $P \geq 1$ is a properly chosen function of μ, d, γ, p nonincreasing in $p > d$ and (3.15) takes place, then, first, the assumption (3.18) is satisfied and, second, the right hand side in (3.30) does not exceed the quantity

$$PL\left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,d,q)} = PL\left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,d,q)} D^{d\lambda(p,k,d,q)}(B)$$

(see (3.16) and take into account that we are in the situation $q \geq \frac{2k+d}{d}p$, so that $\lambda(p,k,d,q) = 0$). We have obtained the bound (3.16) for the case of $d < p < \infty$, $\infty > q \geq \frac{2k+d}{d}p$; passing to limit as $q \to \infty$, we get the desired bound for $q = \infty$ as well.

4^0. Now let $d < p < \infty$ and $1 \leq q \leq q_* \equiv \frac{2k+d}{d}p$. By Hölder inequality,

$$\| g \|_{q,B_\gamma} \leq \| g \|_{q_*,B_\gamma} |B_\gamma|^{\frac{1}{q}-\frac{1}{q_*}},$$

whence

$$\widehat{\mathcal{R}}_q\left(\widehat{f}_n;\mathcal{S}\right) \leq \widehat{\mathcal{R}}_{q_*}\left(\widehat{f}_n;\mathcal{S}\right) D^{d(1/q-1/q_*)}(B);$$

combining this observation with the (already proved) bound (3.16) associated with $q = q_*$, we see that (3.16) is valid for all $q \in [1,\infty]$, provided that $d < p < \infty$. Passing in the resulting bound to limit as $p \to \infty$, we conclude that (3.16) is valid for all $p \in (d,\infty]$, $q \in [1,\infty]$. ■

3.4 Optimality index of the adaptive estimate

Let us first point out lower bounds for the minimax risks of estimating functions from local Sobolev balls. These bounds can be immediately derived from Theorem 2.1.1: by "similarity arguments", to recover functions from $\mathcal{S}_d^{k,p}(B;L)$ via n observations (3.1) is clearly the same as to recover functions from $\mathcal{S}_d^{k,p}([0,1]^d, L')$ via $nD^d(B)$ similar observations, where L' is readily given by the parameters of the local Sobolev ball (in fact, $L' = D^{d\pi+k}(B)L$). The results are as follows:

Theorem 3.4.1 *Let $\sigma, L > 0$, $\gamma \in (0,1)$, (p,k,d), $p > d$, $q \in [1,\infty]$ and a cube $B \subset [0,1]^d$ be given. Assume that the volume of observations n is large enough, namely,*

$$\begin{array}{c} D^{-\frac{2k-2d\pi+d}{2}}(B) \le \frac{L\sqrt{n}}{\sigma} \\ \left[\pi = \frac{1}{p}\right] \end{array} \tag{3.31}$$

Then the minimax q-risk (3.2) *of estimating functions f from the local Sobolev ball $\mathcal{S} = \mathcal{S}_d^{k,p}(B;L)$ via observations* (3.1) *can be bounded from below as*

$$\widehat{\mathcal{R}}_q^*(n;\mathcal{S}) \ge O_{k,d,\gamma}(1) L \left(\frac{\sigma}{L\sqrt{n}}\right)^{2\beta(p,k,d,q)} D^{d\lambda(p,k,d,q)}(B), \tag{3.32}$$

where $\beta(\cdot), \lambda(\cdot)$ are given by (3.16).

If the volume of observations n is so large that

$$n^{\varepsilon} D^{-\frac{2k-2d\pi+d}{2}}(B) \le \frac{L\sqrt{n}}{\sigma}. \tag{3.33}$$

for some positive ε, then in the case of "large" ratios q/p, namely, $\frac{q}{p} \ge \frac{2k+d}{d}$, the lower bound can be strengthened to

$$\widehat{\mathcal{R}}_q^*(n;\mathcal{S}) \ge O_{k,d,\gamma,\varepsilon}(1) L \left(\frac{\sigma\sqrt{\ln n}}{L\sqrt{n}}\right)^{2\beta(p,k,d,q)} D^{d\lambda(p,k,d,q)}(B). \tag{3.34}$$

Comparing the statements of Theorems 3.3.1 and 3.4.1, we come to the following

Theorem 3.4.2 *Let us fix the dimensionality d of the regression problem, a real $p > d$, a nonnegative integer μ, and let us associate with these data the nested family of local Sobolev balls*

$$\mathbf{S} \equiv \mathbf{S}^{p,d,\mu} = \{\mathbf{S}_n\}_{n=1}^{\infty}$$

defined as follows:

$$\mathbf{S}_n = \left\{ \mathcal{S}_d^{k,p'}(B;L) \left| \begin{array}{lrcl} (a) & p' & \ge & p, \\ (b) & 1 \le k & \le & \mu+1, \\ (c) & P^{-1} n^{\frac{2-2d\pi+d}{2d}} \ge \frac{L}{\widehat{\sigma}_n} & \ge & P D^{-\frac{2(\mu+1)-2d\pi+d}{2}}(B) \end{array} \right. \right\} \tag{3.35}$$

where P is given by Theorem 3.3.1, $\pi = \frac{1}{p}$ and

$$\widehat{\sigma}_n = \frac{\sigma\sqrt{\ln n}}{\sqrt{n}}.$$

The $\mathbf{S}$-nonoptimality index of the adaptive estimation method $\{\widehat{f}_n\}_{n=1}^{\infty}$ from Section 3.2 is not worse than the logarithmic in n function

$$\Phi(n) = (\ln n)^{\frac{\mu+1}{2(\mu+1)+d}}. \tag{3.36}$$

We see that the nonoptimality index of our adaptive estimate on certain nested families of local Sobolev balls is "not too large" – it grows with n logarithmically. We are about to demonstrate that this logarithmic growth is, in a sense, unavoidable price for "reasonable adaptive abilities". For the sake of definiteness, in the below statement the parameter γ from (3.2) is assumed to be 0.5.

Proposition 3.4.1 *Let d, p, μ be the same as in Theorem 3.4.2, and let $\varepsilon \in (0,1)$. Consider the nested family $\mathbf{S}$ of local Sobolev balls given by*

$$\mathbf{S}_n = \left\{ \mathcal{S}_d^{\mu+1,p}(B;L) \,\Big|\, P^{-1} n^{\varepsilon \frac{2(\mu+1)-2d\pi+d}{2d}} \geq \tfrac{L}{\widehat{\sigma}_n} \;\geq\; P D^{-\frac{2(\mu+1)-2d\pi+d}{2}}(B) \right\} \tag{3.37}$$

where, as always,

$$\widehat{\sigma}_n = \frac{\sigma\sqrt{\ln n}}{\sqrt{n}}$$

(note that for small enough ε this nested family is contained in the one of Theorem 3.4.2).

There exist positive constants C, N such that for every estimation method $\{\widehat{f}_n\}_{n=1}^{\infty}$ one has

$$n \geq N \Rightarrow$$

$$\sup_{\mathcal{S} \in \mathbf{S}_n} \frac{\widehat{\mathcal{R}}_p\left(\widehat{f}_n; \mathcal{S}\right)}{\widehat{\mathcal{R}}_p^*\left(n; \mathcal{S}\right)} \geq C(\ln n)^{\frac{(\mu+1)}{2(\mu+1)+d}}. \tag{3.38}$$

Thus, the $\mathbf{S}$-nonoptimality index of every estimation method with respect to the nested family $\mathbf{S}$ is at least $O\left((\ln n)^{\frac{\mu+1}{2(\mu+1)+d}}\right)$.

In particular, the adaptive estimation method from Section 3.2 possesses the best possible $\mathbf{S}$-nonoptimality index.

Proof of Proposition is similar to the one used by Lepskii [18] to demonstrate that it is impossible to get optimal in order adaptive to smoothness estimator of the value of a smooth regression function at a given point.

Let us fix $\kappa \in (0, \varepsilon)$ and an estimation method $\{\widehat{f}_n\}$, and let

$$k = \mu + 1.$$

1^0. Given n, consider the Sobolev ball $\mathcal{S}^n$ from $\mathbf{S}_n$ with the largest possible B, namely, $B = [0,1]^d$, and the smallest possible, for our B and n, value of L – namely,

$$L = L(n) = \widehat{\sigma}_n P. \tag{3.39}$$

Let

$$r(n) = \widehat{\mathcal{R}}_p(\widehat{f}_n; \mathcal{S}^n)$$

be the p-risk of the estimate $\widehat{f}_n$ on this ball, and let

$$\rho(n) = \widehat{\mathcal{R}}_p^*(n; \mathcal{S}^n)$$

be the corresponding minimax risk. From the results of Section 2.1 (see (2.15)) we know that

$$\begin{array}{rcl} \rho(n) & \leq & O_{p,\mu,d}(1) L(n) \left(\frac{s_n}{L(n)}\right)^{\frac{2k}{2k+d}} \\ & = & O_{p,\mu,d}(1) s_n \left(\sqrt{\ln n}\right)^{\frac{d}{2k+d}}, \\ s_n & = & \frac{\sigma}{\sqrt{n}}. \end{array} \tag{3.40}$$

Now let us set

$$h(n) = n^{-\kappa/d}; \tag{3.41}$$

for all large enough values of n, the collection $\mathbf{S}_n$ contains the family $\mathcal{F}_n$ of all local Sobolev balls $\mathcal{S}_d^{k,p}(\widehat{L}(n), B)$ with

$$\begin{array}{rcl} \widehat{L}(n) & = & \widehat{\sigma}_n P(2h(n))^{-\frac{2k-2d\pi+d}{2}}, \\ D(B) & = & 2h(n) \qquad [B \subset [0,1]^d]. \end{array} \tag{3.42}$$

Let $\widehat{r}(n)$ be the upper bound of the risks $\widehat{\mathcal{R}}_p(\widehat{f}_n;\cdot)$ over all these balls. Let also $\widehat{\rho}(n)$ be the upper bound of the minimax risks $\widehat{\mathcal{R}}_p^*(n;\cdot)$ over the same family of local Sobolev balls. From (2.15) we know that for large enough values of n one has

$$\begin{array}{rcl} \widehat{\rho}(n) & \leq & O_{p,\mu,d}(1)\widehat{L}(n)\left(\frac{s_n}{\widehat{L}(n)}\right)^{\frac{2k}{2k+d}}(2h(n))^{\frac{2\pi kd}{2k+d}} \\ & \leq & O_{p,\mu,d}(1)s_n\left(\sqrt{\ln n}\right)^{\frac{d}{2k+d}} h^{d\pi-d/2}(n). \end{array} \tag{3.43}$$

Finally, let

$$\delta = \delta(n) = \frac{2^{-d}\sigma\sqrt{\kappa \ln n}}{5\sqrt{n}} h^{-d/2}(n). \tag{3.44}$$

2^0. We claim that for all large enough values of n one has

$$\max\left[\frac{r(n)n^{\kappa/4}}{\delta}; \frac{\widehat{r}(n)}{\delta h^{d\pi}(n)}\right] \geq \frac{1}{4}. \tag{3.45}$$

Postponing for a moment the justification of our claim, let us derive from (3.45) the assertion of Proposition. Indeed, by (3.45),

– either

$$\begin{array}{rcll} r(n) & \geq & \frac{1}{4}n^{-\kappa/4}\delta & \\ & \geq & O_{p,\mu,d,\kappa}(1)s_n\sqrt{\ln n}n^{-\kappa/4}h^{-d/2}(n) & \\ & \geq & O_{p,\mu,d,\kappa}(1)s_n\sqrt{\ln n}n^{-\kappa/4}n^{\kappa/2} & \text{[see (3.41)]} \\ & \geq & O_{p,\mu,d,\kappa}(1)\left(\ln n\right)^{\frac{k}{2k+d}} n^{\kappa/4}\rho(n) & \text{[see (3.40)]}, \end{array}$$

– or

$$\begin{array}{rcll} \widehat{r}(n) & \geq & \frac{1}{4}\delta h^{d\pi}(n) & \\ & \geq & O_{p,\mu,d,\kappa}(1)s_n\sqrt{\ln n}h^{d\pi-d/2}(n) & \text{[see (3.44)]} \\ & \geq & O_{p,\mu,d,\kappa}(1)(\ln n)^{\frac{k}{2k+d}}\widehat{\rho}(n) & \text{[see (3.43)]} \end{array}$$

In both cases, the worst-case, over the local Sobolev balls from $\mathbf{S}_n$, ratio of the risks of $\widehat{f}_n$ and the minimax risks associated with the balls is at least $O_{p,\mu,d,\kappa}(1)(\ln n)^{\frac{k}{2k+d}}$, as stated by Proposition (recall that $k = \mu + 1$).

3^0. To establish (3.45), let us look what happens when $\widehat{f}_n$ is used to recover a particular function from $\mathcal{S}^n$ – namely, the function $f \equiv 0$. The result will be some random function $\tilde{f}_n$ depending deterministically on the observation noises ξ; by definition of $r(n)$, we have

$$\mathcal{E}_\xi\left\{\| \tilde{f}_n \|_p\right\} \leq r(n). \tag{3.46}$$

Lemma 3.4.1 *For all large enough values of n there exists a cube $B \subset [0,1]^d$ with edge length $h(n)$ such that the twice larger concentric cube $B^+(B)$ is contained in $[0,1]^d$ and*

$$\text{Prob}\left\{\| \tilde{f}_n \|_{p,B} > 2n^{\kappa/4}r(n)h^{d\pi}(n)\right\} \leq 2n^{-\kappa/8}, \tag{3.47}$$

the probability being taken w.r.t. the distribution of observations (3.1) *associated with* $f \equiv 0$.

Proof. Let $t = n^{\kappa/4}$, $u = t^{p/(p+1)}$, $v = t^{1/(p+1)}$, and let χ be the characteristic function of the event $\| \tilde{f}_n \|_p \leq ur(n)$. From (3.46) it follows that

$$\mathrm{Prob}\{\chi = 0\} \leq u^{-1}. \tag{3.48}$$

On the other hand, assuming that n is so large that $h(n) < 0.1$, we have

$$\begin{array}{rrcl}
 & \| (\chi\tilde{f}_n) \|_p^p & \leq & u^p r^p(n) \\
\Rightarrow & \mathcal{E}\left\{ \int\limits_{[0,1]^d} |(\chi\tilde{f}_n)(x)|^p dx \right\} & \leq & u^p r^p(n) \\
\Rightarrow & \multicolumn{3}{l}{\exists B : D(B) = h(n), B^+(B) \subset [0,1]^d \text{ and}} \\
 & \mathcal{E}\left\{ \int\limits_{B} |(\chi\tilde{f}_n)(x)|^p dx \right\} & \leq & 2u^p r^p(n) h^d(n) \\
\Rightarrow & \mathrm{Prob}\left\{ \| \chi\tilde{f}_n \|_{p,B}^p > 2v^p u^p r^p(n) h^d(n) \right\} & \leq & \frac{1}{v^p} \\
\Rightarrow & \mathrm{Prob}\left\{ \| \tilde{f}_n \|_{p,B} > 2uvr(n) h^{d\pi}(n) \right\} & \leq & \mathrm{Prob}\{ \ \| \chi\tilde{f}_n \|_{p,B}^p > \\
 & & & 2v^p u^p r^p(n) h^d(n)\} + \mathrm{Prob}\{\chi = 0\} \\
 & & \leq & \frac{1}{v^p} + \frac{1}{u} \qquad \text{[see (3.48)]}
\end{array}$$

It remains to note that $uv = t = n^{\kappa/4}$ and $u^{-1} + v^{-p} = 2t^{-p/(p+1)} \leq 2n^{-\kappa/8}$. □

Let B be given by Lemma 3.4.1 and g be a continuous function taking values between 0 and $\delta(n)$ and such that g is equal to $\delta(n)$ on $B^+(B)$ and vanishes outside twice larger than $B^+(B)$ concentric to $B^+(B)$ cube. Consider the following two hypotheses on the distribution of observations (3.1): H_0 says that the observations come from $f \equiv 0$, while H_1 says that they come from $f = g$. Let us associate with $\widehat{f}_n$ the following procedure for distinguishing between H_0 and H_1 via observations (3.1):

> *Given y, we build the function $\widehat{f}_n(\cdot, y)$ and restrict it on the cube B. If the p-norm of this restriction is $\leq 0.5\delta h^{d\pi}(n)$, we accept H_0, otherwise we accept H_1.*

We claim that if (3.45) is *not* valid, then our procedure possesses the following properties:

(a) probability $p_{1|1}$ to accept H_1 in the case when H_1 is true is at least 1/2;
(b) probability $p_{1|0}$ to accept H_1 in the case when H_0 is true is at most $2n^{-\kappa/8}$.

Indeed, we have $g \in \mathcal{S}_d^{\mu+1,p}(\widehat{L}(n), B^+(B))$. Now, whenever H_1 is true and is rejected by our procedure, the $\| \cdot \|_{p,B}$-error of estimate $\widehat{f}_n$, the true regression function being g, is at least $0.5\delta h^{d\pi}(n)$; since the expectation of this error is at most $\widehat{r}(n)$ by origin of the latter quantity, $1 - p_{1|1}$ is at most $2\widehat{r}(n)(\delta h^{d\pi}(n))^{-1}$; if (3.45) is not valid, the latter quantity is $\leq 1/2$, so that $p_{1|1} \geq 1/2$, as claimed in (a). Now, whenever H_0 is true and is rejected by our procedure, we have $\| \tilde{f}_n \|_{p,B} \geq 0.5\delta h^{d\pi}(n)$. When (3.45) is not valid, we have $0.5\delta h^{d\pi}(n) > 2r(n) n^{\kappa/4} h^{d\pi}(n)$, so that here $\| \tilde{f} \|_{p,B} \geq 2r(n) n^{\kappa/4} h^{d\pi}(n)$, and the H_0-probability of the latter event, by (3.47), does not exceed $2n^{-\kappa/8}$, as claimed in (b).

On the other hand, the Kullback distance between the distributions of observations associated with the hypotheses H_i, $i = 0, 1$, by construction does not exceed

$$\mathcal{K} = (4h(n))^d \sigma^{-2} \delta^2 n = \frac{\kappa \ln n}{25}.$$

As we remember from the proof of the Fano inequality (see Remark 1.4.1), the Kullback distance may only decrease when we pass from the original pair of distributions to their "deterministic transforms" – to the distributions of the results of our routine for hypotheses testing. Thus, denoting by $p_{i|j}$ the probability to accept H_i when the true hypothesis is H_j, $i, j = 0, 1$, we get

$$\begin{array}{rcl}
\frac{\kappa \ln n}{25} & \geq & \mathcal{K} \\
& \geq & p_{1|1} \ln\left(\frac{p_{1|1}}{p_{1|0}}\right) + p_{0|1} \ln\left(\frac{p_{0|1}}{p_{0|0}}\right) \\
& = & \left[p_{1|1} \ln p_{1|1} + p_{0|1} \ln p_{0|1}\right] - p_{1|1} \ln p_{1|0} - p_{0|1} \ln p_{0|0} \\
& \geq & -\ln 2 - p_{1|1} \ln p_{1|0} \\
& \geq & -\ln 2 + \frac{1}{2} \ln\left(\frac{n^{\kappa/8}}{2}\right) \qquad \text{[we have used (a) and (b)]}
\end{array}$$

The resulting inequality cannot be valid for large values of n, so that for these values of n (3.45) does take place. ∎

We conclude this chapter with demonstrating a reasonably good numerical behaviour of the adaptive estimate we have built (for details of implementation, see [9]). Our numerical results deal with univariate functions and two-dimensional images. As the test univariate signals, we used the functions *Blocks*, *Bumps*, *HeaviSine* and *Doppler* given in [6, 5]. The level of noise in experiments is characterized by the *signal-to-noise ratio*

$$\left(\frac{\sum_\iota f^2(x_\iota)}{n\sigma^2}\right)^{1/2};$$

the less it is, the more difficult is to recover the regression function.

Figure 3.1: “Blocks”, $n = 2048$.

True signal

Observations Recovered signal

Signal-to-noise ratio = 7

Observations Recovered signal

Signal-to-noise ratio = 3

Figure 3.2: “Bumps”, $n = 2048$.

True signal

Observations Recovered signal

Signal-to-noise ratio = 7

Observations Recovered signal

Signal-to-noise ratio = 3

Figure 3.3: “HeavySine”, $n = 2048$.

True signal

Observations Recovered signal

Signal-to-noise ratio = 7

Observations Recovered signal

Signal-to-noise ratio = 3

Figure 3.4: “Doppler”, $n = 2048$.

True signal

Observations Recovered signal

Signal-to-noise ratio = 7

Observations Recovered signal

Signal-to-noise ratio = 3

Figure 3.5: “Ball”, $n = 512^2$.

True image

Observations Recovered image

Signal-to-noise ratio = 3

Observations Recovered image

Signal-to-noise ratio = 1

Figure 3.6: “Lennon”, $n = 256^2$.

True image

Observations Recovered image

Signal-to-noise ratio = 3

Observations Recovered image

Signal-to-noise ratio = 1

Chapter 4

Estimating signals satisfying differential inequalities

4.1 The goal

Let us look again at the problem of recovering a univariate function $f : [0,1] \to \mathbf{R}$ via n equidistant observations

$$y = y_f(\xi) = \{y_t = f(t/n) + \sigma\xi_t\}_{t=1}^n,$$

$\xi = \{\xi_t\}_{t=1}^n$ be a collection of independent $\mathcal{N}(0,1)$ noises. To the moment we have developed a number of theoretically efficient techniques for solving this problem in the case when f is a smooth function – it belongs to a (local) Sobolev ball. At the same time, these techniques fail to recover regression functions, even very simple ones, possessing "bad" parameters of smoothness. Assume, e.g., that $f(x) = \sin(\omega x)$ and the frequency of this sine may be large. In spite of the fact that sine is extremely "regular", there does not exist a single Sobolev ball containing sines of all frequencies, As a result, with the techniques we have to the moment (as with all other traditional regression estimation techniques – all of them are aimed at estimating functions with somehow fixed smoothness parameters) the quality of recovering a sine is the worse the larger is the frequency, and no uniform in frequency rate of convergence is guaranteed. In fact, all our estimates as applied to a sine of high frequency will recover it as zero. The same unpleasant phenomenon occurs when the function f to be recovered is an "amplitude modulation" of a smooth (belonging to a given Sobolev ball) signal g:

$$f(x) = g(x)\sin(\omega x + \phi) \tag{4.1}$$

and the frequency ω is large:

We are about to extend our estimation techniques from the classes of smooth functions to wider classes including, in particular, the signals of the type (4.1). This extension comes from the following simple observation:

A Sobolev ball $\mathcal{S}^{k,p}(B;L) \equiv \mathcal{S}_1^{k,p}(B;L)$ is comprised of functions satisfying the "differential inequality":

$$\| \, r\left(\frac{d}{dx}\right) f \, \|_{p,B} \leq L \tag{4.2}$$

Figure 4.1: An “amplitude modulated” signal.

> *associated with the linear differential operator*
>
> $$r\left(\frac{d}{dx}\right) = \frac{d^k}{dx^k}$$
>
> *of order k. A natural way to extend this particular family of functions is to consider functions satisfying differential inequalities associated with other linear differential operators of order k, each function being “served” by its own operator.*

Guided by the above observation, we come to the family of functions $\mathcal{D}^{k,p}(B;L)$ defined as follows:

Definition 4.1.1 *Let k be a positive integer, $p \in [1,\infty]$, $L > 0$ and B be a segment contained in $[0,1]$.*

We say that a function

$$f : [0,1] \to \mathbf{C} \quad {}^{1)}$$

is contained in the family $\mathcal{D}^{k,p}(B;L)$, if f is $k-1$ times continuously differentiable, $f^{(k-1)}$ is absolutely continuous and there exists a linear differential operator with constant (perhaps complex-valued) coefficients

$$r\left(\frac{d}{dx}\right) = \frac{d^k}{dx^k} + r_1\frac{d^{k-1}}{dx^{k-1}} + ... + r_{k-1}\frac{d}{dx} + r_k$$

such that

$$\parallel r\left(\frac{d}{dx}\right) f \parallel_{p,B} \leq L.$$

The families $\mathcal{D}^{k,p}(B;L)$ are wide enough to contain both functions from the usual Sobolev ball $\mathcal{S}^{k,p}(B;L)$ and the sines of arbitrary frequencies: a sine is a solution of a homogeneous differential equation of order 2, so that $\sin(\omega t + \phi) \in \mathcal{D}^{2,p}([0,1];0)$.

[1)] In this chapter it is more convenient to deal with complex-valued functions than with real-valued ones.

As about "modulated signals" (4.1), each of them can be represented as a sum of two signals from $\mathcal{D}^{k,p}(\cdot,\cdot)$. Indeed,

$$\begin{array}{rcl} g & \in & \mathcal{S}^{k,p}(L) \Rightarrow \\ f(x) \equiv g(x)\sin(\omega x+\phi) & = & f_1(x)+f_2(x), \\ f_1(x) & = & \frac{1}{2i}g(x)\exp\{i(\omega x+\phi)\}, \\ f_2(x) & = & -\frac{1}{2i}g(x)\exp\{-i(\omega x+\phi)\}; \end{array}$$

setting

$$r^1(z)=(z-i\omega)^k, \quad r^2(z)=(z+i\omega)^k,$$

we have

$$\begin{array}{c} \left(r^j\left(\frac{d}{dx}\right)f_j\right)(x)=\varepsilon_j\frac{1}{2i}\exp\{\varepsilon_j i(\omega x+\phi)\}g^{(k)}(x),\ \varepsilon_j=(-1)^{j-1} \\ \Downarrow \\ \| r^j\left(\frac{d}{dx}\right)f_j \|_p \leq \frac{1}{2}L, \end{array}$$

so that function (4.1) associated with $g\in\mathcal{S}^{k,p}(B;L)$ is the sum of two functions from $\mathcal{D}^{k,p}(B;L/2)$.

Motivated by the latter example, we see that it makes sense to know how to recover regression functions from the families $\mathcal{W}^{l,k,p}(B;L)$ defined as follows:

Definition 4.1.2 *Let k,l be positive integers, $p\in[1,\infty]$, $L>0$ and B be a segment contained in* $[0,1]$.

We say that a function $f:[0,1]\to\mathbf{C}$ belongs to the family $\mathcal{W}^{l,k,p}(B;L)$, if f can be represented as

$$f(x)=\sum_{j=1}^{l}f_j(x)$$

with $f_j\in\mathcal{D}^{k,p}(B;L_j)$ and

$$\sum_{j=1}^{l}L_j\leq L.$$

Below, we build estimates for regression functions from classes $\mathcal{W}^{l,k,p}(B;L)$; since we have agreed to work with complex-valued functions, it makes sense to speak about complex-valued noises, so that our model of observations from now on will be

$$y=y_f(\xi)=\{y_t=f(t/n)+\sigma\xi_t\}_{t=1}^{n}, \tag{4.3}$$

where $\xi=\{\xi_t\}_{t=1}^n$ is a collection of independent complex-valued standard Gaussian noises (i.e., of random 2D real Gaussian vectors with zero mean and the unit covariance matrix). Of course, if the actual observations are real, we always can add to them artificial imaginary Gaussian noises to fit the model (4.3).

Note that when recovering a highly oscillating function f via observations (4.3), we may hope to say something reasonable *only about the restriction of f on the observation grid* $\Gamma_n=\{x_t=t/n\}_{t=1}^n$, and not on the behaviour of f outside the grid. Indeed, it may happen that f is a sine of amplitude 1 which vanishes on Γ_n, so that observations (4.3) give no hint that f is not identically zero. By the just indicated reason, in what follows we are interested to recover functions *on the observation grid* Γ_n only, and we measure the estimation error in the "discrete versions" of q-norms

$$|g|_{q,B} = \left(\frac{1}{n}\sum_{x\in\Gamma_n\cap B}|g(x)|^q\right)^{1/q},$$

with the standard interpretation of the right hand side in the case of $q=\infty$; here g is a complex-valued function defined at least on Γ_n, and $B\subset[0,1]$ is a segment.

We shall see that our possibilities to recover functions from class $\mathcal{W}^{l,k,p}(B;L)$ are essentially the same as in the case when the functions belong to the Sobolev ball $\mathcal{S}_1^{k,p}(B;L)$ (up to the fact that now we are recovering the restriction of a function on Γ_n rather than the function itself), in spite of the fact that the former class is "incomparably wider" than the latter one.

Our strategy will be as follows. When estimating a smooth function f – one satisfying the differential inequality

$$\parallel r_k\left(\frac{d}{dx}\right)f\parallel_p\leq L \qquad [r_k(z)=z^k]$$

– at a point x, we observe that locally it can be well approximated by a polynomial of degree $k-1$, i.e., by a solution of the *homogeneous* differential equation

$$r_k\left(\frac{d}{dx}\right)p=0$$

associated with our differential inequality; and when estimating $f(x)$, we act as if f were equal to its local polynomial approximation in the neighbourhood of x used by the estimate.

Basically the same strategy can be used for estimating a regression function satisfying a general differential inequality

$$\parallel r\left(\frac{d}{dx}\right)f\parallel_p\leq L, \qquad [\deg r=k]$$

with the only difference that now a "local model" of f should be a solution of the associated homogeneous equation

$$r\left(\frac{d}{dx}\right)p=0 \tag{4.4}$$

rather than an algebraic polynomial. This is, however, an essential difference: it is easy to act "as if f were an algebraic polynomial", because we know very well how to recover algebraic polynomials of a given order from noisy observations. Now we need to solve similar recovering problem for a solution to *unknown* homogeneous differential equation of a given order, which by itself is a nontrivial problem. We start with this problem; after it is resolved, the remaining part of the job will be carried out in the same manner as in the standard case of estimating smooth regression functions.

4.2 Estimating solutions of homogeneous equations

When restricting a solution of a homogeneous differential equation (4.4) on an equidistant grid, we get a sequence satisfying a homogeneous finite-difference equation. Since we are interested to recover signals on the grid only, we may temporarily forget about "continuous time" and focus on estimating sequences satisfying finite-difference equations.

4.2.1 Preliminaries

Space of sequences. Let $\mathcal{F}$ be the space of two-sided complex-valued sequences $\phi = \{\phi_t\}_{t\in\mathbf{Z}}$, and $\mathcal{F}_*$ be the subspace of "finite" sequences – those with finitely many nonzero entries. In what follows we identify a sequence $\phi = \{\phi_t\} \in \mathcal{F}_*$ with the rational function

$$\phi(z) = \sum_t \phi_t z^t.$$

The space $\mathcal{F}$ is equipped with the natural linear operations - addition and multiplication by scalars from $\mathbf{C}$, and $\mathcal{F}_*$ – also with multiplication

$$(\phi\psi)(z) = \phi(z)\psi(z)$$

(which corresponds to the convolution in the initial "sequence" representation of the elements of $\mathcal{F}_*$). For $\phi \in \mathcal{F}_*$ we denote by $\deg(\phi)$ the minimum of those $\tau \geq 0$ for which $\phi_t = 0$, $|t| > \tau$, so that

$$\phi(z) = \sum_{|t|\leq \deg(\phi)} \phi_t z^t;$$

if ϕ is a sequence with infinitely many nonzero entries then by definition $\deg(\phi) = \infty$. Let $\mathcal{F}_N$ denote the subspace of $\mathcal{F}$ comprised of all ϕ with $\deg(\phi) \leq N$; clearly, one always has $\phi \in \mathcal{F}_{\deg(\phi)}$ (by definition $\mathcal{F}_\infty \equiv \mathcal{F}$).

Further, let Δ stand for the backward shift operator on $\mathcal{F}$:

$$(\Delta\phi)_t = \phi_{t-1}.$$

Given $\phi \in \mathcal{F}_*$, we can associate with ϕ the finite difference operator $\phi(\Delta)$ on $\mathcal{F}$:

$$\phi(\Delta)\psi = \left\{\sum_s \phi_s \psi_{t-s}\right\}_{t\in\mathbf{Z}}, \quad \psi \in \mathcal{F}.$$

Discrete Fourier transformation. Let N be a nonnegative integer, and let G_N be the set of all roots

$$\zeta_k = \exp\left\{i\frac{2\pi k}{2N+1}\right\}, \quad k = 0, 1, ..., 2N,$$

of the unity of the degree $2N+1$. Let $\mathbf{C}(G_N)$ be the space of complex–valued functions on G_N, i.e., the vector space $\mathbf{C}^{2N+1}$ with the entries of the vectors indexed by the elements of G_N. We define the discrete Fourier transformation $F_N : \mathcal{F} \to \mathbf{C}(G_N)$ by the usual formula

$$(F_N\phi)(\zeta) = \frac{1}{\sqrt{2N+1}} \sum_{|t|\le N} \phi_t \zeta^t, \quad \zeta \in G_N.$$

Clearly, for $\phi \in \mathcal{F}_N$ one has

$$(F_N\phi)(\zeta) = \frac{1}{\sqrt{2N+1}} \phi(\zeta), \quad \zeta \in G_N.$$

The inverse Fourier transformation is given by

$$\phi_t = \frac{1}{\sqrt{2N+1}} \sum_{\zeta \in G_N} (F_N\phi)(\zeta)\zeta^{-t}, \ \ |t| \le N.$$

Norms on $\mathcal{F}$. For $0 \le N \le \infty$ and $p \in [1,\infty]$ let

$$\| \phi \|_{p,N} = \left(\sum_{t=-N}^{N} |\phi_t|^p \right)^{1/p}$$

(if $p = \infty$, then the right hand side, as usual, is $\max_{|t|\le N} |\phi_t|$). This is the standard p-seminorm on $\mathcal{F}$; restricted on $\mathcal{F}_N$, this is an actual norm. We shall omit explicit indicating N in the notation of the norm in the case of $N = \infty$; thus, $\| \phi \|_p$ is the same as $\| \phi \|_{p,\infty}$.

Let $\phi \in \mathcal{F}$ be such that there exists a positive integer k satisfying

(i) $\| \phi \|_{\infty,k} = 1$,

(ii) the smallest of t's with nonzero ϕ_t is zero, and the largest is $\le k$;

in this case we say that $\phi \in \mathcal{F}$ is *normalized polynomial of the degree* $\le k$. In the other words, the sequence ϕ from $\mathcal{F}$ is normalized polynomial of the degree $\le k$ if it can be identified with polynomial $\phi(z) = \sum_{t=0}^{k} \phi_t z^t$ with $\max_{0\le t\le k} |\phi_t| = 1$.

It is well–known that the Fourier transformation F_N being restricted on $\mathcal{F}_N$ is an isometry in 2-norms:

$$\langle \phi, \psi \rangle_N \equiv \sum_{|t|\le N} \phi_t \overline{\psi}_t = \langle F_N\phi, F_N\psi \rangle \equiv \sum_{\zeta \in G_N} (F_N\phi)(\zeta)\overline{(F_N\psi)(\zeta)}, \quad \phi, \psi \in \mathcal{F}, \qquad (4.5)$$

where $\bar{a}$ denotes the conjugate of $a \in \mathbf{C}$. The space $\mathbf{C}(G_N)$ also can be equipped with p-norms

$$\| g(\cdot) \|_p = \left(\sum_{\zeta \in G_N} |g(\zeta)|^p \right)^{1/p}$$

with the already indicated standard interpretation of the right hand side in the case of $p = \infty$. Via Fourier transformation, the norms on $\mathbf{C}(G_N)$ can be translated to $\mathcal{F}$, and we set

$$\| \phi \|_{p,N}^{*} = \| F_N\phi \|_p;$$

these are seminorms on $\mathcal{F}$, and their restrictions on $\mathcal{F}_N$ are norms on the latter subspace.

Useful inequalities. We list here several inequalities which are used repeatedly in the sequel.

$$\| \phi \|_{2,N} = \| \phi \|^*_{2,N}, \tag{4.6}$$

$$\| \phi\psi \|_{p,N} \leq \| \phi \|_1 \| \psi \|_{p,N+\deg(\phi)}, \tag{4.7}$$

$$\| \phi \|_{1,N} \leq \| \phi \|^*_{1,N} \sqrt{2N+1}, \tag{4.8}$$

$$\| \phi \|^*_{\infty,N} \leq (2N+1)^{1/2-1/p} \| \phi \|_{p,N}, \tag{4.9}$$

$$\deg(\phi) + \deg(\psi) \leq N \Rightarrow \| \phi\psi \|^*_{1,N} \leq \| \phi \|_{1,N} \| \psi \|^*_{1,N}, \tag{4.10}$$

Proofs of the above inequalities are straightforward; we note only that (4.6) is the Parseval equality, and (4.7) is the Young inequality.

4.2.2 Estimating sequences

The problem we want now to focus on is as follows. Assume we are given noisy observations

$$y = y_f(\xi) = \{y_t = f_t + \sigma\xi_t\}_{t\in\mathbf{Z}} \tag{4.11}$$

of a sequence $f \in \mathcal{F}$; here $\{\xi_t\}$ is a sequence of independent random Gaussian 2D noises with zero mean and the unit covariance matrix.

Assume that f "nearly satisfies" an (unknown) finite-difference equation of a given order k:

$$|\phi(\Delta)f| \leq \varepsilon, \tag{4.12}$$

for some normalized polynomial ϕ of degree $\leq k$; here ε is small. We want to recover a given entry of f, say, f_0, via a given number of observations (4.11) around the time instant $t = 0$. For our purposes it is convenient to parameterize the number of observations we use to estimate f_0 as

$$8\mu T + 1,$$

where μ is a once for ever a priori fixed positive integer ("order" of the estimate to be built) and $T \in \mathbf{N}$ is the parameter ("window width"). Thus, we want to estimate f_0 via the vector of observations

$$y^T = \{y_t\}_{|t|\leq 4\mu T}.$$

The idea of the estimate we are about to build is very simple. Assume for a moment that our signal satisfies a homogeneous difference equation – $\varepsilon = 0$. If we knew the underlying difference operator ϕ, we could use the Least Squares approach to estimate f_τ, and the resulting estimator would be linear in observations. By analogy, let us postulate a "filter" form

$$\widehat{f}_\tau = -\sum_{|s|\leq 2\mu T} \psi_s y_{\tau-s}. \tag{4.13}$$

of estimate of f_τ in the case of unknown ϕ as well (By reasons which will become clear in a moment, our filter recovers f_τ via reduced number of observations – $4\mu T+1$ observations around τ instead of the allowed number $8\mu T + 1$.)

If we knew ϕ, we could specify "good weights" ψ_s in advance, as we did it when estimating algebraic polynomials. Since we do not know ϕ, we should determine the weights ψ_s on the basis of observations. The first requirement to the weights is that $\sum\limits_{|s|\le 2\mu T} |\psi_s|^2$ should be small enough in order to suppress the observation noises. Imposing such a restriction on the weights ψ_s, we can determine the weights themselves by a kind of "bootstrapping" – by fitting the output $\{\widehat{f}_\tau\}$ of our filter to its input – to the sequence of observations $\{y_t\}$. Our hope is that if our filter suppresses the noises, then the only possibility for the output to "reproduce" the input is to reproduce its deterministic component f – since the "white noise" component of the input (the sequence of observation noises) is "irreproducible". In other words, let us form the *residual* $g[T,\psi,y] \in \mathcal{F}$ according to

$$g_t[T,\psi,y] = \begin{cases} y_t + \sum\limits_{|s|\le 2\mu T} \psi_s y_{t-s}, & |t| \le 2\mu T \\ 0, & |t| > 2\mu T \end{cases}, \tag{4.14}$$

and let us choose the weights by minimizing a properly chosen norm of this residual in ψ under the restriction that the filter associated with ψ "suppresses the observation noises". After these weights are found, we use them to build the estimate $\widehat{f}_0$ of f_0 according to (4.13).

Note that a procedure of the outlined type indeed recovers f_0 via y^T, since our residual depends on y^T rather than on the entire sequence of observations (the reason to reduce the number of observations used by $\widehat{f}$ was exactly the desire to ensure the latter property).

We have outlined our "estimation strategy" up to the following two issues:

(a) what is an appropriate for us form of the restriction "the filter with weights ψ suppresses the observations noises";

(b) what is a proper choice of the norm used to measure the residual.

Surprisingly enough, it turns out that it makes sense to ensure (a) by imposing an upper bound on the $\|\cdot\|_1$-norm of the Fourier transform of ψ, and to use in (b) the $\|\cdot\|_\infty$-norm of the Fourier transform of the residual. The "common sense" reason for such a choice is that the difference between a highly oscillating "regular" signal observed in noise and the noise itself is much better seen in the frequency domain than in the time domain (look at the plots below!)

The estimate we have outlined formally is defined as follows. Let us fix a positive integer μ – the *order* of our estimate. For every positive T we define the estimate

$$\widehat{f}[T,y]$$

of f_0 via observations $y^T = \{y_t\}_{|t|\le 4\mu T}$, namely,

- We associate with T, y the optimization problem

$$\begin{array}{ll} (P_T[y]): & \\ & \| g[T,\psi,y] \|^*_{\infty,2\mu T} \to \min \\ \text{s.t.} & \\ (a) & \psi \in \mathcal{F}_{2\mu T}; \\ (b) & \| \psi \|^*_{1,2\mu T} \le \alpha(T) \equiv 2^{2\mu+2}\sqrt{\frac{\mu}{T}}. \end{array}$$

As we remember, for $\psi \in \mathcal{F}_{2\mu T}$ the residual $g[T,\psi,y]$ depends on y^T only, so that our optimization problem involves only the observations y_t with $|t| \le 4\mu T$. The problem

$$\overbrace{\Rightarrow}^{F_{128}}$$

$$\overbrace{\Rightarrow}^{F_{128}}$$

Figure 4.2: Who is who?

Up: a noisy sum of 3 sines and the modulus of its Fourier transform
(257 observations, signal-to-noise ratio 1)

Down: noise and the modulus of its Fourier transform clearly is a convex optimization program, and its solution $\widehat{\psi}[T, y^T]$ can be chosen to be a Borel function of observations. By definition,

$$\widehat{f}[T, y] = -\sum_{|s| \le 2\mu T} \widehat{\psi}_s[T, y^T] y_{-s}.$$

The main result on the estimate we have built is as follows.

Theorem 4.2.1 *Let*

- *k, l be two positive integers such that $kl \le \mu$, μ being the order of the estimate $\widehat{f}[\cdot, \cdot]$;*
- *T be a positive integer;*
- *$\varepsilon \ge 0$.*

Assume that the sequence f underlying observations (4.11) *can be decomposed as*

$$f = \sum_{j=1}^{l} f^j \tag{4.15}$$

and for every component f^j there exists normalized polynomial η^j of degree $\le k$ such that

$$\sum_{j=1}^{l} \| \eta^j(\Delta) f^j \|_{p,4\mu T} \le \varepsilon. \tag{4.16}$$

Then the inaccuracy of the estimate $\widehat{f}[T, \cdot]$ of f_0 can be bounded from above as follows:

$$|f_0 - \widehat{f}[T, y]| \le C \left[T^{k-1/p} \varepsilon + \sigma T^{-1/2} \Theta^T(\xi) \right] \tag{4.17}$$

where C depends on μ only and

$$\begin{array}{rcl} \Theta^T(\xi) & = & \max_{|s| \le 2\mu T} \| \Delta^s \xi \|^*_{\infty, 2\mu T} \\ & = & \max\limits_{|s| \le 2\mu T} \max\limits_{\zeta \in G_{2\mu T}} \dfrac{1}{\sqrt{4\mu T + 1}} \left| \sum\limits_{t=-2\mu T}^{2\mu T} \xi_{t-s} \zeta^t \right|. \end{array} \tag{4.18}$$

Proof. Let us fix f satisfying the premise of our theorem, and let η^j, f_j be the associated sequences.

1^0. We start with

Lemma 4.2.1 *There exists $\eta \in \mathcal{F}_{2\mu T}$ such that*

(i) $\eta(z) = \delta(z) + \omega(z)$, $\delta(z) \equiv 1$ *being the convolution unit, with*

$$\| \eta \|_1 \le 2^\mu \tag{4.19}$$

and

$$\| \omega \|^*_{1,N} \le 2^\mu \frac{\sqrt{2N+1}}{T} \quad \forall N \ge 2\mu T; \tag{4.20}$$

(ii) *for every $j = 1, ..., l$ there exists representation*

$$\eta(z) = \eta^j(z) \rho^j(z) : \quad \rho^j \in \mathcal{F}_{2\mu T}, \| \rho^j \|_\infty \le 2^{2\mu} T^{k-1}. \tag{4.21}$$

The proof of Lemma is placed in Section 4.4.

2^0. Let ψ be a feasible solution of the optimization problem $(P_T[y_f(\xi)])$. We claim that the value of the objective at ψ can be bounded from above as follows:

$$\begin{array}{rcl}\| \, g[T,\psi,y_f(\xi)] \, \|^*_{\infty,2\mu T} & \leq & \| \, g^*[\psi,f] \, \|^*_{\infty,2\mu T} + 2^{2\mu+4}\mu\sigma\Theta^T(\xi), \\ g^*_t[\psi,f] & = & \left\{\begin{array}{ll} f_t + \sum\limits_{|s|\leq 2\mu T} \psi_s f_{t-s}, & |t|\leq 2\mu T \\ 0, & |t| > 2\mu T \end{array}\right. . \end{array} \tag{4.22}$$

Indeed, we have

$$\begin{array}{lrcl} (a) & g[T,\psi,y_f(\xi)] & = & g^*[\psi,f] + h[\psi,\xi], \\ (b) & h[\psi,\xi] & = & \left\{\begin{array}{ll} \sigma\xi_t + \sum\limits_{|s|\leq 2\mu T} \psi_s \sigma\xi_{t-s}, & |t|\leq 2\mu T \\ 0, & |t| > 2\mu T \end{array}\right. \\ \Rightarrow & & & \\ & \| \, g[T,\psi,y_f(\xi)] \, \|^*_{\infty,2\mu T} & \leq & \| \, g^*[\psi,f] \, \|^*_{\infty,2\mu T} + \| \, h[\psi,\xi] \, \|^*_{\infty,2\mu T}; \\ (c) & \| \, h[\psi,\xi] \, \|^*_{\infty,2\mu T} & \leq & \| \, \sigma\xi \, \|^*_{\infty,2\mu T} + \sum\limits_{|s|\leq 2\mu T} |\psi_s| \, \| \, \sigma\Delta^s\xi \, \|^*_{\infty,2\mu T} \\ & & & \text{[by definition of } h\text{]} \\ & & \leq & \sigma\Theta^T(\xi)\left[1 + \sum\limits_{|s|\leq 2\mu T} |\psi_s|\right] \\ & & & \text{[see (4.18)]} \\ & & = & \sigma\Theta^T(\xi)\left[1 + \| \, \psi \, \|_{1,2\mu T}\right] \\ & & \leq & \sigma\Theta^T(\xi)\left[1 + \| \, \psi \, \|^*_{1,2\mu T}\sqrt{4\mu T+1}\right] \\ & & & \text{[by (4.8)]} \\ & & \leq & 2^{2\mu+4}\mu\sigma\Theta^T(\xi) \\ & & & \text{[in view of the constraints in } (P_T[\cdot])\text{],} \end{array} \tag{4.23}$$

and (4.22) follows.

3^0. We claim that the optimal value P^*_ξ in $(P_T[y_f(\xi)])$ can be bounded from above as follows:

$$P^*_\xi \leq 2^{2\mu+3}\mu^{3/2}T^{1/2+k-1/p}\varepsilon + 2^{2\mu+4}\mu\sigma\Theta^T(\xi). \tag{4.24}$$

Indeed, let $\eta,\omega \in \mathcal{F}_{2\mu T}$ be given by Lemma 4.2.1. Applying (4.20) with $N = 2\mu T$, we conclude that ω is a feasible solution of $P_T[y_f(\xi)]$. In view of (4.22), to prove (4.24) it suffices to verify that

$$\| \, g^*[\omega,f] \, \|^*_{\infty,2\mu T} \leq 2^{2\mu+3}\mu^{3/2}T^{1/2+k-1/p}\varepsilon \tag{4.25}$$

which is given by the following computation. Let

$$\phi^j = \eta^j(\Delta)f^j, \; j = 1,...,l.$$

We have

$$\begin{array}{rcl}
\parallel g^*[\omega,f] \parallel^*_{\infty,2\mu T} & \leq & \sum\limits_{j=1}^{l} \parallel g^*[\omega,f^j] \parallel^*_{\infty,2\mu T} \\
& & \text{[since } g^*[\omega,\cdot] \text{ is linear in the second argument]} \\
& = & \sum\limits_{j=1}^{l} \parallel f^j + \omega(\Delta) f^j \parallel^*_{\infty,2\mu T} \\
& = & \sum\limits_{j=1}^{l} \parallel \eta(\Delta) f^j \parallel^*_{\infty,2\mu T} \\
& = & \sum\limits_{j=1}^{l} \parallel \rho^j(\Delta)\, [\eta^j(\Delta) f^j] \parallel^*_{\infty,2\mu T} \\
& & \text{[the origin of } \rho^j\text{, see Lemma 4.2.1.(ii)]} \\
& = & \sum\limits_{j=1}^{l} \parallel \rho^j \phi^j \parallel^*_{\infty,2\mu T} \\
& \leq & (4\mu T+1)^{1/2-1/p} \sum\limits_{j=1}^{l} \parallel \rho^j \phi^j \parallel_{p,2\mu T} \\
& & \text{[by (4.9) applied with } N=2\mu T\text{]} \\
& \leq & (4\mu T+1)^{1/2-1/p} \sum\limits_{j=1}^{l} \parallel \rho^j \parallel_1 \parallel \phi^j \parallel_{p,4\mu T} \\
& & \text{[by (4.7) and since } \deg(\rho^j) \leq 2\mu T\text{]} \\
& \leq & (4\mu T+1)^{1/2-1/p} \sum\limits_{j=1}^{l} (2\mu T+1) \parallel \rho^j \parallel_\infty \parallel \phi^j \parallel_{p,4\mu T} \\
& & \text{[since } \rho^j \in \mathcal{F}_{2\mu T}\text{]} \\
& \leq & (4\mu T+1)^{1/2-1/p} (2\mu T+1) 2^{2\mu} T^{k-1} \sum\limits_{j=1}^{l} \parallel \phi^j \parallel_{p,4\mu T} \\
& & \text{[by (4.21)]} \\
& \leq & 2^{2\mu+3} \mu^{3/2} T^{1/2+k-1/p} \varepsilon \\
& & \text{[see (4.16)],}
\end{array}$$

as required in (4.25).

4^0. We claim that

$$\parallel g^*[\omega,f] \parallel_\infty \leq 2^{2\mu+3} \mu T^{k-1/p} \varepsilon. \tag{4.26}$$

Indeed, similar to the preceding computation,

$$\begin{array}{rcl}
\| g^*[\omega, f] \|_\infty & \leq & \sum\limits_{j=1}^{l} \| g^*[\omega, f^j] \|_\infty \\
& = & \sum\limits_{j=1}^{l} \| f^j + \omega(\Delta) f^j \|_{\infty, 2\mu T} \\
& = & \sum\limits_{j=1}^{l} \| \eta(\Delta) f^j \|_{\infty, 2\mu T} \\
& = & \sum\limits_{j=1}^{l} \| \rho^j(\Delta) \, [\eta^j(\Delta) f^j] \|_{\infty, 2\mu T} \\
& = & \sum\limits_{j=1}^{l} \| \rho^j \phi^j \|_{\infty, 2\mu T} \\
& \leq & \sum\limits_{j=1}^{l} \| \rho^j \|_\infty \max\limits_{|s| \leq 2\mu T} \| \Delta^s \phi^j \|_{1, 2\mu T} \\
& & [\text{since } \deg(\rho^j) \leq 2\mu T] \\
& \leq & 2^{2\mu} T^{k-1} \sum\limits_{j=1}^{l} \max\limits_{|s| \leq 2\mu T} \| \Delta^s \phi^j \|_{1, 2\mu T} \\
& & [\text{by } (4.21)] \\
& \leq & 2^{2\mu} T^{k-1} (4\mu T + 1)^{1-1/p} \sum\limits_{j=1}^{l} \| \phi^j \|_{p, 4\mu T} \\
& & [\text{by Hölder inequality}] \\
& \leq & 2^{2\mu+3} \mu T^{k-1/p} \varepsilon \\
& & [\text{see } (4.16)],
\end{array}$$

as required.

5^0. Let us fix a realization ξ of the noises, and let $\widehat{\psi}$ be the corresponding optimal solution of (P_T). By (4.24) one has

$$\begin{array}{rcl}
2^{2\mu+3} \mu^{3/2} T^{1/2+k-1/p} \varepsilon + 2^{2\mu+4} \mu \sigma \Theta^T(\xi) & \geq & P^*_\xi \\
& = & \| g^*[\widehat{\psi}, f] + h[\widehat{\psi}, \xi] \|^*_{\infty, 2\mu T} \\
& & [\text{see } (4.23.a)] \\
& \geq & \| g^*[\widehat{\psi}, f] \|^*_{\infty, 2\mu T} - \| h[\widehat{\psi}, \xi] \|^*_{\infty, 2\mu T},
\end{array}$$

whence

$$\begin{array}{rcl}
\| g^*[\widehat{\psi}, f] \|^*_{\infty, 2\mu T} & \leq & A(\xi) \\
& \equiv & 2^{2\mu+3} \mu^{3/2} T^{1/2+k-1/p} \varepsilon + 2^{2\mu+4} \mu \sigma \Theta^T(\xi) + \| h[\widehat{\psi}, \xi] \|^*_{\infty, 2\mu T} \\
& \leq & 2^{2\mu+3} \mu^{3/2} T^{1/2+k-1/p} \varepsilon + 2^{2\mu+5} \mu \sigma \Theta^T(\xi) \\
& & [\text{see } (4.23.c)]
\end{array} \tag{4.27}$$

6^0. Let η, ω be the same as in 3^0–4^0, and let

$$\alpha = (1 + \widehat{\psi}(\Delta)) f.$$

Note that by the definition of g^* one has

$$g^*_t[\widehat{\psi}, f] = \alpha_t \quad \forall t : |t| \leq 2\mu T. \tag{4.28}$$

We claim that

$$|(\eta(\Delta)\alpha)_0| \leq 2^{4\mu+5} \mu^2 T^{k-1/p} \varepsilon. \tag{4.29}$$

Indeed, we have

$$\begin{array}{rrcl}
 & \eta(\Delta)\alpha & = & \eta(\Delta)(1+\widehat{\psi}(\Delta))f \\
 & & = & (1+\widehat{\psi}(\Delta))[\eta(\Delta)f] \\
\Rightarrow & \left(\eta(\Delta)\alpha\right)_0 & = & \left(\eta(\Delta)f\right)_0 + \sum\limits_{|s|\leq 2\mu T} \widehat{\psi}_s \left(\eta(\Delta)f\right)_{-s} \\
 & & = & g^*_0[\omega,f] + \sum\limits_{|s|\leq 2\mu T} \widehat{\psi}_s g^*_{-s}[\omega,f] \\
 & & & \text{[since } g^*_s[\omega,f] = \left(\eta(\Delta)f\right)_s,\ |s|\leq 2\mu T\text{]} \\
\Rightarrow & |\left(\eta(\Delta)\alpha\right)_0| & \leq & |g^*_0[\omega,f]| + \parallel \widehat{\psi} \parallel^*_{1,2\mu T}\parallel g^*[\omega,f] \parallel^*_{\infty,2\mu T} \\
 & & & \text{[by Parseval equality and since } |\langle u,v\rangle| \leq\parallel u \parallel_1\parallel v \parallel_\infty, \\
 & & & u,v\in \mathbf{C}(G_{2\mu T})\text{]} \\
 & & \leq & 2^{2\mu+3}\mu T^{k-1/p}\varepsilon + \parallel \widehat{\psi} \parallel^*_{1,2\mu T}\parallel g^*[\omega,f] \parallel^*_{\infty,2\mu T} \\
 & & & \text{[by (4.26)]} \\
 & & \leq & 2^{2\mu+3}\mu T^{k-1/p}\varepsilon + 2^{2\mu+2}\mu^{1/2}T^{-1/2}2^{2\mu+3}\mu^{3/2}T^{1/2+k-1/p}\varepsilon \\
 & & & \text{[since } \psi \text{ is feasible for } (P_T) \text{ and by (4.25)],}
\end{array}$$

as claimed.

7^0. Now – the concluding step. Setting $\widehat{f} = \widehat{f}[T,y]$, we have

$$\begin{array}{rcl}
f_0 - \widehat{f} & = & f(0) + \left(\widehat{\psi}(\Delta)y\right)_0 \\
 & & \text{[the construction of the estimate]} \\
 & = & \left((1+\widehat{\psi}(\Delta))f\right)_0 + \sigma\left(\widehat{\psi}(\Delta)\xi\right)_0 \\
 & = & \alpha_0 + \sigma\left(\widehat{\psi}(\Delta)\xi\right)_0 \\
 & & \text{[the definition of } \alpha\text{]} \\
\Rightarrow & & \\
|f_0 - \widehat{f}| & \leq & |\alpha_0| + \sigma\left|\sum\limits_{|s|\leq 2T} \widehat{\psi}_s \xi_{-s}\right| \\
 & \leq & |\alpha_0| + \sigma \parallel \widehat{\psi} \parallel^*_{1,2\mu T}\parallel \xi \parallel^*_{\infty,2\mu T} \\
 & & \text{[same as in the previous computation]} \\
 & \leq & |\alpha_0| + 2^{2\mu+2}\mu^{1/2}T^{-1/2}\sigma\Theta^T(\xi) \\
 & & \text{[since } \widehat{\psi} \text{ is feasible for } (P_T) \text{ and by definition of } \Theta^T\text{];}
\end{array}$$

Thus,

$$|f_0 - \widehat{f}| \leq |\alpha_0| + 2^{2\mu+2}\mu^{1/2}\sigma T^{-1/2}\Theta^T(\xi). \tag{4.30}$$

It remains to bound $|\alpha_0|$. We have

$$\begin{array}{rcl}
\alpha_0 & = & (\eta(\Delta)\alpha)_0 - (\omega(\Delta)\alpha)_0 \Rightarrow \\
|\alpha_0| & \leq & |(\eta(\Delta)\alpha)_0| + |(\omega(\Delta)\alpha)_0| \\
& \leq & 2^{4\mu+5}\mu^2 T^{k-1/p}\varepsilon + |(\omega(\Delta)\alpha)_0| \\
& & \text{[see (4.29)]} \\
& \leq & 2^{4\mu+5}\mu^2 T^{k-1/p}\varepsilon + \parallel \omega \parallel_{1,2\mu T}^* \parallel \alpha \parallel_{\infty,2\mu T}^* \\
& & \text{[as in the previous two computations]} \\
& = & 2^{4\mu+5}\mu^2 T^{k-1/p}\varepsilon + 2^{2\mu+2}\mu^{1/2}T^{-1/2} \parallel \alpha \parallel_{\infty,2\mu T}^* \\
& & \text{[by (4.20) applied with } N = 2\mu T\text{]} \\
& \leq & 2^{4\mu+5}\mu^2 T^{k-1/p}\varepsilon + 2^{2\mu+2}\mu^{1/2}T^{-1/2} \parallel g^*[\widehat{\psi}, f] \parallel_{\infty,2\mu T}^* \\
& & \text{[by (4.28)]} \\
& \leq & 2^{4\mu+5}\mu^2 T^{k-1/p}\varepsilon \\
& & +2^{2\mu+2}\mu^{1/2}T^{-1/2}\left[2^{2\mu+3}\mu^{3/2}T^{1/2+k-1/p}\varepsilon + 2^{2\mu+5}\mu\sigma\Theta^T(\xi)\right] \\
& & \text{[by (4.27)]}
\end{array}$$

Thus,

$$|\alpha_0| \leq 2^{4\mu+6}\mu^2 T^{k-1/p}\varepsilon + 2^{4\mu+7}\mu^{3/2}\sigma T^{-1/2}\Theta^T(\xi).$$

Combining this inequality with (4.30), we come to (4.17). ■

4.2.3 Discussion

Theorem 4.2.1 has a number of important consequences already in the "parametric case" – when the signal f we observe according to (4.11) satisfies a homogeneous finite difference equation with constant coefficients:

$$\eta(\Delta)f \equiv 0, \tag{4.31}$$

η being normalized polynomial of degree $\leq \mu$. In the notation of Theorem 4.2.1, this is in the case when $l = 1$, $k \leq \mu$ and $\varepsilon = 0$.

A) In the case of (4.31) relation (4.17) becomes

$$|f_0 - \widehat{f}[T, y_f(\xi)]| \leq C\sigma T^{-1/2}\Theta^T(\xi). \tag{4.32}$$

Due to the origin of Θ^T, we have

$$\left(\mathcal{E}\left\{(\Theta^T(\xi))^2\right\}\right)^{1/2} \leq O(1)\sqrt{\ln T},$$

so that

$$\left(\mathcal{E}\left\{|f_0 - \widehat{f}[T, y_f(\xi)]|^2\right\}\right)^{1/2} \leq O_\mu(1)\sigma\sqrt{\frac{\ln T}{T}}. \tag{4.33}$$

We see that

> (!) *For every T, it is possible to recover an entry f_t in a sequence f satisfying unknown homogeneous difference equation with constant coefficients of a given order μ via $O_\mu(1)T$ noisy observations of the entries of the sequence around the instant t with "nearly parametric risk" $O_\mu(1)\sigma\sqrt{\frac{\ln T}{T}}$.*

It should be stressed that the result is uniform with respect to *all* solutions of *all* difference equations of a given order, which is rather surprising. Note that if the equation were known in advance, the quality of recovering f_t could be slightly improved – we could get rid of the $\sqrt{\ln T}$-factor, thus coming to the result completely similar to the case of recovering algebraic polynomials of order $\mu - 1$ (their restrictions on an equidistant observation grid are the solutions of a particular finite difference equation of order μ, namely, $(1-\Delta)^\mu f = 0$).

In the case when the equation is unknown, the logarithmic factor turns out to be unavoidable: it is proved in [22] that when the signal to be recovered is known to be a harmonic oscillation $f_t = c\sin(\omega t + \phi)$, the uniform, with respect to all values of c, ω, ϕ, risk of an arbitrary estimate of f_0 via $2T+1$ observations (4.11) around the time instant $t = 0$ is at least $O(1)\sigma\sqrt{\frac{\ln T}{T}}$. Note that the problem of recovering a harmonic oscillation $c\sin(\omega t + \phi)$ is a *parametric* problem; indeed, all we need is to recover the triple of parameters c, ω, ϕ. As we see, the minimax risk associated with this parametric estimation problem is *not* the parametric risk $O(T^{-1/2})$.

B) The estimate we have built solves an "interpolation" problem – it recovers f_0 via observations "placed symmetrically" around the time instant $t = 0$ we are interested in. In some applications we should solve the "forecast" problem – we would like to estimate f_0 via a given number of observations (4.11) placed at least τ units of time before the instant $t = 0$, i.e., via the observations $y_{-\tau-4\mu T}, y_{-\tau-4\mu T+1}, ..., y_{-\tau}$. What can be done in this situation?

Slight modification of the construction we have presented demonstrates the following:

> (!!) *In addition to the premise of Theorem 4.2.1, assume that every finite-difference equation*
>
> $$\eta^j(\Delta)h = 0$$
>
> *is "quasi-stable": every solution of this equation grows with t no faster than an algebraic polynomial (equivalently: all roots of the polynomial $\eta^j(z)$ are ≥ 1 in absolute value). Then the result of the theorem is valid for a properly chosen "forecast" estimate, namely, for the estimate*
>
> $$\widehat{f}^+[T,y] = -\sum_{s=T}^{4\mu T} \widehat{\psi}_s y_{-s},$$
>
> *where $\widehat{\psi}$ is an optimal solution to the optimization program*
>
> $$\begin{array}{rl} & \| \Delta^{4\mu T}(I + \psi(\Delta))y \|^*_{\infty,4\mu T} \to \min \\ s.t. & \\ & \begin{array}{rcl} \psi & \in & \mathcal{F}_{4\mu T}; \\ \psi_s & = & 0, -4\mu T \leq s < T; \\ \| \psi \|^*_{1,4\mu T} & \leq & B(\mu)T^{-1/2} \end{array} \end{array}$$
>
> *with properly chosen $B(m)$.*

As a consequence, given N subsequent noisy observations (4.11) of a solution to an *unknown quasi-stable* homogeneous difference equation of order $\leq \mu$, we may predict the value of the solution $O(N/\mu)$ units of time forward, with the worst-case, over *all* solutions of *all* quasi-stable equations of order $\leq \mu$, risk not exceeding $O_\mu(1)\sigma\sqrt{\frac{\ln N}{N}}$.

Note that the assumption of quasi-stability of the underlying finite-difference equation is crucial in the forecast problem. E.g., given *all* observations y_t, $t < 0$, of a solution to a *known* (unstable) equation

$$f_{t+1} - 2f_t = 0,$$

you cannot say definitely what is the solution at 0 (provided, of course, that $\sigma > 0$).

4.3 From sequences to functions

The main step in passing from estimating sequences "nearly satisfying" homogeneous difference equations to estimating functions satisfying differential inequalities is given by the following simple

Lemma 4.3.1 *Let*

- n, k, $\mu \geq k$ *and* T *be positive integers;*

- $g : (-\infty, \infty) \to \mathbf{C}$ *be a* $k-1$ *times continuously differentiable function with absolute continuous* $g^{(k-1)}$*;*

- $g^n \in \mathcal{F}$ *be the restriction of* g *on the grid* $\Gamma^n = \{t/n\}_{t=-\infty}^{\infty}$*:*

$$g^n_t = g(t/n)\ t \in \mathbf{Z};$$

- $q(z) = z^k + q_1 z^{k-1} + ... + q_k$ *be a polynomial of degree* k *with unit leading coefficient;*

- B *be a segment centered at the origin and containing at least* $8\mu T + 2k + 1$ *points of the grid* Γ^n*;*

- $p \in [1, \infty]$.

There exists a normalized polynomial $\theta(z)$ *of degree* k *such that*

$$\| \theta(\Delta)g^n \|_{p,4\mu T} \leq O_k(1)n^{-k+1/p} \| q\left(\frac{d}{dx}\right) g \|_{p,B} . \tag{4.34}$$

The proof is placed in Section 4.4.

Combining Lemma 4.3.1 with Theorem 4.2.1, we can extend – in a quite straightforward manner – basically all estimation techniques we have considered so far to the case of functions satisfying unknown differential inequalities. We shall focus on "the best" – the spatial adaptive – estimate.

4.3.1 Spatial adaptive estimate: preliminaries

The recovering routine we are about to build, same as the spatial adaptive estimate from Chapter 3, is specified by a single parameter – its *order* μ which should be a positive real. Let observations (4.3) be given, and let $x = t/n$ be a point from the observation grid. For every positive integer T such that the grid contains $8\mu T + 1$ observations around x – i.e., such that $0 < t - 4\mu T$, $t + 4\mu T \leq n$ – we have built in the previous Section an estimate $\widehat{f}^T(x; y)$ of $f(x)$ via the segment of observations $\{y_{t-4\mu T}, y_{t-4\mu T+1}, ..., y_{t+4\mu T}\}$. Let us associate with the estimate $\widehat{f}^T(x; y)$ its *window*

$$B_T(x) = [x - (4T + 2)\mu n^{-1}, x + (4T + 2)\mu n^{-1}].$$

From Theorem 4.2.1 and Lemma 4.3.1 we know that

> **(*)** *Let f be the function underlying observations* (4.3), *$x = t/n$ be a point from the observation grid Γ_n, and let $T \geq 1$ be such that the window $B_T(x)$ is contained in* $[0, 1]$.
>
> (i) *For every collection $\mathcal{U}$ comprised of*
>
> - *positive integers k, l with $kl \leq \mu$;*
> - *l polynomials η^j, $j = 1, ..., l$, normalized of degree k each;*
> - *a decomposition*
>
> $$f(u) = \sum_{j=1}^{l} f^j(u), \quad u \in B_T(x);$$
>
> - *$p \in [1, \infty]$*
>
> *the error of the estimate $\widehat{f}^T(x; y)$ can be bounded from above as*
>
> $$\begin{array}{rcl} |\widehat{f}^T(x; y_f(\xi)) - f(x)| & \leq & C_1(\mu)\left[\varepsilon(T, \mathcal{U}) + \sigma T^{-1/2}\Theta_n(\xi)\right], \\ \varepsilon(T, \mathcal{U}) & = & T^{k-1/p} \sum\limits_{j=1}^{l} \parallel \eta^j(\Delta)\widetilde{f}^j \parallel_{p,4\mu T}, \\ \widetilde{f}^j_s & = & f^j\left(\frac{s-t}{n}\right), \end{array} \qquad (4.35)$$
>
> *where $\Theta_n(\xi)$ is the maximum of the $\parallel \cdot \parallel_\infty$-norms of discrete Fourier transforms of all segments, of odd cardinality, of the sequence $\{\xi_s\}_{s=1}^n$ (so that Θ_n is the maximum of norms of $\leq n^2$ standard Gaussian 2D vectors with zero mean and unit covariance matrix).*
>
> (ii) *Let l, k be positive integers with $kl \leq \mu$, let $p \in [1, \infty]$ and let $f \in \mathcal{W}^{l,k,p}(B_T(x); A)$ for some A. Then there exists a collection $\mathcal{U}$ of the type described in* (i) *such that*
>
> $$\varepsilon(T, \mathcal{U}) \leq C_2(\mu)(T/n)^{k-1/p}A \leq C_3(\mu)D^{k-1/p}(B_T(x))A; \qquad (4.36)$$
>
> *here, as always $D(B)$ is the length of a segment B.*

Combining (*.i) and (*.ii) and observing that $T^{-1/2}$, up to a factor depending on μ only, is the same as $\frac{1}{\sqrt{nD(B_T(x))}}$, we come to the conclusion as follows:

(**) *Given a positive integer μ, a function f and a segment $B \in [0,1]$, let us set*

$$\begin{array}{rcl}\Phi_\mu(f,B) & = & \inf\{D^{k-1/p}(B)A \mid p \in [1,\infty]; k,l \in \mathbf{N}, kl \le \mu; \\ & & A \ge 0, f \in \mathcal{W}^{l,k,p}(B;A)\}.\end{array}$$

Then for every point $x = t/n$ from the observation grid Γ_n and every integer $T \ge 1$ such that $B_T(x) \subset [0,1]$ one has

$$|\widehat{f}^T(x;y_f(\xi)) - f(x)| \le C(\mu)\left[\Phi_f(x,B_T(x)) + \frac{\sigma}{\sqrt{nD(B_T(x))}}\Theta_n(\xi)\right]. \tag{4.37}$$

Besides this,

$$\forall w \ge 1: \quad \mathrm{Prob}\left\{\Theta_n \ge O_\mu(1)w\sqrt{\ln n}\right\} \le \exp\left\{\frac{-w^2\ln n}{2}\right\}. \tag{4.38}$$

Finally, from the definitions of the classes $\mathcal{W}$ and the quantity Φ_μ it immediately follows that

If $f \in \mathcal{W}^{l,k,p}(B;L)$ with $lk \le \mu$, then there exists a function $\widetilde{f}: B \to \mathbf{R}_+$ such that

$$\begin{array}{rrcl} & \|\widetilde{f}\|_{p,B} & \le & L; \\ \forall B' \subset B: & \Phi_\mu(f,B') & \le & D^{k-1/p}(B')\,\|\widetilde{f}\|_{p,B'}\end{array} \tag{4.39}$$

Note that (4.37), (4.38), (4.39) are completely similar to the basic relations (3.5), (3.17), (3.6), respectively, underlying all developments of Chapter 3.

4.3.2 Spatial adaptive estimate: construction and quality

The construction. Let us choose $\omega = \omega(\mu)$ so large that

$$\mathrm{Prob}\left\{\Theta_n > \omega\sqrt{\ln n}\right\} \le n^{-4\mu} \tag{4.40}$$

(cf. (3.8)).

The adaptive estimate $\widehat{f}_n(x;y)$ of the value $f(x)$ at a point $x \in \Gamma_n$ is as follows (cf. Section 3.2). Let us say that a positive integer T is *admissible* for x, if the segment $B_T(x)$ is contained in $[0,1]$. Assume that x admits admissible T's, i.e., that

$$6\mu n^{-1} < x < 1 - 6\mu n^{-1}. \tag{4.41}$$

We already have associated with every T admissible for x certain estimate $\widehat{f}^T(x;\cdot)$ of $f(x)$ via observations (4.3). Given these observations y, let us call a positive integer T x *normal* for x (cf. (3.13)), if it is admissible for x and

$$|\widehat{f}^{T'}(x;y) - \widehat{f}^T(x;y)| \le 4C(\mu)\frac{\sigma\omega\sqrt{\ln n}}{\sqrt{nD(B_{T'}(x))}} \quad \forall T',\ 1 \le T' \le T,$$

$C(\mu)$ being the constant from (4.37). Normal for x values of T clearly exist (e.g., $T = 1$); let $T(x;y)$ be the largest of these values; note that this indeed is a well-defined deterministic function of x,y. Our *order μ adaptive estimate* of $f(x)$, by construction, is

$$\widehat{f}_n(x;y) = \widehat{f}^{T(x;y)}(x;y). \tag{4.42}$$

The quality of our adaptive estimate $\widehat{f}_n$ is given by the following

Theorem 4.3.1 *Let $\gamma \in (0,1)$, let μ be a positive integer, and let $\mathcal{W} = \mathcal{W}^{l,k,p}(B;L)$, where $kl \le \mu$ and $pk > 1$. For properly chosen $P \ge 1$ depending solely on μ, p, γ and nonincreasing in p the following statement takes place:*

If the volume n of observations (4.3) is large enough, namely,

$$\begin{gathered} P^{-1} n^{\frac{2k-2\pi+1}{2}} \ge \frac{L}{\widehat{\sigma}_n} \ge P D^{-\frac{2k-2\pi+1}{2}}(B) \\ \left[\widehat{\sigma}_n = \sigma\sqrt{\frac{\ln n}{n}}, \quad \pi = \frac{1}{p}\right] \end{gathered} \tag{4.43}$$

($D(B)$ is the length of segment B), then for every $q \in [1,\infty]$ the worst case, with respect to $\mathcal{W}$, discrete q-risk of the order μ adaptive estimate $\widehat{f}_n(\cdot;\cdot)$ can be bounded as follows (cf. (3.16)):

$$\begin{array}{rcl} \widetilde{\mathcal{R}}_q\left(\widehat{f}_n;\mathcal{W}\right) & \equiv & \sup_{f\in\mathcal{W}}\left(\mathcal{E}\left\{|\widehat{f}(\cdot;y_f(\xi)) - f(\cdot)|^2_{q,B_\gamma}\right\}\right)^{1/2} \\ & \le & PL\left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,q)} D^{\lambda(p,k,q)}(B), \\ \beta(p,k,q) & = & \begin{cases} \frac{k}{2k+1}, & \theta \ge \pi\frac{1}{2k+1} \\ \frac{k+\theta-\pi}{2k-2\pi+1}, & \theta \le \pi\frac{1}{2k+1} \end{cases}, \\ \theta & = & \frac{1}{q}, \\ \lambda(p,k,q) & = & \begin{cases} \theta - \frac{\pi}{2k+1}, & \theta \ge \pi\frac{1}{2k+1} \\ 0, & \theta \le \pi\frac{1}{2k+1} \end{cases}; \end{array} \tag{4.44}$$

here B_γ is the concentric to B γ times smaller segment and

$$|g|_{q,B} = \left(\frac{1}{n}\sum_{x\in\Gamma_n\cap B}|g(x)|^q\right)^{1/q}.$$

Proof of the theorem repeats word by word the proof of Theorem 3.3.1, with (4.37), (4.38), (4.39) playing the role of (3.5), (3.17), (3.6), respectively.

Optimality issues. The upper bounds on risks given by Theorem 4.3.1 are exactly the univariate ($d = 1$) versions of bounds from Theorem 3.3.1. Since now we are working with wider families of functions wider than local Sobolev balls, all results of Chapter 3 (see Section 3.4) on the non-optimality index of the adaptive estimate remain valid for our new estimate as considered on the nested family of collections of regression functions (cf. Theorem 3.4.2)

$$\mathbf{W} \equiv \mathbf{W}^{p,\mu} = \{\mathbf{W}_n\}_{n=1}^{\infty}$$

($p \in (1,\infty]$, $\mu \in \mathbf{N}$) defined as follows:

$$\begin{gathered} \mathbf{W}_n = \left\{ \mathcal{W}^{l,k,p'}(B;L) \,\middle|\, \begin{array}{lrcl} (a) & p' & \ge & p, \\ (b) & 1 \le kl & \le & \mu, \\ (c) & P^{-1}n^{\frac{2-2\pi+1}{2}} \ge \frac{L}{\widehat{\sigma}_n} & \ge & PD^{-\frac{2\mu-2\pi+1}{2}}(B), \end{array} \right\} \\ \left[\begin{array}{c} P \text{ is given by Theorem 4.3.1, } \pi = \frac{1}{p} \\ \widehat{\sigma}_n = \frac{\sigma\sqrt{\ln n}}{\sqrt{n}} \end{array} \right] \end{gathered} \tag{4.45}$$

the non-optimality index of our estimate on $\mathbf{W}$ does not exceed

$$\Phi(n) = (\ln n)^{\frac{\mu}{2\mu+1}}.\quad ^{2)}$$

4.3.3 "Frequency modulated signals"

A function f from class $\mathcal{W}^{l,k,p}([0,1];L)$ is a sum of l functions f^j satisfying each its own differential inequality of order k on the entire segment $[0,1]$. What happens if we "localize" this property, allowing the decomposition to vary from point to point? The precise definition is as follows:

Definition 4.3.1 *Let us fix positive integers k, l and reals $p \in [1,\infty]$, $L > 0$, $d \in (0,1/6]$. We say that a function $f : [0,1] \to \mathbf{C}$ belongs to the class $\mathcal{A}^{l,k,p,d}(L)$, if there exists a function $L_f(x) \in L_p[d, 1-d]$ such that*

$$\| L_f \|_{p,[d,1-d]} \leq L \tag{4.46}$$

and

$$\forall x \in [d, 1-d] : \quad f \in \mathcal{W}^{l,k,p}([x-d, x+d]; (2d)^{1/p} L_f(x)). \tag{4.47}$$

Note that the classes $\mathcal{A}$ extend our previous classes $\mathcal{W}$:

$$\begin{array}{l} \forall d \in (0,1/6] : \\ \mathcal{W}^{l,k,p}([0,1];L) \subset \mathcal{A}^{l,k,p,d}(L) \end{array} \tag{4.48}$$

Indeed, let $f \in \mathcal{W}^{l,k,p}([0,1];L)$, let $f = \sum\limits_{j=1}^{l} f^j$ be the corresponding decomposition, and let $q^j(z) = z^k + q_1^j z^{k-1} + ... + q_k^j$ be the associated polynomials:

$$\sum_{j=1}^{l} \| q^j\left(\frac{d}{dx}\right) f^j \|_p \leq L.$$

Let us set

$$L_f(x) = (2d)^{-1/p} \sum_{j=1}^{l} \| q^j\left(\frac{d}{dx}\right) f^j \|_{p,[x-d,x+d]}$$

and let us verify that this choice fits (4.46), (4.47). The latter relation is evident, while the former one is given by the following computation: setting

$$L(\cdot) = \sum_{j=1}^{l} \left| q^j\left(\frac{d}{dx}\right) f^j \right|,$$

[2] Formally, the announced statement is *not* a straightforward corollary of the lower bounds on the minimax risk established in Chapter 3: there we were dealing with the usual q-norms of the estimation errors, while now we are speaking about discrete versions of these norms. However, looking at the proofs of the lower bounds, one can observe that they remain valid for the discrete versions of q-risks as well.

Figure 4.3: A "frequency modulated" signal.

and assuming $p < \infty$, we have

$$\begin{array}{rcl} \| L_f \|^p_{p,[d,1-d]} & = & (2d)^{-1} \int\limits_{d}^{1-d} \left\{ \int\limits_{x-d}^{x+d} L^p(u)du \right\} dx \\ & = & (2d)^{-1} \int\limits_{0}^{1} \left\{ \int\limits_{\max[0,u-d]}^{\min[1,u+d]} dx \right\} L^p(u)du \\ & \leq & \| L(\cdot) \|^p_{p,[0,1]} \Rightarrow \\ \| L_f \|_{p,[d,1-d]} & \leq & L, \end{array}$$

as required in (4.46). We have established the latter relation in the case of $p < \infty$; by continuity, it is valid in the case $p = \infty$ as well.

Our interest in classes $\mathcal{A}$ comes from the fact that they contained not only "amplitude modulated", but also "frequency modulated" signals. Consider, e.g., the following construction. Given a positive integer N, we partition the segment $[0, 1]$ into N non-overlapping segments B_t, $t = 1, ..., N$, of the length $2d = \frac{1}{N}$ each; let $x_t = 2td$, $t = 1, ..., N$, be the right endpoints of these segments. Now let $g \in \mathcal{S}^{k,p}([0, 1]; L)$ be an "amplitude" which is supposed to vanish, along with its derivatives of order $< k$, at all points x_t, $t = 1, ..., N$. Consider the family of functions obtained from g by "frequency modulation": A function f from the family on every segment B_t is of the form

$$g(x)\sin(\omega_t x + \phi_t)$$

with somehow chosen frequency ω_t and phase ϕ_t. One can immediately verify that all functions from this family belong to $\mathcal{A}^{4,k,p,d}(4L)$.

It turns out that the quality of our adaptive estimate on classes $\mathcal{A}$ is basically the same as on narrower classes $\mathcal{W}$:

Theorem 4.3.2 *Let $\gamma \in (0, 1)$, let μ be a positive integer, and let $\mathcal{A} = \mathcal{A}^{l,k,p,d}(L)$, where $kl \leq \mu$ and $pk > 1$. Assume that the volume of observations n is large enough (the critical value depends on $\mu, p, L/\sigma$ only and is independent of d), and that d is* *not too small, namely,*

$$d \geq \left(\frac{\widehat{\sigma}_n}{L}\right)^{\frac{2}{2k+1}}, \qquad \widehat{\sigma}_n = \frac{\sigma\sqrt{\ln n}}{\sqrt{n}}. \tag{4.49}$$

Then, for every $q \in [1, \infty]$, the $|\cdot|_{q,[0,1]_\gamma}$-risk of the order μ adaptive estimate $\widehat{f}_n$ on the class $\mathcal{A}$ can be bounded as follows:

$$\begin{array}{rcl}
\widetilde{\mathcal{R}}_q\left(\widehat{f}_n; \mathcal{A}\right) & \equiv & \sup_{f\in\mathcal{A}}\left(\mathcal{E}\left\{|\widehat{f}(\cdot; y_f(\xi)) - f(\cdot)|^2_{q,[0,1]_\gamma}\right\}\right)^{1/2} \\
& \leq & PL\left(\frac{\widehat{\sigma}_n}{L}\right)^{2\beta(p,k,q)}, \\
\beta(p,k,q) & = & \left\{\begin{array}{ll} \frac{k}{2k+1}, & \theta \geq \pi\frac{1}{2k+1} \\ \frac{k+\theta-\pi}{2k-2\pi+1}, & \theta \leq \pi\frac{1}{2k+1} \end{array}\right., \\
\pi & = & \frac{1}{p}, \\
\theta & = & \frac{1}{q}.
\end{array} \tag{4.50}$$

Here $[0,1]_\gamma = [0.5(1-\gamma), 1 - 0.5(1-\gamma)]$ is the γ-shrinkage of the segment $[0,1]$ to its center and P depends on μ, p, γ only and is nonincreasing in p.

If (4.49) *is equality rather than inequality, then, for all large enough values of n, the upper bound* (4.50) *coincides (up to a factor depending on μ, p, γ only) with the minimax $|\cdot|_{q,[0,1]_\gamma}$-risk of estimating functions from $\mathcal{A}$ via observations* (4.3).

For proof, see [10].

4.4 Appendix: Proofs of Lemmas 4.2.1, 4.3.1

4.4.1 Proof of Lemma 4.2.1

1^0. To simplify notation, let us assume that every polynomial η^j is of degree k (the modifications in the case of $\deg \eta^j < k$ are quite straightforward), and let $\lambda_{j\ell}$, $\ell = 1, ..., k$, be the roots of the polynomial η_j (taken with their multiplicities). For every j, let $\mathcal{L}_j$ be the set of those ℓ for which $\lambda_{j\ell}$ are ≥ 1 in absolute value, and let $\mathcal{S}_j$ be the set of the remaining indices from $[\overline{1, k}]$.

2^0. Let us fix $j \leq l$, and let

$$\nu^j(z) = \left(\prod_{\ell\in\mathcal{L}_j}(1 - z/\lambda_{j\ell})\right)\left(\prod_{\ell\in\mathcal{S}_j}(z - \lambda_{j\ell})\right). \tag{4.51}$$

Then

$$\eta^j(z) = c_j\nu^j(z). \tag{4.52}$$

We claim that

$$|c_j| \geq 2^{-k}. \tag{4.53}$$

Indeed, it is clear that the maximum of absolute values of the coefficients of ν^j is not greater than the one of the polynomial $(z+1)^k$, i.e., is $\leq 2^k$; since the product $c_j\pi^j(z)$ is a normalized polynomial (i.e., with the maximum of modulae of coefficients equal to 1), the factor c_j must satisfy (4.53).

3^0. Mini-lemma. *Let $\lambda \in \mathbf{C}$. Then there exists a polynomial $\pi_T^\lambda(z)$ of degree $2T$ such that*

(i) *If $|\lambda| \geq 1$, then*

$$\begin{array}{lrcl} & 1+\pi_T^\lambda(z) & = & (1-z/\lambda)r_T^\lambda(z) \\ \textit{with} & & & \\ (a) & r_T^\lambda & \in & \mathcal{F}_{2T}, \\ (b) & \| r_T^\lambda \|_1 & \leq & 2T, \\ (c) & \| r_T^\lambda \|_\infty & \leq & 2; \end{array} \tag{4.54}$$

If $|\lambda| < 1$, then

$$\begin{array}{lrcl} & z^{2T}+\pi_T^\lambda(z) & = & (z-\lambda)r_T^\lambda(z) \\ \textit{with} & & & \\ (a) & r_T^\lambda & \in & \mathcal{F}_{2T}, \\ (b) & \| r_T^\lambda \|_1 & \leq & 2T, \\ (c) & \| r_T^\lambda \|_\infty & \leq & 2. \end{array} \tag{4.55}$$

(ii) *One has*

$$\forall N \geq 2T: \quad \| \pi_T^\lambda \|_{1,N}^* \leq \frac{\sqrt{2N+1}}{T}. \tag{4.56}$$

and

$$\| \pi_T^\lambda \|_1 \leq 1. \tag{4.57}$$

Indeed, let us set

$$\begin{array}{rcl} \psi(z) & = & \begin{cases} \frac{1}{T}\sum\limits_{t=1}^{T}(z/\lambda)^t, & \text{if } |\lambda| \geq 1 \\ \frac{1}{T}\sum\limits_{t=0}^{T-1} z^t\lambda^{T-t}, & \text{otherwise} \end{cases}, \\ \pi_T^\lambda(z) & = & -\psi^2(z). \end{array}$$

Note that in the case of $|\lambda| \geq 1$ we have

$$\begin{array}{rcl} 1+\pi_T^\lambda(z) & = & \left(T^{-1}\sum\limits_{t=1}^{T}[1-(z/\lambda)^t]\right)\left(1+T^{-1}\sum\limits_{t=1}^{T}(z/\lambda)^t\right) \\ & = & (1-z/\lambda)\underbrace{\left(T^{-1}\sum_{t=1}^{T}\sum_{\tau=0}^{t-1}(z/\lambda)^\tau\right)}_{q_1(z)}\underbrace{\left(1+T^{-1}\sum_{t=1}^{T}(z/\lambda)^t\right)}_{q_2(z)} \\ & \equiv & (1-z/\lambda)r_T^\lambda(z) \end{array}$$

and

$$\begin{array}{rcll} r_T^\lambda & \in & \mathcal{F}_{2T}, & \\ \| r_T^\lambda \|_1 & \leq & \| q_1 \|_1 \| q_2 \|_1 & [\text{by (4.8) with } p=1] \\ & \leq & T \times 2 & [\text{since } |\lambda| \geq 1] \\ & = & 2T, & \\ \| r_T^\lambda \|_\infty & \leq & \| q_1 \|_\infty \| q_2 \|_1 & [\text{by (4.8) with } p=\infty] \\ & \leq & 2 & [\text{since } |\lambda| \geq 1]. \end{array}$$

as required in (4.54). Completely similar computation demonstrates (4.55) in the case of $|\lambda| < 1$.

Now, by construction, $\| \pi_T^\lambda \|_1 \leq \| \psi \|_1^2 = 1$, as required in (4.57). To check (4.56), note that for $N \geq 2T$ and $\zeta \in G_N$ we have

$$\begin{array}{rrcll}
 & |(F_N\pi_T^\lambda)(\zeta)| & = & \frac{1}{\sqrt{2N+1}}|\pi_T^\lambda(\zeta)| & \\
 & & = & \frac{1}{\sqrt{2N+1}}|\psi(\zeta)|^2 & \\
 & & = & \sqrt{2N+1}\left[\frac{1}{\sqrt{2N+1}}|\psi(\zeta)|\right]^2 & \\
 & & = & \sqrt{2N+1}|(F_N\psi)(\zeta)|^2 & \\
\Rightarrow & \| \pi_T^\lambda \|_{1,N}^* & = & \sqrt{2N+1} \| F_N\psi \|_2^2 & \\
 & & = & \sqrt{2N+1} \| \psi \|_2^2 & [\text{by (4.6) and in view of } \psi \in \mathcal{F}_T] \\
 & & = & \frac{\sqrt{2N+1}}{T} & [\text{by construction of } \psi]
\end{array}$$

□

4^0. Now let us set

$$\begin{array}{rcl}
\pi_+^j(z) & = & \prod\limits_{s\in\mathcal{L}_j}(1+\pi_T^{\lambda_{js}}(z)), \\
\pi_-^j(z) & = & \prod\limits_{s\in\mathcal{S}_j}(z^{2T}+\pi_T^{\lambda_{js}}(z)), \\
\pi^j(z) & = & \pi_+^j(z)\pi_-^j(z), \\
\pi(z) & = & \prod\limits_{j=1}^{l}\pi^j(z), \\
\eta(z) & = & z^{-2TM}\pi(z),
\end{array}$$

where M is the sum, over $j \leq l$, of the cardinalities of the sets $\mathcal{S}_j$. Let us verify that η meets all requirements of Lemma 4.2.1.

4^0.1) By construction, we have $\deg(\pi_+^j(z)\pi_-^j(z)) \leq 2Tk$, whence $\pi \in \mathcal{F}_{2lkT}$; since $M \leq lk$, we have $2TM \leq 2lkT$ as well. Since $\pi(z)$ is a polynomial, we have $\eta \in \mathcal{F}_{2klT} \subset \mathcal{F}_{2\mu T}$, as required.

4^0.2) By (4.57) we have $\| \pi_T^\lambda \|_1 \leq 1$, whence

$$\| \pi^j \|_1 \leq 2^k, \tag{4.58}$$

so that $\| \eta \|_1 = \| \pi \|_1 \leq (2^k)^l = 2^{kl} \leq 2^\mu$, as required in (4.19).

4^0.3) Let $N \geq 2\mu T$. Let us fix $j \leq l$, and let $M_j = \mathrm{Card}(\mathcal{S}_j)$. By construction, we have

$$\pi^j(z)z^{-2TM_j} = \prod_{s=1}^{k}(1+\theta_s^j(z)), \tag{4.59}$$

where (see Mini-lemma) $\theta_s^j \in \mathcal{F}_{2T}$ are of $\| \cdot \|_1$-norms not exceeding 1 and of $\| \cdot \|_{1,N}^*$ norms not exceeding $\gamma \equiv \frac{\sqrt{2N+1}}{T}$. By (4.7) applied with $p = 1$ and (4.10), every nonempty product of a number of $\theta_s^j(z)$ with distinct values of the index s is of $\| \cdot \|_1$-norm not exceeding 1 and of $\| \cdot \|_{1,N}^*$-norm not exceeding γ. When opening parentheses in the right hand side of (4.59), we get the sum of 1 and $2^k - 1$ products of the just outlined type; consequently,

$$\theta^j(z) \equiv \pi^j(z)z^{-2TM_j} - 1 \in \mathcal{F}_{2kT}$$

is of $\| \cdot \|_1$-norm not exceeding 2^k and of $\| \cdot \|_{1,N}^*$-norm not exceeding $2^k\gamma$. Observing that

$$\eta(z) = \prod_{j=1}^{l}(1+\theta^j(z))$$

and repeating the reasoning we just have used, we conclude that $\omega(z) = \eta(z) - 1$ is of $\| \cdot \|_{1,N}^*$-norm not exceeding $2^{kl}\gamma$, as required in (4.20).

4^0.4) It remains to verify (4.21). To save notation, let us prove this relation for $j = l$. By construction (see Mini-lemma and (4.51)) we have

$$\begin{array}{rcl}\pi(z) & = & \pi_1(z)\nu^l(z)\prod\limits_{s=1}^{k} r_T^{\lambda_{ls}}(z),\\ \pi_1(z) & = & \prod\limits_{j=1}^{l-1}\pi^j(z),\end{array}$$

whence, in view of (4.52),

$$\eta(z) = \eta^l(z)\underbrace{c_l^{-1}z^{-2MT}\pi_1(z)\prod_{s=1}^{k} r_T^{\lambda_{ls}}(z)}_{\rho^l(z)}.$$

Due to its origin, $\rho^l \in \mathcal{F}_{2\mu T}$. Furthermore, we have

$$\begin{array}{rcll}|c_l|^{-1} & \leq & 2^k & \text{[by (4.53)]}\\ \parallel \pi_1 \parallel_1 & \leq & 2^{k(l-1)} & \text{[by (4.58)]}\\ \parallel r_T^{\lambda_{ls}}(z) \parallel_1 & \leq & 2T & \text{[by Mini-lemma]}\\ \parallel r_T^{\lambda_{ls}}(z) \parallel_\infty & \leq & 2 & \text{[by Mini-lemma];}\end{array}$$

applying (4.7), we come to (4.21). ■

4.4.2 Proof of Lemma 4.3.1

Proof. We may assume that q has k distinct roots $\lambda_1, ..., \lambda_k$ [3]. Let the first ν of the roots belong to the closed left half-plane, and the rest of the roots belong from the open right half-plane.

Let us set

$$\begin{array}{rcl}\mu_s & = & \exp\{\lambda_s n^{-1}\},\ s = 1, ..., k;\\ \widehat{\theta}(z) & = & \left(\prod\limits_{s=1}^{\nu}(1 - z\mu_s)\right)\left(\prod\limits_{s=\nu+1}^{k}(z - 1/\mu_s)\right);\\ \chi_s(x) & = & \left\{\begin{array}{ll}\left\{\begin{array}{ll}0, & x \leq 0\\ \exp\{\lambda_s x\}, & x > 0\end{array}\right., & s \leq \nu\\ \left\{\begin{array}{ll}-\exp\{\lambda_s x\}, & x \leq 0\\ 0, & x > 0\end{array}\right., & k \geq s > \nu\end{array}\right.;\end{array}$$

note that the fundamental solutions $\exp\{\lambda_s x\}$ of the homogeneous equation $q\left(\frac{d}{dx}\right)p = 0$, being restricted on the grid $\{t/n\}_{t\in\mathbf{Z}}$, are proportional to the progressions $\{\mu_s^t\}_{t\in\mathbf{Z}}$ and therefore satisfy the homogeneous difference equation

$$\widehat{\theta}(\Delta)g \equiv 0.$$

Let

$$(a * b)(x) = \int_{-\infty}^{\infty} a(u)b(x-u)du$$

[3] The case when q has multiple roots can be obtained from the one with simple roots by perturbing q to make its roots distinct and then passing to limit as the perturbation tends to 0.

be the usual convolution, $\delta(x)$ be the Dirac delta-function and

$$\gamma(x) = \chi_1 * ... * \chi_k.$$

We have $q(z) = (z-\lambda_1)...(z-\lambda_n)$, so that

$$\begin{array}{rcl} q\left(\frac{d}{dx}\right)\gamma &=& \left((\frac{d}{dx}-\lambda_1)\chi_1\right) * ... * \left((\frac{d}{dx}-\lambda_k)\chi_k\right) \\ &=& \underbrace{\delta * ... * \delta}_{k \text{ times}} \\ &=& \delta, \end{array}$$

whence, setting

$$\begin{array}{rcl} h(x) &=& \begin{cases} \left(q\left(\frac{d}{dx}\right)g\right)(x), & x \in B \\ 0, & x \notin B \end{cases}, \\ r &=& \gamma * h, \end{array}$$

we get

$$q\left(\frac{d}{dx}\right)r = \left(q\left(\frac{d}{dx}\right)\gamma\right) * h = h = \left(q\left(\frac{d}{dx}\right)g\right)\chi_{x\in B}.$$

Thus, for $x \in \text{int } B$ we have

$$\left(q\left(\frac{d}{dx}\right)(g-r)\right)(x) = 0 \Rightarrow g(x) - r(x) = \sum_{s=1}^{k} c_s \exp\{\lambda_s x\},$$

whence

$$\left(\widehat{\theta}(\Delta)(g^n - r^n)\right)_t = 0 \quad \forall t : \frac{t \pm k}{n} \in \text{int } B \Rightarrow \| \widehat{\theta}(\Delta)(g^n - r^n) \|_{p,4\mu T} = 0 \qquad (4.60)$$

(recall that for $|t| \leq 4\mu T$ the points $(t \pm k)/n$ belong to B, since B is centered at the origin and contains at least $8\mu T + 2k + 1$ points of the grid Γ^n).

Now let us compute $\widehat{\theta}(\Delta)r^n$. Let $\overline{\Delta}$ be the shift by n^{-1} in the space of functions on the axis:

$$(\overline{\Delta}f)(x) = f(x - n^{-1}).$$

Then

$$\begin{array}{rcl} \left(\widehat{\theta}(\Delta)r^n\right)_t &=& \left(\widehat{\theta}(\overline{\Delta})r\right)(tn^{-1}), \\ \widehat{\theta}(\overline{\Delta})r &=& \widehat{\theta}(\overline{\Delta})(\gamma * h) \\ && [\text{since } r = \gamma * h] \\ &=& \left(\widehat{\theta}(\overline{\Delta})\gamma\right) * h \\ && [\text{since } \overline{\Delta}(f * e) = (\overline{\Delta}f) * e] \\ &=& \underbrace{\left(\chi_1 - \mu_1\overline{\Delta}\chi_1\right)}_{\psi_1} *... * \underbrace{\left(\chi_\nu - \mu_\nu\overline{\Delta}\chi_\nu\right)}_{\psi_\nu} \\ && * \underbrace{\left(\overline{\Delta}\chi_{\nu+1} - \mu_{\nu+1}^{-1}\chi_{\nu+1}\right)}_{\psi_{\nu+1}} *... * \underbrace{\left(\overline{\Delta}\chi_k - \mu_k^{-1}\chi_k\right)}_{\psi_k} *h. \end{array}$$

Now note that every one of the functions $\psi_s(\cdot)$ in absolute value does not exceed 1, and that it vanishes outside $[0, n^{-1}]$. It follows that the function $\psi = \psi_1 *...* \psi_k$ vanishes outside $[0, kn^{-1}]$ and does not exceed in absolute value the quantity $\| \psi_1 \|_1 ... \| \psi_{k-1} \|_1 \| \psi_k \|_\infty \leq n^{-(k-1)}$.

Assuming $1 \le p < \infty$, we have

$$\begin{aligned}
\| \widehat{\theta}(\Delta) r^n \|_{p,4\mu T}^p &= \sum_{t=-4\mu T}^{4\mu T} \left| \left(\widehat{\theta}(\overline{\Delta}) r\right)(tn^{-1}) \right|^p \\
&= \sum_{t=-4\mu T}^{4\mu T} \left| (\psi * h)(tn^{-1}) \right|^p \\
&= \sum_{t=-4\mu T}^{4\mu T} \left| \int\limits_0^{kn^{-1}} \psi(u) h(tn^{-1} - u) du \right|^p \\
&\le \sum_{t=-4\mu T}^{4\mu T} \left(\int\limits_0^{kn^{-1}} n^{-(k-1)} |h(tn^{-1} - u)| du \right)^p \\
&\le \sum_{t=-4\mu T}^{4\mu T} n^{-(k-1)p} \int\limits_0^{kn^{-1}} |h(tn^{-1} - u)|^p du (kn^{-1})^{p-1} \\
&= k^{p-1} n^{-kp+1} \int |h(u)|^p C(u) du \\
&\quad [C(u) = \mathrm{Card}\left(\{t \in \mathbf{Z} : |t| \le 4\mu T, (-k+t)n^{-1} \le u \le tn^{-1}\}\right)] \\
&\le k^p n^{-kp+1} \| h \|_{p,B}^p \\
\Rightarrow \quad \| \widehat{\theta}(\Delta) r^n \|_{p,4\mu T} &\le k n^{-k+1/p} \| q\left(\tfrac{d}{dx}\right) g \|_{p,B} .
\end{aligned}$$

Combining the resulting inequality with (4.60), we get

$$\| \widehat{\theta}(\Delta) g^n \|_{p,4\mu T} \le k n^{-k+1/p} \| q\left(\frac{d}{dx}\right) g \|_{p,B} . \qquad (4.61)$$

This inequality was obtained in the case of $p < \infty$; by continuity reasons, it is valid for $p = \infty$ as well.

Relation (4.61) is nearly what we need; the only bad thing is that the polynomial $\widehat{\theta}$ is not normalized. Let w be the maximum of absolute values of the coefficients of $\widehat{\theta}$; setting

$$\theta = w^{-1} \widehat{\theta},$$

we get a normalized polynomial of degree k such that

$$\| \theta(\Delta) g^n \|_{p,4\mu T} \le w^{-1} k n^{-k+1/p} \| q\left(\frac{d}{dx}\right) g \|_{p,B} . \qquad (4.62)$$

It remains to bound from above the quantity w^{-1}, which is immediate: there exists a point z^* in the unit circle which is at the distance at least $d = (1+\sqrt{k})^{-1}$ from all points $\mu_{\nu+1}^{-1}, ..., \mu_k^{-1}$ and at at least at the same distance from the boundary of the circle (otherwise the circles of the radius d centered at the points were covering the circle of the radius $1 - d$ centered at the origin, which is clearly impossible – compare the areas!). From the formula for $\widehat{\theta}$ it follows that

$$\sum_{s=0}^{k} |\widehat{\theta}_s| \ge |\widehat{\theta}(z^*)| \ge (1 + \sqrt{k})^{-k},$$

whence

$$w^{-1} \le (k+1)(1 + \sqrt{k})^k.$$

Combining this inequality with (4.62), we come to (4.34). ■

Chapter 5

Aggregation of estimates, I

5.1 Motivation

The non-parametric regression estimates we have built so far heavily depend on a priori assumptions on the structure of the function to be recovered. As a matter of fact, this dependence of estimation techniques on a priori hypotheses concerning the structure of "true signals" is a characteristic feature of the non-parametric regression estimates; we can reduce sometimes the "size" of the required a priori knowledge, but we never can get rid of it completely. Now, typically there are many "concurrent" a priori hypotheses on the structure of a signal rather than a single hypothesis of this type; if we knew which one of our a priori hypotheses indeed takes place, we would know how to recover the signal. The difficulty, however, is that we do not know in advance which one of our concurrent hypotheses actually takes place. We already met this situation in adaptive estimation of smooth functions, where the hypotheses were parameterized by the smoothness parameters of local Sobolev balls, a particular hypothesis saying that the signal belongs to a particular Sobolev ball (and similarly in the case of recovering functions satisfying differential inequalities). As another example of this type, assume that we are recovering a smooth regression function f of $d > 1$ variables and that we have reasons to suppose that in fact f depends on $d' < d$ "properly chosen" variables:

$$f(x) = F(P^T x), \tag{5.1}$$

where P is a $d' \times d$ matrix. If we knew P in advance, we could reduce the problem of recovering f to the one of recovering F. Since the rates of convergence of non-parametric estimates rapidly slows down with the dimensionality of the problem (e.g., for Lipschitz continuous functions of d variables the convergence rate is $O(n^{-\frac{1}{2+d}})$ – think how many observations we need to get a reasonable accuracy when estimating a Lipschitz continuous function of just 4 variables), such an opportunity would look very attractive. But what to do when we know that a representation (5.1) exists, but do not know the matrix P?

The "general form" of the situation we are interested in is as follows. We have a family $\mathcal{H}$ of a priori hypotheses on the signal f, and we know in advance that at least one of these hypotheses is true. If we knew that f fits a particular hypothesis $H \in \mathcal{H}$, we would know how to recover f – in other words, every hypothesis H is associated with a recovering routine $\hat{f}^H$ which "works fine" when f fits H. However, we do not know what is the hypothesis the observed signal fits. What to do?

Sometimes (e.g., in the case of recovering smooth functions or functions satisfying differential inequalities) we may act as if we knew the "true" hypothesis, but this possibility heavily depends on the specific nature of the corresponding family of hypotheses $\mathcal{H}$; for other families $\mathcal{H}$, no results of this type are known. This is the case, e.g., for the family associated with representation (2.1) with given d, d' and varying P.

In the general case we could act as follows: we could partition our observations y into two groups and use the observations of the first group, y^I, to build all estimates $f^H(\cdot) = \hat{f}^H(\cdot, y^I)$, $H \in \mathcal{H}$, of f; after this is done, we could use the second group of observations, y^{II}, in order to "aggregate" the estimates f^H – to build a new estimate which reproduces f (nearly) as good as the best of the functions f^H, $H \in \mathcal{H}$. Since in our approach the family of hypotheses/estimates is given in advance and is therefore beyond our control, our problem is how to implement the "aggregation" stage; how we resolve this problem, it depends on what exactly is our target. Mathematically natural targets could be to find an "aggregated" estimate which is nearly as good as

L. The closest to f *linear* combination of the functions f^H, $H \in \mathcal{H}$;

C. The closest to f *convex* combination of the functions f^H, $H \in \mathcal{H}$;

V. The closest to f of the functions f^H, $H \in \mathcal{H}$.

To the moment, the three outlined versions of the Aggregation problem were investigated in the case when

- The number of "basic estimates" is finite.
- The estimation error is measured in $L_2(X, \mu)$, X being a space on which f, f^H are defined and μ being a probability measure on this space.

The majority of known results relate to the version **V** of the aggregation problem (see [11] and references therein). In our course, we prefer to start with the version **C**, postponing the versions **L**, **V** till Chapter 6.

5.2 The problem and the main result

5.2.1 Aggregation problem

We are about to consider the following

Aggregation problem **C**. Let

- $\Lambda \subset \mathbf{R}^M$ be a convex compact set contained in the $\|\cdot\|_1$-ball, i.e., let

$$\max\{\| \lambda \|_1 |\ \lambda \in \Lambda\} \leq 1;$$

- X be a Polish space equipped with Borel probability measure μ;
- $f_j : X \to \mathbf{R}$, $j = 1, ..., M$, $M \geq 3$, be given Borel functions;
- $f : X \to \mathbf{R}$ be a Borel function.

Assume that we are given n noisy observations of f:

$$z = \{z_t = (x_t, y_t = f(x_t) + e_t)\}_{t=1}^n, \tag{5.2}$$

where x_t are mutually independent random points from X, each of them being distributed according to μ, and e_t are independent of each other and of $\{x_t\}$ random noises such that

$$\mathcal{E}\{e_t\} = 0 \text{ and } \mathcal{E}\{e_t^2\} \le \sigma^2 < \infty, \ t = 1, ..., n. \tag{5.3}$$

Let f_Λ be the closest to f, in $L_2(X, \mu)$, linear combination of functions $f_1, ..., f_M$ with coefficients from Λ:

$$\begin{array}{rcl} f_\Lambda & = & \sum\limits_{j=1}^M \lambda_j^* f_j, \\ \lambda_* & \in & \mathop{\mathrm{Argmin}}\limits_{\lambda \in \Lambda} \Psi(\lambda), \\ \Psi(\lambda) & = & \int\limits_X (f(x) - \sum\limits_{j=1}^M \lambda_j f_j(x))^2 \mu(dx). \end{array} \tag{5.4}$$

Our goal is to find, given $f_1, ..., f_M$ and n observations (5.2), a combination $\sum\limits_j \lambda_j f_j$ with $\lambda \in \Lambda$ which is nearly as close to f as f_Λ.

It should be stressed that we do *not* assume that the measure μ is known in advance.

From now on, we make the following crucial for us

Boundedness assumption: Functions $f, f_1, ..., f_M$ are bounded.

From now on, we set

$$L = \max\{\| f \|_\infty, \| f_1 \|_\infty, ..., \| f_M \|_\infty\} < \infty, \tag{5.5}$$

the ∞-norm being associated with the measure μ.

5.2.2 The recovering routine

Our recovering routine is extremely simple. The function $\Psi(\lambda)$ from (5.4) is a convex quadratic form of λ:

$$\begin{array}{rcl} \Psi(\lambda) & = & \Psi^*(\lambda) + c^*, \\ \Psi^*(\lambda) & = & \sum\limits_{i,j=1}^M Q_{ij}^* \lambda_i \lambda_j - \sum\limits_{j=1}^M q_j^* \lambda_j, \\ Q_{ij}^* & = & \int\limits_X f_i(x) f_j(x) \mu(dx), \\ q_j^* & = & 2\int\limits_X f(x) f_j(x) \mu(dx), \\ c^* & = & \int\limits_X f^2(x) \mu(dx). \end{array} \tag{5.6}$$

Given a quadratic form $\Phi(\lambda)$ on $\mathbf{R}^M$ with $\Phi(0) = 0$:

$$\Phi(\lambda) = \sum_{i,j=1}^M Q_{ij} \lambda_i \lambda_j - \sum_{j=1}^M q_j \lambda_j \qquad [Q_{ij} = Q_{ji}]$$

let us denote by

$$\mathrm{Coef}(\Phi) = (\{Q_{ij}\}_{1\le j\le i\le M}, \{q_j\}_{j=1}^M)$$

the $\left(\frac{M(M+1)}{2} + M\right)$-dimensional vector of coefficients of the form.

Note that every observation $(x_t, y_t = f(x_t) + e_t)$ provides us with a noisy observation

$$\zeta^t = (\{f_i(x_t)f_j(x_t)\}_{1\le j\le i\le M}, \{2y_t f_j(x_t)\}_{j=1}^M) \tag{5.7}$$

of the vector

$$\zeta^* = \mathrm{Coef}(\Psi^*),$$

and that ζ^t is the vector of coefficients of the convex quadratic form of rank 1

$$\Psi^{z_t}(\lambda) = \left(y_t - \sum_{j=1}^M \lambda_j f_j(x_t)\right)^2 - y_t^2.$$

Our aggregation procedure is as follows: given observations (5.2), we

1) build the form

$$\Psi^z(\lambda) = \frac{1}{n}\sum_{t=1}^n \Psi^{z_t}(\lambda),$$

and

2) solve the convex optimization problem

$$\Psi^z(\lambda) \to \min \mid \lambda \in \Lambda. \tag{P_z}$$

An optimal solution $\lambda(z)$ to this problem clearly can be chosen to be Borel in z;

3) We define our "aggregated estimate" as

$$\widehat{f}(\cdot; z) = \sum_{j=1}^M \lambda_j(z) f_j(\cdot).$$

5.2.3 Main result

Our main result bounds the difference between the quality of the "ideal", as far as closeness to f is concerned, aggregate of f_j with coefficients from Λ and the expected quality of the aggregate $\widehat{f}$ we have built, i.e., the difference between the quantities

$$\Psi(\lambda^*) = \min_{\lambda\in\Lambda} \int\limits_X (f(x) - \sum_{j=1}^M \lambda_j f_j(x))^2 \mu(dx)$$

and

$$\mathcal{E}\{\Psi(\lambda(z))\} = \mathcal{E}\left\{\int\limits_X (f(x) - \widehat{f}(x;z))^2 \mu(dx)\right\}.$$

Note that a meaningful "quality measure" for an aggregation routine should be exactly of this type – it should bound the *difference* between the expected distance from f to the result of the aggregation routine in question and the distance from f to the "ideal" aggregate f_Λ, not the distance from f to the result of the aggregation routine separately. Indeed, since we make no assumptions on how well the "ideal" aggregate approximates f, we have no hope to ensure that the result of an aggregation routine (which cannot be closer to f than the ideal aggregate) is a good approximation of f; all we should worry about is to get an aggregate which is nearly as good as the ideal one.

Theorem 5.2.1 *For the aggregation routine $\widehat{f}$ we have built, one has*

$$\varepsilon_n \equiv \mathcal{E}\left\{\Psi(\lambda(z))\right\} - \Psi(\lambda^*) \le O(1)\frac{(L^2 + L\sigma)\sqrt{\ln M}}{\sqrt{n}} \tag{5.8}$$

with absolute constant $O(1)$.

Discussion

The quantity ε_n in the left hand side of (5.8) can be treated as the "aggregation price" – the loss in accuracy of approximating f by a linear combination of f_j (with coefficients from Λ) coming from the fact that we do not know the "true" optimal combination (since neither f nor even μ are known in advance) and are enforced to recover a nearly optimal combination from observations. Note that ε_n is the expected loss in the *squared* $\| \cdot \|_2$-distance from f ($\| \cdot \|_2$ is associated with the measure μ). A more natural price is the loss in the $\| \cdot \|_2$-distance itself – the quantity

$$\nu_n = \| f - \widehat{f} \|_2 - \| f - f_\Lambda \|_2 .$$

Since for $0 \le a \le b$ one has $(b-a)^2 \le b^2 - a^2$, (5.8) implies that

$$E_n \equiv \left(\mathcal{E}\left\{\nu_n^2\right\}\right)^{1/2} \le \sqrt{\varepsilon_n} \le O(1)\frac{(L + \sqrt{L\sigma})(\ln M)^{1/4}}{n^{1/4}}. \tag{5.9}$$

A good news about the latter bound is that it is "nearly independent of the number M of functions we are estimating" – it is proportional to $(\ln M)^{1/4}$. Thus, if our aggregation problem comes from the desire to aggregate estimates associated with a number of concurrent hypotheses on the signal f to be recovered, this number can be "very large". From the applied viewpoint, it means that our abilities to handle many concurrent hypotheses are limited not by the statistics – by growth of the aggregation price with the number of hypotheses – but by the necessity to process these hypotheses computationally. And a bad news about our aggregation routine is that the aggregation price E_n decreases rather slowly (as $n^{-1/4}$) as the volume n of observations used for aggregation grows. We shall see, however, that in our setting of the aggregation problem this rate is unimprovable.

Note that one can replace the "off-line" aggregation routine we have described (where we first accumulate all observations (5.2) and only then solve a (large-scale, for large M) convex optimization problem (P_z) to build the desired aggregate) with a Stochastic Approximation-type on-line routine where neither the observations should be stored, nor a separate stage of solving a convex optimization problem is needed (for details, see [16]).

Proof of Main result

Proof of Theorem 5.2.1 is given by combination of two simple observations; the second of them is interesting by its own right.

The first observation is given by

Lemma 5.2.1 *The random vectors ζ^t given by (5.7) are mutually independent and unbiased estimates of ζ^*:*

$$\mathcal{E}\{\zeta^t\} = \zeta^*. \tag{5.10}$$

Besides this,

$$\mathcal{E}\left\{\| \zeta^t - \zeta^* \|_\infty^2\right\} \leq 4(2L^2 + \sigma L)^2, \tag{5.11}$$

(From now on, for $\xi = (\xi_1, ..., \xi_K) \in \mathbf{R}^K$ $\| \xi \|_\infty$ is the norm $\max_k |\xi_k|$ of the vector ξ).

Proof. Mutual independence of $\{\zeta^t\}_{t=1}^n$ and relation (5.10) are evident. To establish (5.11), note that

$$\begin{array}{rcl} |Q^*_{ij} - f_i(x_t)f_j(x_t)| & \leq & 2L^2, \\ |q_j - 2(f(x_t) + e_t)f_j(x_t)| & \leq & 4L^2 + 2L|e_t| \Rightarrow \\ \| \zeta^t - \zeta^* \|_\infty^2 & \leq & 4(2L^2 + L|e_t|)^2 \Rightarrow \quad \blacksquare \\ \mathcal{E}\left\{\| \zeta^t - \zeta^* \|_\infty^2\right\} & \leq & 4(2L^2 + L\sigma)^2. \end{array}$$

2^0. Our second observation is an extremely useful "Tschebyshev inequality in the ∞-norm". Recall that the usual Tschebyshev inequality gives a rough upper bound on the probability of the event $|\sum\limits_{t=1}^n \xi^t| > a$, where ξ^t are independent scalar random variables with zero mean and finite variance; this inequality is an immediate consequence of the observation that in the case in question

$$\mathcal{E}\left\{|\sum_{t=1}^n \xi^t|^2\right\} = \sum_{t=1}^n \mathcal{E}\{|\xi^t|^2\}.$$

Similar equality *with respect to the Euclidean norm* takes place if ξ^t are independent vectors with zero mean and bounded variances:

$$\mathcal{E}\left\{\| \sum_{t=1}^n \xi^t \|_2^2\right\} = \sum_{t=1}^n \mathcal{E}\{\| \xi^t \|_2^2\}, \tag{$*$}$$

where for $\xi = (\xi_1, ..., \xi_K) \in \mathbf{R}^K$

$$\| \xi \|_p = \begin{cases} \left(\sum\limits_{i=1}^K |\xi_i|^p\right)^{1/p}, & 1 \leq p < \infty \\ \max_i |\xi_i|, & p = \infty \end{cases}.$$

Now, (*) reflects specific algebraic properties of the Euclidean norm $\| \cdot \|_2$ and fails to be valid for the standard norms $\| \cdot \|_p$ with $p \neq 2$. As far as statistical consequences are concerned, the "$\| \cdot \|_p$-version" of (*) is played by the following result[1]:

Lemma 5.2.2 *Let $\xi_t \in \mathbf{R}^K$, $t = 1, ..., n$, be independent random vectors with zero means and finite variance, and let $K \geq 3$. Then for every $p \in [2, \infty]$ one has*

$$\mathcal{E}\left\{\| \sum_{t=1}^n \xi^t \|_p^2\right\} \leq O(1) \min[p, \ln K] \sum_{t=1}^n \mathcal{E}\left\{\| \xi^t \|_p^2\right\}; \tag{5.12}$$

here, as always, $O(1)$ is an absolute constant.

[1] I am using this fact for more than 20 years; all this time I was (and still am) sure that the fact is well-known, all this time I was looking for a reference and found none.

Proof. Given $\pi \in [2,\infty)$, let us set

$$V_\pi(\xi) = \| \xi \|_\pi^2 : \mathbf{R}^K \to \mathbf{R}.$$

The function V_π is continuously differentiable with Lipschitz continuous gradient; it can be easily verified (for the proof, see [21]) that

$$V_\pi(\xi+\eta) \leq V_\pi(\xi) + \eta^T \nabla V_\pi(\xi) + C\pi V_\pi(\eta) \tag{5.13}$$

with absolute constant C. We conclude that

$$\begin{array}{rcl}
\mathcal{E}\left\{ V_\pi(\sum\limits_{t=1}^{k+1} \xi^t) \right\} & \leq & \mathcal{E}\left\{ V_\pi(\sum\limits_{t=1}^{k} \xi^t) + (\xi^{k+1})^T \nabla V_\pi(\sum\limits_{t=1}^{k} \xi^t) \right\} + C\pi \mathcal{E}\left\{ V_\pi(\xi^t) \right\} \\
& = & \mathcal{E}\left\{ V_\pi(\sum\limits_{t=1}^{k} \xi^t) \right\} + C\pi \mathcal{E}\left\{ V_\pi(\xi^t) \right\} \\
& & [\text{since } \mathcal{E}\{\xi^{k+1}\} = 0 \text{ and } \xi^{k+1} \text{ is independent of } \xi^1, ..., \xi^k]
\end{array}$$

The resulting recurrence implies that whenever $p \in [2,\infty)$, one has

$$\mathcal{E}\left\{ \| \sum_{t=1}^{n} \xi^t \|_p^2 \right\} \leq Cp \sum_{t=1}^{n} \mathcal{E}\left\{ \| \xi^t \|_p^2 \right\}. \tag{5.14}$$

To complete the proof of (5.12), it suffices to verify that we can replace the factor Cp in the right hand side by a factor of the type $O(1)\ln K$. This is immediate: there is nothing to prove when $p \leq p(K) \equiv 2\ln K$. Now let us assume that $p > 2\ln K$. Since for $p \geq p' \geq 1$ one has

$$\| \xi \|_p \leq \| \xi \|_{p'} \leq K^{\frac{1}{p'}-\frac{1}{p}} \| \xi \|_p \qquad \forall \xi \in \mathbf{R}^K$$

we have

$$\begin{array}{rcl}
\mathcal{E}\left\{ \| \sum\limits_{t=1}^{n} \xi^t \|_p^2 \right\} & \leq & \mathcal{E}\left\{ \| \sum\limits_{t=1}^{n} \xi^t \|_{p(K)}^2 \right\} \\
& \leq & Cp(K) \sum\limits_{t=1}^{n} \mathcal{E}\left\{ \| \xi^t \|_{p(K)}^2 \right\} \\
& & [\text{by (5.14) applied with } p = p(K)] \\
& \leq & Cp(K) \sum\limits_{t=1}^{n} \mathcal{E}\left\{ K^{\frac{2}{p(K)}-\frac{2}{p}} \| \xi^t \|_p^2 \right\} \\
& \leq & Cp(K) K^{\frac{2}{p(K)}} \sum\limits_{t=1}^{n} \mathcal{E}\left\{ \| \xi^t \|_p^2 \right\} \\
& = & 2Ce \ln K \sum\limits_{t=1}^{n} \mathcal{E}\left\{ \| \xi^t \|_p^2 \right\} \\
& & [\text{since } p(K) = 2\ln K]
\end{array}$$

■

3^0. We are basically done. Indeed, since Λ is contained in the unit $\| \cdot \|_1$-ball in $\mathbf{R}^M$, the uniform, on Λ, distance between a pair of quadratic forms Ψ, Ψ' of λ, both forms being with zero constant terms, does not exceed 3 times the $\| \cdot \|_\infty$-distance between the coefficient vectors of the forms:

$$\begin{array}{rcl}
\Psi(\lambda) & = & \sum\limits_{i,j=1}^{M} Q_{ij}\lambda_i\lambda_j - \sum\limits_{j=1}^{M} q_j\lambda_j, \\
\Psi'(\lambda) & = & \sum\limits_{i,j=1}^{M} Q'_{ij}\lambda_i\lambda_j - \sum\limits_{j=1}^{M} q'_j\lambda_j \\
\Rightarrow \max\limits_{\lambda\in\Lambda} |\Psi(\lambda) - \Psi'(\lambda)| & \leq & 3 \,\| \operatorname{Coef}(\Psi) - \operatorname{Coef}(\Psi') \|_\infty .
\end{array}$$

It follows that if λ' is a minimizer of Ψ' on Λ and $\|\ \mathrm{Coef}(\Psi) - \mathrm{Coef}(\Psi')\ \|_\infty$ is small, then λ' is a "nearly minimizer" of Ψ on Λ:

$$\begin{array}{rcl}\lambda' & \in & \underset{\Lambda}{\mathrm{Argmin}}\Psi'(\cdot) \Rightarrow \\ \Psi(\lambda') - \min\limits_{\Lambda}\Psi(\cdot) & \leq & 2\max\limits_{\lambda\in\Lambda}|\Psi(\lambda) - \Psi'(\lambda)| \\ & \leq & 6\ \|\ \mathrm{Coef}(\Psi) - \mathrm{Coef}(\Psi')\ \|_\infty .\end{array} \tag{5.15}$$

Now, the output of our aggregation routine – the vector of aggregation weights $\lambda(z)$ – by construction is a minimizer, on Λ, of a random quadratic form $\Psi^z(\lambda) = \frac{1}{n}\sum\limits_{t=1}^{n}\Psi^{z_t}(\lambda)$, so that our quality measure – the "aggregation price" – can be bounded as follows:

$$\begin{array}{rcl}\Psi(\lambda(z)) - \min_\Lambda \Psi(\cdot) & = & \Psi^*(\lambda(z)) - \min_\Lambda \Psi^*(\cdot) \\ & & [\text{since } \Psi \text{ differs from } \Psi^* \text{ by a constant}] \\ & \leq & 6\ \|\ \mathrm{Coef}(\Psi^*) - \mathrm{Coef}(\Psi^z)\ \|_\infty \\ & & [\text{by (5.15)}] \\ & = & \frac{6}{n}\ \|\ \sum\limits_{t=1}^{n}[\zeta^* - \zeta^t]\ \|_\infty \\ & & [\text{by construction}] \\ \Rightarrow & & \\ \varepsilon_n & \equiv & \mathcal{E}\left\{\Psi(\lambda(z)) - \min_\Lambda \Psi(\cdot)\right\} \\ & \leq & \frac{6}{n}\mathcal{E}\left\{\|\ \sum\limits_{t=1}^{n}[\zeta^* - \zeta^t]\ \|_\infty\right\} \\ & \leq & \frac{6}{n}\left(\mathcal{E}\left\{\|\ \sum\limits_{t=1}^{n}[\zeta^* - \zeta^t]\ \|_\infty^2\right\}\right)^{1/2} \\ & \leq & \frac{6}{n}\left(O(1)\ln M\left[\sum\limits_{t=1}^{n}4(2L^2+\sigma L)^2\right]\right)^{1/2} \\ & & [\text{by Lemmas 5.2.1, 5.2.2}] \\ & \leq & O(1)\frac{(L^2+\sigma L)\sqrt{\ln M}}{\sqrt{n}},\end{array}$$

as required. ∎

5.2.4 "Concentration"

From the computational viewpoint, a drawback of our aggregation routine is that the resulting aggregate $\widehat{f}$ can involve all our M functions $f_1, ..., f_M$. If M is very large (and this is the case we indeed are interested in), such an aggregate is computationally difficult to use.

We are about to prove that in fact the aggregate $\widehat{f}$ can be enforced to involve at most $O(n)$ or even $O(n^{1/2})$ of the functions $f_1, ..., f_M$, provided that Λ is "simple", e.g.,

$$\Lambda = \{\lambda \in \mathbf{R}^M \mid \|\ \lambda\ \|_1 \leq 1\} \tag{5.16}$$

$$\Lambda = \{\lambda \in \mathbf{R}^M \mid \lambda \geq 0, \|\ \lambda\ \|_1 \leq 1\} \tag{5.17}$$

$$\Lambda = \{\lambda \in \mathbf{R}^M \mid \lambda \geq 0, \|\ \lambda\ \|_1 = 1\} \tag{5.18}$$

"n-concentrated" aggregation. Given an M-dimensional vector ω with coordinates ± 1, let us set

$$\mathbf{R}^M_\omega = \{\lambda \in \mathbf{R}^M \mid \omega_j \lambda_j \geq 0,\ j = 1, ..., M\}.$$

Let us call Λ *k-simple*, if the intersection of Λ with every one of 2^M "orthants" $\mathbf{R}^M_\omega$ is a polyhedral set cut off $\mathbf{R}^M_\omega$ by at most k linear equalities and inequalities (in addition to M "sign constraints" which define $\mathbf{R}^M_\omega$ itself). E.g., every one of the sets (5.16) – (5.18) is 1-simple.

Note that the weight vector $\lambda(z)$ yielded by our aggregation routine is not necessarily unique. Indeed, we can choose as $\lambda(z)$ *any* minimizer (on Λ) of the quadratic form $\Psi^z(\cdot)$. The quadratic part of each of the forms $\Psi^{z_t}(\cdot)$, $t = 1, ..., n$, is of rank 1, so that the rank of the quadratic part of the form $\Psi^z(\cdot)$ is of rank at most n. It follows that there exists a linear subspace $E^z \subset \mathbf{R}^M$ of codimension at most $n + 1$ such that $\Psi^z(\cdot)$ is constant along every translation of this subspace. In particular, after we have found a minimizer $\lambda(z)$ of $\Psi^z(\cdot)$ on Λ, we can "refine" it as follows. Let ω be such that $\lambda(z) \in \mathbf{R}^M_\omega$. Consider the set

$$P = \Lambda \cap \mathbf{R}^M_\omega \cap [E^z + \lambda(z)].$$

Every point of this set (which contains $\lambda(z)$) is a minimizer of $\Psi^z(\cdot)$ on Λ, along with $\lambda(z)$ (since Ψ^z is constant on $E^z + \lambda(z)$). Assuming that Λ is k-simple, we observe that P is a compact polyhedral set given by M "sign constraints" defining $\mathbf{R}^M_\omega$ and no more than $k+n+1$ additional linear inequalities and equations (at most k linear constraints which cut off $\Lambda \cap \mathbf{R}^M_\omega$ from $\mathbf{R}^M_\omega$ plus $n+1$ linear equation defining the affine plane $E^z + \lambda(z)$). As any compact polyhedral set, P has extreme points, and by the standard results of Linear Programming every extreme point of P fits at least M of equations/inequalities defining P as equations. We are in our right to choose, as a minimizer of $\Psi^z(\cdot)$ on Λ, any one of these extreme points, let the chosen point be denoted $\lambda^+(z)$, and to treat $\lambda(z)$ as an intermediate, and $\lambda^+(z)$ – as the actual output of our aggregation routine. It remains to note that among $\geq M$ of equations/inequalities defining P which are satisfied at $\lambda^+(z)$ as equalities, at least $M - (k + n + 1)$ must come from the sign constraints defining the orthant $\mathbf{R}^M_\omega$, i.e., at least $M - (k + n + 1)$ coordinates in $\lambda^+(z)$ must be zero. We have arrived at the following

Proposition 5.2.1 *Assume that Λ is k-simple. Then in our aggregation routine we can specify the rules for choosing the weight vector $\lambda(z)$ in such a way that the aggregate*

$$\widehat{f}(\cdot; z) = \sum_{j=1}^{M} \lambda_j(z) f_j(\cdot)$$

will include, with positive weights $\lambda_j(z)$, no more than $k + n + 1$ of the functions f_j.

"$n^{1/2}$-concentrated" aggregation. The construction we are about to present goes back to Maurey [28]. We shall implement the construction under the assumption that Λ is the $\|\cdot\|_1$-unit ball (5.16); however, our reasoning can be easily modified to handle the case of simplices (5.17), (5.18).

Our new aggregation routine is randomized. Namely, we first apply our basic routine to get the vector of aggregation weights $\lambda(z)$. After it is found, we set

$$\nu(z) = \sum_{j=1}^{M} |\lambda_j(z)|$$

(note that $\nu(z) \leq 1$) and define a probability measure $\{\pi_j^z\}_{j=0}^M$ on the set $\{0, 1, ..., M\}$ as follows:

$$\pi_j^z = \begin{cases} |\lambda_j(z)|, & j > 0 \\ 1 - \nu, & j = 0 \end{cases}$$

For $0 \leq j \leq M$, let us set

$$g_j^z(\cdot) = \begin{cases} 0, & j = 0 \\ f_j(\cdot), & j > 0, \lambda_j(z) \geq 0 \\ -f_j(\cdot), & j > 0, \lambda_j(z) < 0 \end{cases}.$$

Note that we can represent the aggregate $\widehat{f}(\cdot; z) = \sum\limits_{j=1}^{M} \lambda_j(z) f_j(\cdot)$ as the expectation of "random function" g_j^z with respect to the distribution π^z of the index j:

$$\widehat{f}(\cdot; z) = \sum_{j=0}^{M} \pi_j^z g_j^z(\cdot).$$

Now let us draw independently of each other K indices $j_1, ..., j_K$ according to the probability distribution π^z and let us set

$$\widetilde{f}(\cdot; z, \bar{j}) = \frac{1}{K} \sum_{l=1}^{K} g_{j_l}^z(\cdot) \qquad [\bar{j} = (j_1, ..., j_k)]$$

Note that the resulting function is obtained from $f_1, ..., f_M$ by linear aggregation with the weight vector $\widetilde{\lambda}(z, \bar{j}) \in \Lambda$ which is "K-concentrated" – has at most K nonzero entries.

Now let us look at the "aggregation price"

$$\widetilde{\varepsilon}_n(K) \equiv \mathcal{E}_{z,\bar{j}} \left\{ \Psi(\widetilde{\lambda}(z, \bar{j})) - \min_{\Lambda} \Psi(\cdot) \right\}$$

of our new – randomized – aggregation routine. Treating $g_j^z(\cdot)$ as a random element of $L_2(X, \mu)$, the conditional, for z fixed, distribution of j being π^z, we observe that

(a) $g_{j_1}^z, ..., g_{j_K}^z$ are conditionally, z being fixed, independent and identically distributed with conditional expectation $\widehat{f}(\cdot; z)$

(b) The conditional, z being fixed, expectation of $\| g_{j_l}^z(\cdot) - \widehat{f}(\cdot; z) \|_{2,\mu}^2$ does not exceed L^2, where $\| \cdot \|_{2,\mu}$ is the standard norm of $L_2(X, \mu)$.

We now have

$$\begin{array}{rll}
& \mathcal{E}_{z,\bar{j}}\left\{\Psi(\widetilde{\lambda}(z,\bar{j}))\right\} & \\
= & \mathcal{E}_z\left\{\mathcal{E}_{\bar{j}|z}\left\{\| \frac{1}{K}\sum\limits_{l=1}^{K} g^z_{\bar{j}_l} - f \|^2_{2,\mu}\right\}\right\} & \\
= & \mathcal{E}_z\left\{\mathcal{E}_{\bar{j}|z}\left\{\| \left[\frac{1}{K}\sum\limits_{l=1}^{K}[g^z_{\bar{j}_l}(\cdot) - \widehat{f}(\cdot;z)]\right] + \left[\widehat{f}(\cdot;z) - f(\cdot)\right] \|^2_{2,\mu}\right\}\right\} & \\
= & \mathcal{E}_z\left\{\mathcal{E}_{\bar{j}|z}\left\{\| \frac{1}{K}\sum\limits_{l=1}^{K}[g^z_{\bar{j}_l}(\cdot) - \widehat{f}(\cdot;z)] \|^2_{2,\mu}\right\} + \| \widehat{f}(\cdot;z) - f(\cdot) \|^2_{2,\mu}\right\} & \text{[by (a)]} \\
= & \mathcal{E}_z\left\{\frac{1}{K^2}\sum\limits_{l=1}^{K}\mathcal{E}_{\bar{j}|z}\left\{\| g^z_{\bar{j}_l}(\cdot) - \widehat{f}(\cdot;z) \|^2_{2,\mu}\right\} + \| \widehat{f}(\cdot;z) - f(\cdot) \|^2_{2,\mu}\right\} & \text{[by (a)]} \\
\leq & \mathcal{E}_z\left\{\frac{L^2}{K} + \| \widehat{f}(\cdot;z) - f(\cdot) \|^2_{2,\mu}\right\} & \text{[by (b)]} \\
\leq & \frac{L^2}{K} + \mathcal{E}_z\left\{\| \widehat{f}(\cdot;z) - f(\cdot) \|^2_{2,\mu}\right\} & \\
= & \frac{L^2}{K} + \mathcal{E}\left\{\Psi(\lambda(z))\right\}. &
\end{array}$$

Combining the resulting inequality with (5.8), we come to the result as follows:

Proposition 5.2.2 *For the randomized, with parameter K, aggregate $\widetilde{f}(\cdot;z,\bar{j})$, the aggregation price can be bounded from above as*

$$\widetilde{\varepsilon}_n(K) \equiv \mathcal{E}_{z,\bar{j}}\left\{\Psi(\widetilde{\lambda}(z,\bar{j})) - \min_{\Lambda}\Psi(\cdot)\right\} \leq O(1)\frac{(L^2 + L\sigma)\sqrt{\ln M}}{\sqrt{n}} + \frac{L^2}{K}. \tag{5.19}$$

In particular, choosing K as the smallest integer which is $\geq \sqrt{\frac{n}{\ln M}}$, we get a randomized aggregation routine which is "$\sqrt{n}$-concentrated" – the resulting aggregate always is combination of at most $K \leq \sqrt{n}$ of the functions $f_1, ..., f_M$, and the aggregation price of the routine, up to factor 2, is the same as for our basic aggregation routine, see Theorem 5.2.1.

5.3 Lower bound

We have seen that when aggregating M functions on the basis of n observations (5.2), the expected aggregation price

$$\mathcal{E}\left\{\Psi(\lambda(\cdot)) - \min_{\Lambda}\Psi(\cdot)\right\}, \qquad \Psi(\lambda) = \int\limits_X \left(f(x) - \sum_{j=1}^{M}\lambda_j f_j(x)\right)^2 \mu(dx)$$

can be made as small as $O(\sqrt{\ln M}n^{-1/2})$. We are about to demonstrate that this bound is optimal in order in the minimax sense.

Theorem 5.3.1 *For appropriately chosen absolute constant $\kappa > 0$ the following is true.*

Let positive L, σ and integer $M \geq 3$ be given, and let n be a positive integer such that

$$\frac{\sigma^2 \ln M}{L^2} \leq n \leq \kappa\frac{\sigma^2 M \ln M}{L^2}. \tag{5.20}$$

For every aggregation routine $\mathcal{B}$ solving the Aggregation problem $\mathbf{C}$ on the basis of n observations (5.2) one can point out

- *M continuous functions $f_1, ..., f_M$ on the segment $[0,1]$ not exceeding L in absolute value,*
- *a function f which is a convex combination of the functions $f_1, ..., f_M$,*

with the following property. Let

$$\widehat{f}^{\mathcal{B}}(\cdot; z) = \sum_{j=1}^{M} \lambda_j^{\mathcal{B}}(z) f_j(\cdot)$$

be the aggregate yielded by the routine $\mathcal{B}$ as applied to the Aggregation problem with the data given by

– f_j, $j = 1, ..., M$, as the basic functions,

– the uniform distribution on $X = [0,1]$ as the distribution μ of the observation points,

– the $\mathcal{N}(0, \sigma^2)$ observation noises e_t,

– f as the "true" function,

and

– the simplex (5.18) as Λ.

The expected aggregation price of the aggregate $\widehat{f}^{\mathcal{B}}$ can be bounded from below as

$$\mathcal{E}\left\{\Psi(\lambda^{\mathcal{B}}) - \min_{\Lambda} \Psi(\cdot)\right\} = \mathcal{E}\left\{\Psi(\lambda^{\mathcal{B}})\right\} \geq \kappa \frac{L\sigma\sqrt{\ln M}}{\sqrt{n}}. \tag{5.21}$$

In particular, under assumption (5.20) the aggregation price associated with the routines from Section 5.2.2 is optimal in order, in the minimax sense, provided that $L = O(1)\sigma$.

Proof. Let $M \geq 3$, and let

$$f_j(x) = L\cos(2\pi j x), \; j = 1, ..., M.$$

Given a positive integer $p \leq M/2$, let us denote by $\mathcal{F}_p$ the set of all convex combinations of the functions $f_1, ..., f_M$ with the coefficients as follows: $2p$ of the coefficients are equal to $(2p)^{-1}$ each, and the remaining coefficients vanish.

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

I. Every two distinct functions from $\mathcal{F}_p^*$ have at most p common nonzero coefficients in the basis $f_1, ..., f_M$, so that

$$\frac{L^2}{4p} \leq \| f - g \|_2^2 \leq \frac{L^2}{2p} \tag{5.22}$$

(note that $f_1, ..., f_M$ are mutually orthogonal in $L_2[0,1]$ and that $\| f_j \|_2^2 = \frac{1}{2}$);

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

$$K \geq M^{\kappa_1 p} \tag{5.23}$$

(from now on, $\kappa_i > 0$ are appropriate absolute constants).

Now let

$$\varepsilon(p) = \max_{f\in\mathcal{F}_p^*}\left[\mathcal{E}\left\{\Psi_f(\lambda_f^{\mathcal{B}}) - \min_{\Lambda}\Psi_f(\cdot)\right\}\right] = \max_{f\in\mathcal{F}_p^*}\left[\mathcal{E}\left\{\Psi_f(\lambda_f^{\mathcal{B}})\right\}\right], \tag{5.24}$$

where

$$\Psi_f(\lambda) = \| f - \sum_{j=1}^{M}\lambda_j f_j \|_2^2$$

and $\lambda_f^{\mathcal{B}}$ is the vector of aggregation weights yielded by the aggregation routine $\mathcal{B}$, the observations being associated with f. Note that the second equality in (5.24) comes from the fact that Λ is the simplex (5.18) and all $f \in \mathcal{F}_p$ are convex combinations of $f_1, ..., f_M$.

We claim that if $p \leq \sqrt{M}$, then, for properly chosen κ_2, the following implication holds true:

$$\varepsilon(p) < \frac{L^2}{64p} \Rightarrow n \geq \kappa_2 \frac{\sigma^2 p^2 \ln M}{L^2}. \tag{5.25}$$

Note that (5.25) implies the conclusion of the Theorem. Indeed, believing in (5.25), choosing

$$p = \rfloor \frac{L}{\sigma}\sqrt{\frac{n}{\kappa_2 \ln M}} \lfloor$$

and taking into account that in the case of (5.20) the resulting p is $\leq \sqrt{M}$, provided that κ is chosen properly, we see that the conclusion in (5.25) fails to be true, so that

$$\varepsilon(p) \geq \frac{L^2}{64p} \geq O(1)\frac{\sigma L \sqrt{\ln M}}{\sqrt{n}};$$

the latter inequality, in view of the origin of $\varepsilon(p)$, is exactly what we need.

It remains to prove (5.25), which can be done by our standard information-based considerations. Indeed, let p satisfy the premise in (5.25), and let $\mathcal{B}$ be a method for solving the Aggregation problem with the data we have built. Let us associate with $\mathcal{B}$ a method $\mathcal{B}'$ for distinguishing between K hypotheses H_ℓ, $\ell = 1, ..., K$, on the distribution of the observation (5.2), ℓ-th of them saying that the observations are associated with ℓ-th signal f^ℓ from $\mathcal{F}_p^*$. Namely, given observations z, we call $\mathcal{B}$ to solve the Aggregation problem; after the corresponding aggregated estimate $F_{\mathcal{B}} = f_{\mathcal{B}}(z)$ is obtained, we find the $\|\cdot\|_2$-closest to $f_{\mathcal{B}}$ function f^ℓ in $\mathcal{F}_p^*$ (if there are several functions of this type, we choose, say, the first of them) and accept the hypotheses H_ℓ.

Since the pairwise $\|\cdot\|_2$-distances between the signals from $\mathcal{F}_p^*$ are $\geq d \equiv L/\sqrt{4p}$ by (5.22), and for every $f \in \mathcal{F}_p^*$ it holds $\mathcal{E}\{\| f_{\mathcal{B}} - f \|_2^2\} \leq \varepsilon(p)$ by (5.24), we see that the probability to reject hypothesis H_ℓ if it is true is, for every $\ell = 1, ..., K$, at most $\sqrt{\varepsilon(p)}/(d/2) \leq 1/4$. On the other hand, it is immediately seen that

> (!) The Kullback distance between every pair of distributions associated with our K hypotheses does not exceed
>
> $$\mathcal{K} \equiv \frac{n}{2\sigma^2}\max_{f,g\in\mathcal{F}_p^*} \| f - g \|_2^2 \leq \frac{nL^2}{4p\sigma^2}. \tag{5.26}$$

Indeed, let $f, g \in \mathcal{F}_p^*$ and F_f^n, F_g^n be the corresponding distributions of observations (5.2). Since the entries z_t are independent identically distributed, we have

$$\begin{array}{rcl}\mathcal{K}(F_f^n : F_g^n) & = & n\mathcal{K}(F_f^1 : F_g^1) \\ & = & n\int\limits_0^1 dx \left\{ \int\limits_{-\infty}^{\infty} \psi(t-f(x)) \ln \frac{\psi(t-f(x))}{\psi(t-g(x))} dt \right\} \\ & & \left[\psi(t) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{-t^2/(2\sigma^2)\}\right] \\ & = & \frac{n}{2\sigma^2} \int\limits_0^1 (f(x)-g(x))^2 dx \\ & = & \frac{n}{2\sigma^2} \parallel f - g \parallel_2^2,\end{array}$$

and we conclude that the Kullback distance between F_f^n and F_g^n does not exceed the quantity $\mathcal{K}$ defined in (5.26). The inequality in (5.26) is given by (5.22). □

Applying the Fano inequality (Theorem 1.2.1), we come to

$$\frac{nL^2}{4p\sigma^2} \geq \frac{3}{4} \ln(K-1) - \ln 2;$$

taking into account (5.23), we come to the conclusion of (5.25). ■

5.4 Application: Recovering functions from Barron's class

Usually, the "complexity" of approximating a multivariate function (e.g., the number of "simple terms" used in approximation) grows rapidly with dimensionality. This is why the Artificial Intelligence community was happy with the following "dimension-independent" result:

Theorem 5.4.1 [Barron '93 [1]] *Let $f : \mathbf{R}^d \to \mathbf{R}$ be the Fourier transform of a complex-valued measure of variation 1:*

$$f(x) = \int \exp\{i\omega^T x\} F(d\omega), \qquad \int |F(d\omega)| \leq 1,$$

and let μ be a probability distribution on $\mathbf{R}^d$. Then for every $n \geq 1$ there exists an n-term sum of cosines

$$\widetilde{f}(x) = \sum_{j=1}^{n} a_j \cos(\omega_j^T x + \phi_j)$$

such that

$$\int |\widetilde{f}(x) - f(x)|^2 \mu(dx) \leq \frac{1}{n}.$$

In fact, this theorem goes back to Maurey. In order to simplify notation, assume that

$$\int |F(d\omega)| = 1,$$

so that

$$\nu(d\omega) = |F(d\omega)|$$

is a probability distribution. Let

$$p(\omega) = \frac{F(d\omega)}{\nu(d\omega)}$$

be the density of complex-valued measure F with respect to the probability measure ν, and let $g(\cdot)$ be random element of the space $L_2(\mathbf{R}^d, \mu)$ of complex-valued μ-square summable functions on $\mathbf{R}^n$ given by

$$g_\omega(x) = p(\omega)\exp\{i\omega^T x\},$$

ω being distributed according to ν.

The expected value of the random function $g_\omega(\cdot)$ clearly is f, while the second moment of the $L_2(\mathbf{R}^d, \mu)$-norm of this random function does not exceed 1:

$$\mathcal{E}\left\{\int |g_\omega(x)|^2 \mu(d\omega)\right\} \leq 1,$$

since $\| g_\omega(\cdot) \|_\infty \leq 1$ and μ is a probabilistic measure.

It follows that if $\omega_1, ..., \omega_n$ is a sample of n independent random vectors ω_j distributed each according to ν, then

$$\mathcal{E}\left\{\int \left|\frac{1}{n}\sum_{j=1}^n g_{\omega_j}(x) - f(x)\right|^2 \mu(dx)\right\} = \frac{1}{n^2}\sum_{j=1}^n \mathcal{E}\left\{\int |g_{\omega_j}(x) - f(x)|^2 \mu(dx)\right\} \leq \frac{1}{n}$$

and, consequently, there exists a particular collection $\bar{\omega}_1, ...\bar{\omega}_n$ such that

$$\int \left| f(x) - \frac{1}{n}\sum_{j=1}^n g_{\bar{\omega}_j}(x)\right|^2 \mu(dx) \leq \frac{1}{n};$$

it suffices to take, as $\widetilde{f}(\cdot)$, the real part of the function

$$\frac{1}{n}\sum_{j=1}^n g_{\bar{\omega}_j}(x) = \frac{1}{n}\sum_{j=1}^n p(\bar{\omega}_j)\exp\{i\bar{\omega}_j^T x\}. \qquad \blacksquare$$

The advantage of Barron's result is that the quality of approximation in his theorem depends on the "number of simple terms" in the approximating aggregate and is independent of the dimensionality of the function to be approximated. A disadvantage of the construction is that in order to build the approximation, we need complete knowledge of F, or, which is the same, of f.

We are about to demonstrate that the aggregation technique developed in the previous section allows to build a "simple" approximation of f directly from its noisy observations, with basically no a priori knowledge of the function. Namely, assume that all our a priori knowledge about f is that f is the Fourier transform of a complex-valued measure of variation not exceeding a given upper bound $L/2 < \infty$ and vanishing outside the ball of a given radius R:

$$f(x) \in \mathcal{F}(L,R) = \left\{ f(x) = \int\limits_{\|\omega\|_2 \leq R} \exp\{i\omega^T x\}F(d\omega) \,\middle|\, \int |F(d\omega)| \leq L/2 \right\}. \qquad (5.27)$$

Besides this a priori knowledge, we are given n noisy observations

$$z = \{z_t = (x_t, y_t = f(x_t) + e_t)\}_{t=1}^n \tag{5.28}$$

of the values of f, where the observation points x_t are independent of each other and are distributed according to certain probability measure μ, and the observation noises e_t are independent of each other and of $\{x_t\}_{t=1}^n$ and have zero mean and bounded variance:

$$\mathcal{E}\{e_t\} = 0; \quad \mathcal{E}\left\{e_t^2\right\} \le \sigma^2. \tag{5.29}$$

We do not assume the measure μ to be known; all our a priori information on this measure is that

$$\int \| x \|_2^2 \, \mu(dx) \le \sigma_x^2 \tag{5.30}$$

with certain known in advance $\sigma_x < \infty$.

In order to recover f via observations (5.28), we act as follows:

Initialization. Given $\sigma, n, d, L, R, \sigma_x$, we set

$$\begin{array}{llll} (a) & \eta & = & \frac{\sqrt{L^2+L\sigma}}{n^{1/4}}, \\ (b) & \varepsilon & = & \frac{2\eta}{L\sigma_x} = \frac{2\sqrt{L^2+L\sigma}}{n^{1/4}L\sigma_x} \end{array} \tag{5.31}$$

and build an ε-net $\Omega = \{\omega_k\}_{k=1}^K$ in the ball $W_R = \{\omega \in \mathbf{R}^d \mid \| \omega \|_2 \le R\}$.

It is easily seen that the cardinality K of the net can be chosen to satisfy the bound

$$K \le (1 + 2\varepsilon^{-1}R)^d. \tag{5.32}$$

Estimation. We set $M = 2K$, $\Lambda = \{\lambda \in \mathbf{R}^M \mid \| \lambda \|_1 \le 1\}$ and define the basic functions f_j, $j = 1, ..., M$, as

$$f_{2k-1}(x) = L\cos(\omega_k^T x), \; f_{2k}(x) = L\sin(\omega_k^T x), \; k = 1, 2, ..., K.$$

Then we use the aggregation routine from Section 5.2.2 to get "nearly closest to f" weighted combination

$$\widehat{f}_n(\cdot; z) = \sum_{j=1}^M \lambda_j(z) f_j(\cdot) \qquad [\sum_{j=1}^M |\lambda_j(z)| \le 1]$$

of functions f_j and treat this combination as the resulting approximation of f.

Remark 5.4.1 Applying our "n-concentration" technique, we can enforce $\widehat{f}_n$ to be a weighted sum of at most $n + 2$ cosines, similarly to the approximation given by Barron's Theorem.

The rate of convergence of the outlined approximation scheme is given by the following

Theorem 5.4.2 *Let $f \in \mathcal{F}(L, R)$ and let (5.29), (5.30) be satisfied. Then for all n one has*

$$\mathcal{E}\left\{\| \widehat{f}_n(\cdot, z) - f(\cdot) \|_{2,\mu}^2\right\} \le O(1)\frac{(L^2 + L\sigma)\sqrt{d\ln M_n}}{\sqrt{n}}, \quad M_n = 2 + \frac{n^{1/4}LR\sigma_x}{\sqrt{L^2 + L\sigma}}. \tag{5.33}$$

Proof. 1^0. Let us verify that for every $f \in \mathcal{F}(L,R)$ there exists a function

$$\widetilde{f}(x) = \sum_{j=1}^{M} \lambda_j f_j(x)$$

with $\lambda \in \Lambda$ such that

$$\| f - \widetilde{f} \|_{2,\mu} \leq \eta. \tag{5.34}$$

Indeed, we have

$$f(x) = \int\limits_{W_R} \exp\{i\omega^T x\} F(d\omega) \text{ with } \int\limits_{W_R} |F(d\omega)| \leq L/2.$$

Since Ω is an ε-net in W_R, we can partition W_R into K non-overlapping sets Ω_k in such a way that $\omega_k \in \Omega_k$ and Ω_k is contained in the ball of radius ε centered at ω_k, for all k. Setting

$$\begin{array}{rcl} p_k & = & \int\limits_{\Omega_k} F(d\omega) = a_k + b_k i, \; k = 1, ..., K, \\ \widetilde{f} & = & \sum\limits_{k=1}^{K} \Re\left\{ p_k \exp\{i\omega_k^T x\}\right\} = \sum\limits_{k=1}^{K} \left[\lambda_{2k-1} f_{2k-1}(x) + \lambda_{2k} f_{2k}(x)\right], \\ & & \lambda_{2k-1} = \frac{1}{L} a_k, \\ & & \lambda_{2k} = -\frac{1}{L} b_k, \end{array}$$

we get

$$\sum_{j=1}^{M} |\lambda_j| \leq \sqrt{2} L^{-1} \sum_{k=1}^{K} |p_k| \leq \int\limits_{W_R} |F(d\omega)| \leq 1$$

and

$$\begin{array}{rcl} |\widetilde{f}(x) - f(x)| & \leq & \left| \sum\limits_{k=1}^{K} p_k \exp\{i\omega_k^T x\} - f(x) \right| \quad \text{[since } f \text{ is real-valued]} \\ & = & \left| \sum\limits_{k=1}^{K} \int\limits_{\Omega_k} \left[\exp\{i\omega^T x\} - \exp\{i\omega_k^T x\}\right] F(d\omega) \right| \\ & \leq & \sum\limits_{k=1}^{K} \int\limits_{\Omega_k} \left|\exp\{i\omega^T x\} - \exp\{i\omega_k^T x\}\right| |F(d\omega)| \\ & \leq & \varepsilon \| x \|_2 \sum\limits_{k=1}^{K} \int\limits_{\Omega_k} |F(d\omega)| \quad \text{[since } |\omega - \omega_k| \leq \varepsilon \; \forall \omega \in \Omega_k\text{]} \\ & \leq & \varepsilon \| x \|_2 \, L/2 \\ \Rightarrow \; \| \widetilde{f} - f \|_{2,\mu} & \leq & 0.5 \varepsilon L \sigma_x \\ & = & \eta \quad \text{[see (5.31.}(b)\text{)]} \end{array}$$

as required.

2^0. Applying Theorem 5.2.1, we get

$$\begin{array}{rcl} \mathcal{E}\left\{ \| f(\cdot) - \widehat{f}_n(\cdot; z) \|_{2,\mu}^2 \right\} & \leq & O(1) \frac{(L^2 + L\sigma)\sqrt{\ln M}}{\sqrt{n}} + \min\limits_{\lambda \in \Lambda} \| f - \sum\limits_{j=1}^{M} \lambda_j f_j \|_{2,\mu}^2 \\ & \leq & O(1) \frac{(L^2 + L\sigma)\sqrt{\ln M}}{\sqrt{n}} + \| f - \widetilde{f} \|_{2,\mu}^2, \end{array}$$

which combined with (5.34) implies that

$$\mathcal{E}\left\{\| f(\cdot) - \widehat{f}_n(\cdot; z) \|_{2,\mu}^2\right\} \leq O(1)\frac{(L^2 + L\sigma)\sqrt{\ln M}}{\sqrt{n}} + \eta^2.$$

It remains to note that $M \leq 2M_n^d$ by (5.32) and that $\eta^2 \leq \frac{L^2+L\sigma}{\sqrt{n}}$ by (5.31.(a)). ■

Discussion. Theorem 5.4.2 establishes "nearly dimension-independent" rate of convergence of approximations of a function $f \in \mathcal{F}(L, R)$ to the function: when all but the dimension parameters (i.e., σ, L, R, σ_x) are fixed, the rate of convergence (measured as $\mathcal{E}\left\{\| f - \widehat{f}_n \|_{2,\mu}^2\right\}$) is $O(\sqrt{dn^{-1}\ln n})$, so that the volume of observations required to approximate f within a given margin is just proportional to the dimension d. To understand that this linear growth indeed means "nearly dimension-independence" of the complexity of recovering a function, note that for the "usual" functional classes, like Sobolev and Hölder balls, the number of observations (even noiseless) needed to recover a function within a given inaccuracy grows with the dimension d like $\exp\{\alpha d\}$ ($\alpha > 0$ depends on the parameters of smoothness of the class in question). It should be stressed that the rate of convergence given by (5.33) is nearly independent of the parameters R, σ_x; we could allow these parameters to grow with n in a polynomial fashion, still preserving the $O(\sqrt{dn^{-1}\ln n})$-rate of convergence. By similar reasons, we would not loose much when replacing the assumption that the Fourier transform of f vanishes outside a given compact with bounds on the "tails" of this transform, thus coming to the classes like

$$\begin{array}{rcl}\mathcal{F}(L,\gamma) & = & \{f = \int \exp\{i\omega^T x\}F(d\omega)\Big| \int |F(d\omega)| \leq L, \\ & & \int\limits_{\|\omega\|_2 > R} |F(d\omega)| \leq R^{-\gamma}\ \forall R > 0\}.\end{array}$$

As compared to the original result of Barron, the result stated by Theorem 5.4.2 has, essentially, only one drawback: the rate of convergence (5.33) is nearly $O(n^{-1/2})$, while in Barron's theorem the rate of convergence is $O(n^{-1})$. This "slowing down" is an unavoidable price for the fact that Theorem 5.4.2 deals with the case of approximating *unknown* function from Barron's-type class. In this case, the convergence rate $O(n^{-1/2})$ is nearly optimal in the minimax sense, as stated by the following result of [16]:

Theorem 5.4.3 *Let $L > 0$. Consider the problem of estimating a univariate function $f : \mathbf{R} \to \mathbf{R}$ via observations* (5.28)*, where x_t are uniformly distributed on* $[0, 1]$ *and $e_t \sim \mathcal{N}(0, \sigma^2)$. Let $\mathcal{F}_n$ be the class of all real-valued trigonometric polynomials of degree $\leq n$ with the sum of absolute values of the coefficients not exceeding L. Then, for appropriately chosen absolute constant $\kappa > 0$ and for all large enough values of n, for every algorithm $\mathcal{B}$ approximating $f \in \mathcal{F}_n$ via n associated with f observations* (5.28) *it holds*

$$\sup_{f \in \mathcal{F}_n} \mathcal{E}\left\{\| f - \widehat{f}_{\mathcal{B}} \|_2^2\right\} \geq \kappa L\sigma\sqrt{\frac{\ln n}{n}}; \tag{5.35}$$

here $\widehat{f}_{\mathcal{B}}$ is the estimate yielded by $\mathcal{B}$, the function underlying observations being f.

5.5 Numerical example: nonparametric filtration

Following [16], consider a nonlinear time-invariant dynamic system:

$$y_t = f(y_{t-1}, y_{t-2}, ..., y_{t-d}) + e_t, \tag{5.36}$$

$e_0, e_1, ...$ being independent noises. We do not know f, and our target is to predict, given $y_0, ..., y_n$, the state y_{n+1}.

A natural way to approach our target is to recover f from observations and to form the prediction as

$$y^{\mathrm{p}}_{n+1} = \widehat{f}_n(y_n, ..., y_{n-d+1}), \tag{5.37}$$

$\widehat{f}_n$ being the estimate of f built upon the first n observations (5.36). Setting $x_t = (y_{t-1}, ..., y_{t-d})^T$, we can represent the observations accumulated at time instant n as

$$z = \{z_t = (x_t, y_t = f(x_t) + e_t)\}_{t=d}^n. \tag{5.38}$$

The situation resembles the observation scheme (5.2), up to the fact that now the points x_t where we observe f depend on each other in a complicated and unknown fashion rather than to be i.i.d. Let us ignore this "minor difference" (we are not going to *prove* anything, just to *look* how it works) and act as if $\{x_t\}$ were i.i.d.

Assume that the dynamic system in question is known to be *semilinear* ("a system with single output nonlinearity"):

$$f(x) = \phi(p^T x).$$

If p were known, we could project our observation points x_t onto the corresponding axis, thus reducing the situation to the one where we are observing a *univariate* function ϕ. As a result, we would be capable to recover the multivariate function f as if it were a univariate function. In the case when p is unknown (this is the case we are interested in) it makes sense to use the approach outlined in Section 5.1, namely, to choose a "fine finite grid" Π in the space of d-dimensional directions and to associate with every direction $p \in \Pi$ the estimate $\widehat{f}_p$ of f corresponding to the hypothesis that the "true" direction is p. We can use, say, the first half of our n observations to build the associated realizations f_p, $p \in \Pi$, of our estimates, and use the remaining half of observations to aggregate the resulting basic estimates, as described in Section 5.2.2, thus coming to the aggregated estimate $\widehat{f}_n$ to be used in the predictor (5.37).

We are about to present the results yielded by the just outlined scheme as applied to systems of the type

$$(\mathcal{D}_d) : \left\{ \begin{array}{rcl} y_t & = & F(p^T x) + \sigma\eta_t,\ x_t^T = (y_{t-1}, ..., y_{t-d}), \\ F(z) & = & \cos(4\pi z) + \cos(5\pi z), \\ \eta_t & \sim & \mathcal{N}(0,1), \\ p & = & d^{-1/2}(1, ..., 1)^T \in \mathbf{R}^d. \end{array} \right.$$

In our simulations, we dealt with the dynamics $(\mathcal{D}_d)$ with $d = 2, 3$. In the case of $d = 2$, the grid Π of directions was

$$\left\{ p_i = \begin{pmatrix} \cos(\phi_0 + jM^{-1}\pi) \\ \sin(\phi_0 + jM^{-1}\pi \end{pmatrix} \right\}_{j=1}^M,$$

ϕ_0 being a randomly chosen "phase shift"; we used $M = 400$. In the case of $d = 3$, the grid Π was comprised of $M = 3144$ randomly generated directions in $\mathbf{R}^3$. In both cases, the basic estimates f_p were the zero order spatial adaptive estimates from Chapter 3 (modified in an evident manner to get the possibility to work with non-equidistant grids of observation points).

In our experiments, we used the first 1024 observations z_t to build the basic estimates, the next 1024 observations to aggregate these estimates by the aggregation routine from Section 5.2.2, the underlying set Λ being the standard simplex

$$\{\lambda \in \mathbf{R}^M \mid \lambda \geq 0, \sum_j \lambda_j = 1\},$$

and used the resulting predictor (5.37) at 2048 subsequent time instants in order to measure the empirical standard deviation

$$\delta = \sqrt{\frac{1}{2048} \sum_{t=2049}^{4096} (f(x_t) - y_t^{\mathrm{p}})^2}.$$

In order to understand what is the effect of our "structure-based" prediction scheme – one which exploits the a priori knowledge that the actual dynamics is semilinear, we have compared its performance with the one of the "standard" prediction scheme based on the zero order spatial adaptive non-parametric recovering of f (treated as a "general-type" function of d variables) from the first 2048 observations (5.38).

The results of the experiments are as follows:

Method	$\sigma = 0.1$	$\sigma = 0.33$
Structure-based predictor, dynamics $(\mathcal{D}_2)$	0.093	0.275
Standard predictor, dynamics $(\mathcal{D}_2)$	0.483	0.623
Structure-based predictor, dynamics $(\mathcal{D}_3)$	0.107	0.288
Standard predictor, dynamics $(\mathcal{D}_3)$	0.244	1.013

Empirical standard deviation

The histograms of the prediction errors $f(x_t) - y_t^{\mathrm{p}}$ and typical prediction patterns are as follows: Finally, this is how the function f itself was recovered in the case of dynamics $(\mathcal{D}_2)$:

Structure-based predictor, $\sigma = 0.1$ Standard predictor, $\sigma = 0.1$

Structure-based predictor, $\sigma = 0.33$ Standard predictor, $\sigma = 0.33$

Figure 5.1: Distribution of prediction errors, dynamics ($\mathcal{D}_2$).

Structure-based predictor, $\sigma = 0.1$ Standard predictor, $\sigma = 0.1$

Structure-based predictor, $\sigma = 0.33$ Standard predictor, $\sigma = 0.33$

Figure 5.2: Distribution of prediction errors, dynamics $(\mathcal{D}_3)$.

Structure-based predictor, $\sigma = 0.33$ Standard predictor, $\sigma = 0.33$

Figure 5.3: Prediction patterns, dynamics $(\mathcal{D}_2)$.
[circles: $f(x_t)$; crosses: y_t^{p}]

Structure-based predictor, $\sigma = 0.1$ Standard predictor, $\sigma = 0.1$

Figure 5.4: Prediction patterns, dynamics $(\mathcal{D}_3)$.
[circles: $f(x_t)$; crosses: y_t^{p}]

Dynamics $(\mathcal{D}_2)$

Structure-based reconstruction Standard reconstruction

Figure 5.5: Reconstructions of dynamics $\mathcal{D}_2$, $\sigma = 0.1$.

Dynamics ($\mathcal{D}_2$)

Structure-based reconstruction Standard reconstruction

Figure 5.6: Reconstructions of dynamics $\mathcal{D}_2$, $\sigma = 0.33$.

Chapter 6

Aggregation of estimates, II

We proceed with aggregating estimates associated with a number of "concurrent hypotheses" on the observed regression function. In the previous chapter our goal was, essentially, to reproduce the best convex combination of the estimates, while now we focus on reproducing the best of the estimates or their best linear combination.

6.1 Gaussian white noise model of observations

It makes sense now to switch from the "discrete" models of observations we dealt with to the moment to the "continuous" model. In this new model, a signal $f : [0, 1] \to \mathbf{R}$ is observed in continuous time Gaussian white noise of intensity ε^2. In other words, our observation is the random function

$$y(x) = y_{f,\varepsilon}(x) = \int_0^x f(s)ds + \varepsilon W(x), \tag{6.1}$$

$W(x)$ being the standard Wiener process.

Model (6.1) is very popular in Nonparametric Statistics by the reasons as follow. There exists a rich "L_2 regression theory", where the quality of restoring f is measured in the L_2 norm and a priori assumptions on the signal are expressed in geometric form – usually, as hypotheses on the rate at which f can be approximated by elements of a given sequence of finite-dimensional subspaces $E_1 \subset E_2 \subset ...$ of L_2. A typical example is a periodic with derivatives of order $< k$, of the period 1, signal from the Sobolev ball $\mathbf{S}_1^{k,2}(L)$:

$$\int_0^1 (f^{(k)}(x))^2 dx \leq L^2.$$

The indicated properties of f are equivalent to the fact that

$$\sum_{j=1}^{\infty} (2\pi j)^{2k} [f_{2j-1}^2 + f_{2j}^2] \leq L^2, \tag{6.2}$$

where $\{f_j\}_{j=0}^{\infty}$ are the Fourier coefficients of f in the standard trigonometric orthonormal basis of $L_2[0, 1]$

$$\phi_0(x) \equiv 1, \phi_{2j-1}(x) = \sqrt{2}\cos(2\pi jx), \phi_{2j}(x) = \sqrt{2}\sin(2\pi jx),\ j = 1, 2, ...$$

Note that (6.2) is just a way to fix the rate at which f can be approximated, in the L_2-metric, by a trigonometric polynomial of degree j.

As far as the L_2 regression theory is concerned, (6.1) definitely is the most convenient model of observations, since it admits a very transparent and simple "translation" to the language of the L_2-geometry. As a result, with this model we get a "convenient road" to a number of interesting and instructive results. Now the role of "volume of observations" n is played by the quantity ε^{-2}; instead of asking "how well can we recover a signal from a large number n of noisy observations of the signal", we now ask how well we can recover a signal affected by Gaussian white noise of small intensity ε^2.

"Scientific practice" demonstrates that the majority of asymptotic, $\varepsilon \to 0$, results of the L_2 regression theory with observations (6.1) can as well be established (under appropriate technical assumptions) for more (or less?) realistic discrete models of observations like the one where we observe the signal along an equidistant (or random) n-point grid, variance of the noise affecting a particular observation being σ^2. The "translation" of the results obtained for the continuous model of observations to those for the discrete model is given by the correspondence $\sigma^2 n^{-1} = \varepsilon^2$. Which one of these models to use, it is, essentially, the question of mathematical convenience, and in our course we have reached the point when it definitely is easier to deal with model (6.1).

L_2 regression theory: the language. It is well-known that observations (6.1) are equivalent to the possibility to observe the $L_2[0,1]$-inner products of the signal f with functions $\phi \in L_2[0,1]$. Namely, given a function $\phi \in L_2[0,1]$, one can convert a realization of observation (6.1) in a realization of the random variable

$$\int_0^1 \phi(x)dy(x) = (f,\phi) + \varepsilon\xi_\phi, \tag{6.3}$$

where

$$(f,g) = \int_0^1 f(x)g(x)dx$$

is the standard inner product in $L_2[0,1]$. It turns out that the vector of random noises $\{\xi_{\phi_i}\}_{i=1}^k$ corresponding to every finite collection of $\phi_i \in L_2$ is Gaussian, and its covariance matrix is just the Gram matrix of $\phi_1, ..., \phi_k$:

$$\mathcal{E}\left\{\xi_\phi \xi_\psi\right\} = (\phi,\psi) \quad \forall \phi,\psi \in L_2. \tag{6.4}$$

It should be mentioned that for every $\phi \in L_2$ the left hand side in (6.3) is well-defined with probability one, the probability space in question being generated by the underlying Wiener process; thus, it makes no sense to speak simultaneously about values of *all* random noises $\{\xi_\phi \mid \phi \in L_2\}$, but it does make sense to speak about values of any countable collection from this set, and this is the only situation we shall deal with.

The outlined properties of model (6.1) allow to pass from the "functional" language to the geometric one and to represent the situation we are interested in as follows. We fix a real separable Hilbert space H with inner product $(\cdot,\cdot)$ and the associated norm $\|\cdot\|$; the "signals" we are observing are just the elements of this space. An observation y of a signal $f \in H$ is comprised of noisy measurements

$$\{y_\phi(f,\varepsilon) = (f,\phi) + \varepsilon\xi_\phi\}_{\phi \in H} \tag{6.5}$$

of the projections of f on vectors from H, all finite collections of the noises ξ_ϕ being Gaussian random vectors with covariance matrices given by (6.4). In (6.5), ε is given "noise intensity"[1)] Note that a sufficient statistics for (6.5) is given already by the sequence of observations

$$y^{f,\varepsilon} = \left\{ y_i^{f,\varepsilon} \equiv y_{\phi_i}(f,\varepsilon) = (f,\phi_i) + \varepsilon\xi_i \equiv (f,\phi_i) + \varepsilon\xi_{\phi_i} \right\} \tag{6.6}$$

associated with a fixed orthonormal basis $\{\phi_i\}_{i=1}^\infty$ of H; given these observations, we can recover $y_\phi(f,\varepsilon)$ for every $\phi \in H$ according to

$$y_\phi(f,\varepsilon) = \sum_{i=1}^\infty (\phi,\phi_i) y_{\phi_i}(f,\varepsilon).$$

Thus, in fact our observations are just noisy observations of the coordinates of the signal in a somehow fixed orthonormal basis of H, the noises in the observations forming a sequence of independent $\mathcal{N}(0,\varepsilon^2)$ random variables.

Our goal is to recover signal f from the associated observations. A *recovering routine* $\widehat{f}(\cdot)$ is a Borel mapping acting from the set $\mathbf{R}^{\mathbf{Z}}$ of sequences with real entries to H, the set of sequences being equipped with the usual Tikhonov topology of the direct product (in this topology, $\mathbf{R}^{\mathbf{Z}}$ is a Polish space). The reconstruction associated with observations (6.5) is the random vector

$$\widehat{f}(y^{f,\varepsilon}) \in H,$$

where $y^{f,\varepsilon}$ are given by (6.6), $\{\phi_i\}$ being a fixed orthonormal basis in H (it does not matter how we choose this basis: as explained above, the observations associated with a basis can be converted to those associated with any other basis).

Given noise intensity ε, we measure the quality of a recovering routine $\widehat{f}$ *at a signal* $f \in H$ by the quantity

$$\mathcal{R}_\varepsilon(\widehat{f}, f) = \left(\mathcal{E}\left\{ \| \widehat{f}(y^{f,\varepsilon}) - f \|^2 \right\} \right)^{1/2}, \tag{6.7}$$

the expectation being taken over the observation noise. Given a subset $F \subset H$, we measure the quality of the routine $\widehat{f}$ *on the set* F by the corresponding worst-case risk

$$\mathcal{R}_\varepsilon(\widehat{f}, F) = \sup_{f\in F} \mathcal{R}_\varepsilon(\widehat{f}, f). \tag{6.8}$$

The *minimax risk* associated with F is the function

$$\mathcal{R}^*(\varepsilon, F) = \inf_{\widehat{f}} \mathcal{R}_\varepsilon(f, F) = \inf_{\widehat{f}} \sup_{f\in F} \mathcal{R}_\varepsilon(\widehat{f}, f). \tag{6.9}$$

Finally, an *estimation method* is a family $\{\widehat{f}_\varepsilon\}_{\varepsilon>0}$ of recovering routines parameterized by the noise intensity; we say that such a method is asymptotically optimal/optimal in order in the minimax sense on a set $F \subset H$, if

$$\mathcal{R}_\varepsilon(\widehat{f}_\varepsilon, F) \le C(\varepsilon)\mathcal{R}^*(\varepsilon, F)$$

where $C(\varepsilon)$ converges to 1, respectively, remains bounded as $\varepsilon \to +0$.

[1)] In the standard terminology, the intensity of noise in (6.5) is ε^2 rather than ε. In order to get a name for the quantity ε, we prefer to call it, and not its square, the intensity of noise.

6.2 Approximating the best linear combination of estimates

The problem. Assume we observe signals from a separable Hilbert space H according to (6.6) and are given a collection

$$\mathcal{M} = \left\{ \mathcal{M}^j = \{\widehat{f}^j_\varepsilon\}_{\varepsilon>0},\ j = 1, ..., M \right\}$$

of M estimation methods. For every signal $f \in H$, let

$$\mathcal{M}_\varepsilon(f, y) = \min_{\mu\in\mathbf{R}^M} \parallel f - \sum_j \mu_j \widehat{f}^j_\varepsilon(y) \parallel$$

be the distance from f to the linear span of the estimates $\widehat{f}^j_\varepsilon(y)$. When aggregating the given estimates in a linear fashion (however, with the weights which may depend on observations), and being clever enough to find, for every signal f underlying observations and every sequence of observation noises, the best – the closest to f – "mixture" of this type, we would recover f with inaccuracy $\mathcal{M}_\varepsilon(f, y)$; the risk of this "ideal linear aggregation" would be

$$\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, f) = \left(\mathcal{E}\left\{ \mathcal{M}^2_\varepsilon(f, y^{f,\varepsilon}) \right\} \right)^{1/2}. \tag{6.10}$$

The problem we are about to address is as follows:

> Aggregation problem **L**. *Given a collection $\mathcal{M}$ of M estimation methods, find an estimation method with the risk, at every $f \in H$, "close" to the risk $\mathcal{R}^{LA}_{\mathcal{M}}(\varepsilon, f)$ of the "ideal" linear aggregation of the methods from $\mathcal{M}$.*

A solution: the idea. The problem we are interested in admits an extremely simple (and, as we shall see in a while, quite powerful) "solution" as follows. Assume we observe a signal $f \in H$ *twice*, so that we have two realizations y', y'' of observation $y^{f,\cdot}$, the noises affected the realizations being independent of each other; let the intensities of noise in y' and y'' be $\varepsilon', \varepsilon''$, respectively.

Let us use the first realization of observations to build the M estimates $f^j = \widehat{f}^j_{\varepsilon'}(y')$, $j = 1, ..., M$. Consider the linear span

$$L = L(y') = \left\{ g = \sum_{j=1}^M \mu_j f^j \mid \mu \in \mathbf{R}^M \right\} \subset H$$

of these estimates; this is a random linear subspace of H of dimension not exceeding M. To simplify notation, assume that this dimension almost surely is equal to M (what follows can be modified in an evident fashion to capture the general case as well). Applying the orthogonalization process, we may build a basis in L comprised of M *orthonormal* vectors $h^1, ..., h^M$; these vectors are deterministic functions of y'.

Now let us use the second observation, y'', to evaluate the orthogonal projection f_L of f onto $L = L(y')$. The orthogonal projection itself is given by

$$f_L = \sum_{j=1}^M (f, h^j)h^j, \tag{6.11}$$

and is the closest to f linear combination of $f^j = \widehat{f}^j_{\varepsilon'}(y')$:

$$\parallel f - f_L \parallel^2 = \mathcal{M}^2_{\varepsilon'}(f, y'). \tag{6.12}$$

Observation y'' provide us with noisy observations

$$z_j = (f, h^j) + \varepsilon''\xi''_j,$$

ξ''_j being independent of each other and of y' $\mathcal{N}(0,1)$ random noises. Using z_j in (6.11) instead of the "true" Fourier coefficients (f, h^j), we come to the estimate

$$\widetilde{f} = \widetilde{f}(y', y'') = \sum_{j=1}^{M} z_j h^j = f_L + \varepsilon'' \sum_{j=1}^{M} \xi''_j h^j. \tag{6.13}$$

Let us evaluate the quality of the resulting estimate of f. We have

$$\parallel f - \widetilde{f} \parallel^2 = \parallel f - f_L \parallel^2 + 2\varepsilon''(f - f_L, \sum_{j=1}^{M} \xi''_j h^j) + (\varepsilon'')^2 M.$$

Taking expectation over noises affecting y', y'' and taking into account that ξ''_j are independent of y', we get

$$\mathcal{E}\left\{\parallel f - \widetilde{f} \parallel^2\right\} = \left(\mathcal{R}^{\rm LA}_{\mathcal{M}}(\varepsilon', f)\right)^2 + (\varepsilon'')^2 M, \tag{6.14}$$

whence, in particular,

$$\left(\mathcal{E}\left\{\parallel f - \widetilde{f} \parallel^2\right\}\right)^{1/2} \le \mathcal{R}^{\rm LA}_{\mathcal{M}}(\varepsilon', f) + \varepsilon''\sqrt{M}. \tag{6.15}$$

The simple result we have obtained looks as a "nearly solution" to the Aggregation problem **L**: in the right hand side of (6.15) we see the risk $\mathcal{R}^{\rm LA}_{\mathcal{M}}$ of the "ideal" linear aggregation (associated, however, with noise intensity ε' rather than ε), plus the "aggregation price" $\varepsilon''\sqrt{M}$. As we shall see, in many important cases this price is negligible small as compared to the risk of the ideal linear aggregation, so that (6.15) is, basically, what we need. There is, however, a difficulty: our estimation method requires two independent observations of the signal, while in our setting of the Aggregation problem we are allowed to use only one observation. We are about to demonstrate that this difficulty can be easily avoided – we always can "split" a single observation we have into two (or 1000) independent observations.

Splitting observations. Let us start with the following simple situation: we are given a realization ζ of an $\mathcal{N}(a, \sigma^2)$ random variable; σ is known, a is unknown. Can we "split" our observation ζ in a given number k of *independent* of each other realizations ζ_ℓ, $\ell = 1, ..., k$, of $\mathcal{N}(a, \sigma^2_\ell)$ random variables? What could be the corresponding variances σ^2_ℓ?

The answer is immediate: we claim that the required partitioning is possible, provided that

$$\frac{1}{\sigma^2} = \sum_{\ell=1}^{k} \frac{1}{\sigma^2_\ell}. \tag{6.16}$$

Indeed, we claim that under assumption (6.16) there exists a $k \times k$ matrix of the form

$$Q = \begin{pmatrix} 1 & q_{12} & q_{13} & \cdots & q_{1k} \\ 1 & q_{22} & q_{23} & \cdots & q_{2k} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & q_{k2} & q_{k3} & \cdots & q_{kk} \end{pmatrix}$$

such that for the rows $q_1, ..., q_k$ of the matrix it holds

$$q_j^T q_\ell = \frac{\sigma_j^2}{\sigma^2}\delta_{j\ell}, \tag{6.17}$$

$\delta_{j\ell}$ being the Kronecker symbols.

> Matrix Q can be built as follows. Indeed, let $e_1, ..., e_k$ be the standard basic orths in $\mathbf{R}^k$, and let
>
> $$r_j = \frac{\sigma_j}{\sigma}e_j,\ j = 1, ..., k; \quad \bar{u} = \left(\frac{\sigma}{\sigma_1}, ..., \frac{\sigma}{\sigma_k}\right)^T.$$
>
> By construction we have
>
> $$r_j^T \bar{u} = 1,\ j = 1, ..., k,$$
>
> and $\bar{u}$ is a unit vector by (6.16). Let us pass to an orthonormal basis of $\mathbf{R}^k$ where the first vector of the basis is $\bar{u}$, and let q_j be the vector of coordinates of r_j in this new basis. Then
>
> $$(q_j)_1 = r_j^T \bar{u} = 1 \quad \forall j \text{ and } q_j^T q_\ell = r_j^T r_\ell = \frac{\sigma_j^2}{\sigma^2}\delta_{j\ell}, \tag{6.18}$$
>
> as required.

Now assume that we are given $\sigma, \{\sigma_i\}_{i=1}^k$ satisfying (6.16), and a realization ζ of $\mathcal{N}(a, \sigma^2)$ random variable, and our goal is to "split" ζ in a sample of k *independent* $\mathcal{N}(a, \sigma_i^2)$ random variables $\zeta_1, ..., \zeta_k$. To this end let us build matrix Q satisfying (6.17), generate $k-1$ independent of each other and of ζ "artificial" $\mathcal{N}(0,1)$ random variables $\omega_1, ..., \omega_{k-1}$ and set

$$\begin{pmatrix} \zeta_1 \\ \zeta_2 \\ \cdots \\ \zeta_k \end{pmatrix} = Q \begin{pmatrix} \zeta \\ \sigma\omega_1 \\ \sigma\omega_2 \\ \cdots \\ \sigma\omega_{k-1} \end{pmatrix}. \tag{6.19}$$

From (6.17) combined with the fact that the first column of Q is comprised of ones it immediately follows that ζ_j, $j = 1, ..., k$, are independent of each other $\mathcal{N}(a, \sigma_j^2)$ random variables.

After we know how to split a single realization of an $\mathcal{N}(a, \sigma^2)$ random variable, we know how to split a single realization $y^{f,\varepsilon}$ of observation (6.6) into a desired number k of *independent* realizations y^{f,ε_i}, $i = 1, ..., k$, of the same signal with prescribed noise intensities $\sigma_1, ..., \sigma_k$ satisfying the "balance equation"

$$\frac{1}{\varepsilon^2} = \sum_{i=1}^k \frac{1}{\varepsilon_i^2} \tag{6.20}$$

– it suffices to apply the above randomized routine to every one of the observations $y_j^{f,\varepsilon}$, using, for every index $j = 1, 2, ...$, "its own" artificial random ω-variables. Thus, from the statistical viewpoint, we always may assume that instead of observation (6.6) we are given a desired number k of independent of each other similar observations, the noise intensities of the observations being linked by (6.20). From the implementation viewpoint, our "observation splitting" just means that we pass from deterministic recovering routines to randomized ones.

As a byproduct of our "splitting result", we see that as far as model (6.6) of observations is concerned, the quantity ε^{-2} indeed behaves itself as the "volume of observations": given an observation of "volume $n = \varepsilon^{-2}$", we can partition it into a prescribed number k of independent observations of prescribed volumes $n_k = \varepsilon_k^{-2}$, provided that $n = n_1 + ... + n_k$. And of course vice versa: given k independent observations y^{f,ε_i} of the same signal, we can aggregate them into a single observation y of the volume $n = \sum\limits_i n_i \equiv \sum\limits_i \frac{1}{\varepsilon_i^2}$: it suffices to set

$$y = \frac{1}{n_1 + ... + n_k} \sum_{i=1}^{k} n_i y^{f,\varepsilon_i}.$$

Finally, note that in the discrete model of observations similar "splitting" is given by "physical" splitting of observations, like partitioning all observations in subsequent segments, 5 observations per each, and putting the first observation from every segment to the first group, two next – to the second group, and two more – to the third one.

Intermediate summary. Let us come back to Aggregation problem **L**, and let us fix somehow a set $F \subset H$ of signals we are interested in, along with a collection $\mathcal{M} = \left\{\widehat{f}_\varepsilon^j(\cdot)\right\}_{\substack{\varepsilon>0\\ j=1,...,M}}$ of M estimation methods. Assume that

A. *The worst-case, over $f \in F$, risk*

$$\mathcal{R}_{\mathcal{M}}^{\mathrm{LA}}(\varepsilon, F) = \sup_{f\in F} \mathcal{R}_{\mathcal{M}}^{\mathrm{LA}}(\varepsilon, f)$$

of "ideal linear aggregation" of the estimation methods from $\mathcal{M}$ is a "well-behaved" function of ε as $\varepsilon \to 0$: whenever $\delta(\varepsilon) \ge \varepsilon$ is such that $\delta(\varepsilon)/\varepsilon \to 1$, $\varepsilon \to +0$, one has

$$\mathcal{R}_{\mathcal{M}}^{\mathrm{LA}}(\delta(\varepsilon), F) \le (1 + o(1))\mathcal{R}_{\mathcal{M}}^{\mathrm{LA}}(\varepsilon, F),\ \varepsilon \to 0. \tag{6.21}$$

B. *The worst-case, over $f \in F$, risk of "ideal linear aggregation" of the estimation methods from $\mathcal{M}$ is "non-parametric":*

$$\varepsilon^{-1}\mathcal{R}_{\mathcal{M}}^{\mathrm{LA}}(\varepsilon, F) \to \infty,\ \varepsilon \to 0. \tag{6.22}$$

Both of these assumptions are very natural. As about **B**, note that already the minimax risk of estimating *k-parametric* signals

$$f \in F_k(L) = \{f = \sum_{i=1}^{k} f_i\phi_i \mid \sum_{i=1}^{k} f_i^2 \le L^2\}$$

is $(1+o(1))\varepsilon\sqrt{k}$ as $\varepsilon \to +0$. As about **A**, this assumption is satisfied in all applications known to us.

Under assumptions **A**, **B** we can implement the outlined aggregation scheme as follows:

Setup. Choose $\delta_1(\varepsilon), \delta_2(\varepsilon)$ satisfying the relations

$$\begin{array}{lrcl} (a) & \frac{1}{\delta_1^2(\varepsilon)} + \frac{1}{\delta_2^2(\varepsilon)} & = & \frac{1}{\varepsilon^2}; \\ (b) & \frac{\delta_1(\varepsilon)}{\varepsilon} & \to & 1, \ \varepsilon \to +0; \\ (c) & \delta_2^{-1}(\varepsilon)\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, F) & \to & \infty, \ \varepsilon \to 0, \end{array} \tag{6.23}$$

which is possible in view of (6.22).

Aggregation.

1) Given observation $y = y^{f,\varepsilon}$ with known noise intensity ε, we split it into two independent observations $y' = y^{f,\delta_1(\varepsilon)}, y'' = y^{f,\delta_2(\varepsilon)}$.

2) We build M vectors $f^j = \widehat{f}^j_{\delta_1(\varepsilon)}(y')$, $j = 1, ..., M$, and apply to these vectors orthogonalization procedure to get an orthonormal system $h^1, ..., h^M$ with the same linear span.

3) We use the observation y'' to get estimates $z_j = (f, h^j) + \delta_2(\varepsilon)\xi_j$ of the projections of f on the directions $h^1, ..., h^M$ and define the resulting estimate of f as

$$\widetilde{f}_\varepsilon(y) = \sum_{j=1}^{M} z_j h^j.$$

Relation (6.15) immediately yields the following

Proposition 6.2.1 *Under assumptions* **A**, **B** *one can solve the Aggregation problem* **L** *associated with the collection $\mathcal{M}$ of estimation methods and the set of signals F "asymptotically ideally" in the minimax sense. Namely, for the outlined estimation method $\{\widetilde{f}_\varepsilon\}_{\varepsilon>0}$ one has*

$$\mathcal{R}_\varepsilon(\widetilde{f}_\varepsilon, F) \equiv \sup_{f\in F} \left(\mathcal{E}\left\{\| f - \widetilde{f}_\varepsilon(y^{f,\varepsilon}) \|^2\right\}\right)^{1/2} \le (1+o(1))\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, F), \ \varepsilon \to +0. \tag{6.24}$$

Indeed, from (6.15) it follows that

$$\mathcal{R}_\varepsilon(\widetilde{f}_\varepsilon, F) \le \mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\delta_1(\varepsilon), F) + \delta_2(\varepsilon)\sqrt{M}.$$

By (6.23.b) and Assumption **A**, the first term in the right hand side of this bound is $(1+o(1))\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, F)$, while the second term, by (6.23.c), is $o(1)\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, F)$.

Note that we are not restricted to deal with mimicking the best linear aggregation of *once for ever fixed* number M of estimation methods: we can allow the collection to extend at certain, not too high, rate as $\varepsilon \to +0$. Thus, assume that we have a "nested family"

$$\mathcal{M} = \left\{\widehat{f}^j_\varepsilon(\cdot)\right\}^{\varepsilon>0}_{j=1,...,M(\varepsilon)}$$

of estimation methods. The notions of the ideal linear aggregation of the methods from the family and the associated "ideal aggregation risks" $\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, f)$ and $\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, F)$

at a given signal f and at a given family $F \subset H$ of signals can be straightforwardly extended to our new situation. The assumptions **A**, **B** which allowed us to get Proposition 6.2.1 now should be modified as follows: **A** remains unchanged, and **B** is replaced with the assumption

> **B.1.** *The worst-case, over $f \in F$, risk of "ideal linear aggregation" of the estimation methods from $\mathcal{M}$ satisfies the relation*
>
> $$\left(\varepsilon\sqrt{M(\varepsilon)}\right)^{-1}\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, F) \to \infty,\ \varepsilon \to +0 \tag{6.25}$$

which is an upper bound on the rate at which $M(\varepsilon)$ is allowed to grow as $\varepsilon \to +0$. Finally, the setup rule (6.23.c) should be replaced with

$$\left(\delta_2(\varepsilon)\sqrt{M(\varepsilon)}\right)^{-1}\mathcal{R}^{\mathrm{LA}}_{\mathcal{M}}(\varepsilon, F) \to \infty,\ \varepsilon \to +0. \tag{6.26}$$

With these modifications of the assumptions and the construction, we still ensure (6.24), i.e., still are able to get "asymptotically ideal" linear aggregation of our, now extending as $\varepsilon \to +0$, nested family of estimation methods.

6.3 Application: aggregating projection estimates

Recall that we have fixed an orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$ in our "universe" – in the Hilbert space H. To the moment this basis was playing a completely technical role of representing an observation as a countable (and thus – "tractable") sample of random variables. In fact, in the traditional L_2 regression theory basis plays a much more important role – the majority of the traditional estimators are "basis-dependent". As the simplest – and typical – example, consider a *linear* or, better to say, a simple *filter* estimate associated with a given basis $\{\phi_i\}$.

6.3.1 Linear estimates

A *linear estimate* is specified by a square-summable sequence of its *weights* $\lambda = \{\lambda_i\}_{i=1}^{\infty}$ and is just

$$\widehat{f}^{\lambda}(y) = \sum_{i=1}^{\infty} \lambda_i y_i \phi_i. \tag{6.27}$$

As applied to an observation $y^{f,\varepsilon}$, the estimate becomes

$$\widehat{f}^{\lambda}(y^{f,\varepsilon}) = \sum_{i=1}^{\infty} \lambda_i (f, \phi_i)\phi_i + \varepsilon\left[\sum_{i=1}^{\infty} \lambda_i \xi_i \phi_i\right] \tag{6.28}$$

The "stochastic vector series" $\varepsilon\left[\sum\limits_{i=1}^{\infty} \lambda_i \xi_i \phi_i\right]$ in the right hand side of this expression clearly converges in the mean square sense to a random element of H (recall that λ is a square summable sequence), so that the estimate makes sense. One can easily compute the squared expected error of this estimate:

$$\begin{array}{rcl}\mathcal{R}^2_\varepsilon(\widehat{f}^\lambda, f) &=& \sum\limits_{i=1}^\infty (1-\lambda_i)^2(f,\phi_i)^2 + \varepsilon^2\sum\limits_{i=1}^\infty \lambda_i^2 \\ &=& d^2(\lambda, f) + \varepsilon^2 e^2(\lambda), \\ d(\lambda, f) &=& \sqrt{\sum\limits_{i=1}^\infty (1-\lambda_i)^2(f,\phi_i)^2} \\ e(\lambda, f) &=& \sqrt{\sum\limits_{i=1}^\infty \lambda_i^2}. \end{array} \tag{6.29}$$

The "deterministic" component $d^2(\lambda, f)$ of the squared risk depends on the signal f and is nothing but the squared norm of the *bias* of our filter (the difference between the input $f = \sum\limits_i f_i\phi_i$ of the filter and its output $\sum\limits_i \lambda_i f_i \phi_i$ in the absence of errors; from now on,

$$f_i = (f, \phi_i)$$

stand for the coordinates of a signal in our fixed basis). The stochastic component $\varepsilon^2 e^2(\lambda)$ of the squared risk is nothing but the energy of the noise component of the output, the input being affected by white noise of intensity ε.

The simplest linear estimates are the so called *projection estimates* $\widehat{f}^k$ – the weights λ_j are equal to 1 for j not exceeding certain k (called the *degree* of the estimate) and are zero for $j > k$. Note that for the projection estimate of degree k relation (6.29) becomes

$$\mathcal{R}^2_\varepsilon(\widehat{f}^k, f) = \sum_{i=k+1}^\infty f_i^2 + \varepsilon^2 k \tag{6.30}$$

and is very transparent: the larger is the degree of the estimate, the less is the deterministic component of the squared risk and the larger is its stochastic component – the situation similar, "up to vice versa", to the one with window estimates. "Reversed", as compared to the case of window estimates, monotonicity properties of the deterministic and the stochastic components of the squared risk as functions of the "window width" (the role of the latter now is played by the degree k of the estimate) is quite natural: narrow windows in "time domain" correspond to wide windows in the "frequency domain".

From the theoretical viewpoint, the interest in linear and projection estimates comes from the fact that they are minimax optimal in order on very natural classes of signals – on "ellipsoids".

An "ellipsoid" is given by a sequence of its "half-axes"

$$a_1 \geq a_2 \geq ... : \quad a_i \in \mathbf{R} \cup \{+\infty\}, a_i \to 0, i \to \infty$$

and its "radius" – a positive real L – and is defined as

$$E(\{a_i\}, L) = \{f \in H \mid \sum_{i=1}^\infty \frac{f_i^2}{a_i^2} \leq L^2\}.$$

For example, the class of k times differentiable periodic with derivatives of order $< k$, of period 1, signals f with $\int_0^1 |f^(k)(x)|^2 dx \leq L^2$ is an ellipsoid with respect to the standard trigonometric orthonormal basis in $L_2[0,1]$, and its half-axes are $a_i = (1 + o(1))(\pi i)^k$ (see (6.2)). Similarly, a Sobolev ball $\mathcal{S}_1^{k,2}(L)$ (see Chapter 2) is an ellipsoid with respect to a properly chosen (depending on k) orthonormal basis in $L_2[0,1]$, the half-axes of the ellipsoid possessing the same asymptotics as in the periodic case (6.2).

It can be easily seen that when estimating signals from a given ellipsoid, one can find optimal in order estimates already among the simplest – the projection – ones, choosing properly the degree of the projection estimate as a function of noise intensity ε. Moreover, given an ellipsoid E and a noise intensity ε, we can easily build the best, in the minimax sense on E, among all linear estimates. In view of (6.29), to this end it suffices to solve the optimization program

$$\begin{array}{rcl}\Phi_\varepsilon(\lambda) & \to & \min \\ \Phi_\varepsilon(\lambda) & \equiv & \sup\{\sum\limits_{i=1}^{\infty}(1-\lambda_i)^2 f_i^2 + \varepsilon^2 \sum\limits_{i=1}^{\infty} \lambda_i^2 \mid f : \sum\limits_{i=1}^{\infty}(f_i/a_i)^2 \le L^2\} \\ & = & L^2 \max_i (1-\lambda_i)^2 a_i^2 + \varepsilon^2 \sum\limits_{i=1}^{\infty} \lambda_i^2.\end{array}$$

It is immediately seen that the optimal solution to this optimization program has the following structure:

$$\lambda_i^* = \frac{(a_i - t(\varepsilon))_+}{a_i}, \tag{6.31}$$

where $a_+ = \max(a,0)$ and $t(\varepsilon)$ is the minimizer of the univariate function

$$\phi_\varepsilon(t) = L^2 t^2 + \varepsilon^2 \sum_i \frac{[(a_i - t)_+]^2}{a_i^2}$$

over $t > 0$.

A remarkable result of M.S. Pinsker [26] is that under minimal regularity assumptions on the sequence of half-axes of the ellipsoid in question (which are satisfied for the case when the axes decrease as a power sequence $a_j \sim j^{-\alpha}$ or as a geometric progression), the optimal in the minimax sense on E *linear* estimate is asymptotically optimal in the minimax sense among *all possible* estimates. As a byproduct of this fact, we can point out not only the *order* of the principal term of the minimax risk $\mathcal{R}(\varepsilon, E)$ as $\varepsilon \to +0$, but this principal term itself; this is a very rare case in the non-parametric regression when we know both the principal term of the minimax risk and an asymptotically optimal up to factor $(1+o(1))$, not just in order, estimation method.

A shortcoming of the initial results on minimax optimality in order/minimax optimality up to factor $(1+o(1))$ of projection/linear estimates on ellipsoids is that to build an estimate of this type we should know the parameters of the ellipsoid – the sequence of its half-axes and its radius (cf. the case of estimates from Chapters 1 – 2). As far as the projection estimates are concerned, there are several popular, although not too well understood theoretically, techniques for specifying the degree of the estimate from observations; for linear estimates, for a long time no theoretically valid "adaptation schemes" were known. A breakthrough in the area is due to Efroimovich and Pinsker [27] who proposed an *adaptive* estimation method which is asymptotically optimal (up to $(1+o(1))$!) in the minimax sense on a wide spectrum of ellipsoids.

We are about to demonstrate that the results completely similar to those of Efroimovich and Pinsker can be obtained just by linear aggregation of projection estimates.

6.3.2 Aggregating projection estimates

Looking at the structure (6.31) of the optimal, in the minimax sense on an ellipsoid E, linear estimate, we see that this estimate has a nonincreasing sequence of weights belonging to the segment $[0,1]$ [2]. In other words, all these estimates belong to the set Λ of linear estimates with nonincreasing weights from $[0,1]$:

$$\Lambda = \{\lambda \in \ell_2 \mid 1 \geq \lambda_1 \geq \lambda_2 \geq ..., \sum_i \lambda_i^2 < \infty\}.$$

Now, given a signal $f \in H$ and a positive ε, we may ask ourselves what is the best, for these f and ε, linear estimate from the class Λ. The answer is clear: the corresponding weights are given by the solution to the optimization problem

$$(P_{f,\varepsilon}) : \qquad \min\left\{\sqrt{\sum_{i=1}^{\infty}[(1-\lambda_i)^2 f_i^2 + \varepsilon^2\lambda_i^2]} \mid \lambda \in \Lambda\right\}.$$

The optimal value in this problem, i.e., the best quality of reproducing f from observations (6.6) by a linear estimate with nonincreasing weights, the intensity of noise being ε, is certain function $\Phi(f,\varepsilon)$ of f,ε. What we are about to build is an estimation method $\mathcal{B}^m = \{\widehat{f}_\varepsilon^m(\cdot)\}_{\varepsilon>0}$ depending on a single "design parameter" $m \in \mathbf{N}$ with the following property:

> (!) *Whenever* $\| f \| \leq 1$ *and* $0 < \varepsilon < 1$, *the risk of the estimate* $\widehat{f}_\varepsilon^m$ *at* f *can be bounded as follows:*
>
> $$\mathcal{R}_\varepsilon(\widehat{f}_\varepsilon^m, f) \leq (1+\gamma_m(\varepsilon))\Phi(f,\varepsilon) + C_m\varepsilon \mathrm{Ln}_m(1/\varepsilon), \tag{6.32}$$
>
> *where* $\gamma_m(\varepsilon) \to 0$, $\varepsilon \to 0$, *is independent of* f, C_m *depends on* m *only, and*
>
> $$\mathrm{Ln}_m(x) = \underbrace{\ln(1+\ln(1+\ln(1+...+\ln(1+\ln(1+x))...)))}_{m \text{ times}}$$
>
> *is the "m-iterated" logarithm.*

Postponing for the moment the construction which leads to (6.32), let us look what are the consequences. Consider an ellipsoid $E(\{a_i\},L)$ and assume (in fact this assumption can be eliminated) that the ellipsoid is contained in the unit ball. According to Pinsker's result, for a given intensity of noise the best, in the minimax sense on E, linear estimate is minimax optimal up to $(1+o(1))$ factor as $\varepsilon \to +0$, and this estimate, as we have seen, for every ε is given by certain weight sequence $\lambda = \lambda(\varepsilon) \in \Lambda$. Combining this fact with the definition of $\Phi(\cdot,\cdot)$, we conclude that for the minimax risk $\mathcal{R}^*(\varepsilon,E)$ associated with the ellipsoid E it holds

$$\mathcal{R}^*(\varepsilon,E) \geq (1-o(1))\sup_{f\in E}\Phi(f,\varepsilon), \ \varepsilon \to +0.$$

In view of this relation, (6.32) implies that

$$\mathcal{R}_\varepsilon(\widehat{f}_\varepsilon^m, F) \leq (1+o(1))\mathcal{R}^*(\varepsilon,E) + C_m\varepsilon\mathrm{Ln}_m(\varepsilon^{-1}), \ \varepsilon \to +0.$$

Consequently,

[2] From (6.29) it is absolutely clear that there is no sense to speak about linear estimates with part of the weights outside $[0,1]$: replacing a weight $\lambda_i \notin [0,1]$ by the closest weight from this segment, we always improve the quality of the estimate. The actually important observation is that the weights λ_i^* given by (6.31) form a nonincreasing sequence.

> (!!) *The estimation method* $\mathcal{B}^m$ *is asymptotically optimal in the minimax sense, up to* $(1+o(1))$ *factor, on every ellipsoid* E *such that*

$$\left(\varepsilon \mathrm{Ln}_m(\varepsilon^{-1})\right)^{-1} \mathcal{R}^*(\varepsilon, E) \to \infty, \varepsilon \to \infty.$$

(!!) is a very strong "certificate of adaptive optimality", since the minimax risks associated with interesting ellipsoids do not decrease with ε too fast. E.g., in the case when $\{\phi_i\}$ is the standard trigonometric basis in $L_2[0,1]$, it turns out that

- When the half-axes a_i of E decrease sub-linearly:

  $$a_i \geq O(i^{-\alpha}),$$

  for some α, as it is the case for ellipsoids comprised of smooth periodic functions of fixed degree of smoothness, one has

  $$\frac{\mathcal{R}^*(\varepsilon, E)}{\varepsilon \ln(\varepsilon^{-1})} \to \infty, \ \varepsilon \to +0,$$

  so that already the method $\mathcal{B}^1$ is asymptotically optimal on E;

- When a_i decrease at most exponentially:

  $$a_i \geq O(\exp\{-\alpha i\}), \ i \to \infty,$$

  for some α, as it is the case, e.g., for classes of functions $f(x) = \phi(\exp\{2\pi i x\})$, $\phi(z)$ being analytic in a fixed ring containing the unit circumference, one has

  $$\frac{\mathcal{R}^*(\varepsilon, E)}{\varepsilon \mathrm{Ln}_2(\varepsilon^{-1})} \to \infty, \ \varepsilon \to +0,$$

  so that the method $\mathcal{B}^2$ is asymptotically optimal on E;

- When a_i decrease at most double-exponentially:

  $$a_i \geq O(\exp\{-\exp\{O(i)\}\}),$$

  the method $\mathcal{B}^3$ is asymptotically optimal on E, etc.

6.3.3 The construction

We are about to build the estimation method $\mathcal{B}^m$ underlying (!). In what follows, $0 < \varepsilon < 1$.

1^0. Let us set

$$\rho(\varepsilon) = \mathrm{Ln}_m^{-1/6}(10\varepsilon^{-1}). \tag{6.33}$$

Given $\varepsilon > 0$, let us define a sequence of positive integers $\{k_j(\varepsilon)\}_{j=1}^{\infty}$ as follows. Let $\nu(\varepsilon)$ be the first integer ν such that $(1+\rho(\varepsilon))^\nu > \frac{1}{\rho(\varepsilon)}$. We set

$$k_j(\varepsilon) = \begin{cases} j, & j \leq \nu(\varepsilon) \\ k_{j-1} + \lfloor (1+\rho(\varepsilon))^j \rfloor, & j > \nu(\varepsilon) \end{cases}, \tag{6.34}$$

where $\lfloor a \rfloor$ is the largest integer not exceeding a.

The structure of the sequence $\{k_j\}$ is quite transparent:

$$\nu(\varepsilon) = O\left(\rho^{-1}(\varepsilon)\ln(\rho^{-1}(\varepsilon))\right)$$

initial terms of the sequence are just subsequent integers 1,2,..., so that the corresponding differences

$$d_j = k_{j+1}(\varepsilon) - k_j(\varepsilon)$$

are equal to 1. Starting with $j = \nu(\varepsilon) + 1$, the differences d_j of two subsequent terms of our sequence become integer approximations of the geometric progression $(1+\rho(\varepsilon))^{j+1}$. Note that the number $K(n,\varepsilon)$ of terms $k_j(\varepsilon)$ not exceeding a positive integer n is not too large:

$$K(n,\varepsilon) = \max\{j : k_j(\varepsilon) \le n\} \le O_m(1)\rho^{-1}(\varepsilon)\left[\ln(\rho^{-1}(\varepsilon)) + \ln n\right]. \tag{6.35}$$

2^0. Let us set

$$\widetilde{N}_\ell(\varepsilon) = \mathrm{Ln}^2_{\ell-1}(10^2\varepsilon^{-2}),\ \ell = 1,...,m, \tag{6.36}$$

where for $\ell \ge 1$ $\mathrm{Ln}_\ell(\cdot)$ is the ℓ-iterated logarithm and $\mathrm{Ln}_0(x) = x$, and let $N_\ell(\varepsilon)$ be the first integer in the sequence $\{k_j(\varepsilon)\}$ which is $\ge \widetilde{N}_\ell(\varepsilon)$.

For $\ell = 1,...,m$, let $P_\ell(\varepsilon)$ be the set of all projection estimates $\widehat{f}^k$ of degrees belonging to the sequence $\{k_j(\varepsilon) - 1\}$ and not exceeding $N_\ell(\varepsilon) - 1$; note that

$$P_1(\varepsilon) \supset P_2(\varepsilon) \supset ... \supset P_m(\varepsilon).$$

Let $K_\ell(\varepsilon)$ be the cardinality of $P_\ell(\varepsilon)$; according to (6.35), for all small enough values of ε we have

$$K_\ell(\varepsilon) \le O_m(1)\rho^{-1}(\varepsilon)\left[\ln(\rho^{-1}(\varepsilon)) + \ln N_\ell(\varepsilon)\right] \le O_m(1)\rho^{-1}(\varepsilon)\mathrm{Ln}_\ell(\varepsilon^{-1}). \tag{6.37}$$

Our plan is as follows: given $\varepsilon > 0$, we aggregate all projection estimates from $P_\ell(\varepsilon)$ according to the scheme of Section 6.2, thus getting m "aggregated estimates" $\widetilde{f}^\ell_\varepsilon$, $\ell = 1,...,m$, and then aggregate these m estimates, thus coming to the desired estimate $\widetilde{f}_\varepsilon$. The precise description of the construction is as follows:

Setup. We choose in advance three positive functions $\delta_1(\varepsilon), \delta_2(\varepsilon), \delta_3(\varepsilon)$ in such a way that

$$\begin{array}{lrcl} (a) & \sum\limits_{\nu=1}^{3}\frac{1}{\delta_\nu^2(\varepsilon)} & = & \frac{1}{\varepsilon^2} \\ (b) & \frac{\delta_1(\varepsilon)}{\varepsilon} & \to & 1,\ \varepsilon \to +0; \\ (c) & \delta_\nu(\varepsilon) & \le & O_m(1)\frac{\varepsilon}{\rho(\varepsilon)},\ \nu = 2,3, \end{array} \tag{6.38}$$

which of course is possible.

Building estimate $\widetilde{f}_\varepsilon$. 1) Given observation $y = y^{f,\varepsilon}$ with known noise intensity ε, we split it into three independent observations $y^\nu = y^{f,\delta_\nu(\varepsilon)}$, $\nu = 1,2,3$.

2) We build $K_1(\varepsilon)$ vectors $f^j = \widehat{f}^{k_j(\varepsilon)-1}(y^1)$, $j = 1,...,K_1(\varepsilon)$, and apply to the vectors the orthogonalization procedure to get an orthonormal system $\{h^j\}_{j=1}^{K_1(\varepsilon)}$ such that the linear span of $h^1,...,h^s$ is, for every $s = 1,2,...,K_1(\varepsilon)$, coincides with the linear span of $f^1,...,f^s$.

3) We use the observation y^2 to get estimates

$$z_j = (f, h^j) + \delta_2(\varepsilon)\xi_j$$

of the projections of f on the directions h^j, $j = 1,...,K_1(\varepsilon)$ (the noises ξ_j, $j = 1,...,K_1(\varepsilon)$, are independent of each other and of y^1 $\mathcal{N}(0,1)$ random variables), and for every $\ell = 1,...,m$ define $g^\ell \equiv \widetilde{f}^\ell_\varepsilon(y) \in H$ as

$$g^\ell = \sum_{j=1}^{K_\ell(\varepsilon)} z_j h^j.$$

4) We apply to the m vectors $g^1,...,g^m$ the orthogonalization process to get an orthonormal system $\{e^\ell\}_{\ell=1}^m$ with the same linear span, and use the observation y^3 to get estimates $w_\ell = (f, e^\ell) + \delta_3(\varepsilon)\eta_\ell$ of the projections of f onto $e^1,...,e^m$, the noises η_ℓ, $\ell = 1,...,m$, being independent of each other and of y^1, y^2 $\mathcal{N}(0,1)$ random variables, and define the resulting estimate $\widetilde{f}_\varepsilon$ of f as

$$\widetilde{f}_\varepsilon = \widetilde{f}_\varepsilon(y^{f,\varepsilon}) = \sum_{\ell=1}^m w_\ell e^\ell.$$

Accuracy analysis, I. Let

$$\Lambda(\varepsilon) = \left\{\lambda = \{\lambda_j\}_{j=1}^\infty \in \ell_2 : \lambda_l = \lambda_{l'}, k_j(\varepsilon) \le l \le l' < k_{j+1}(\varepsilon),\ \forall j = 1,2,...\right\},$$

and let

$$\Lambda_\ell(\varepsilon) = \{\lambda \in \Lambda(\varepsilon) \mid \lambda_l = 0, l \ge N_\ell(\varepsilon)\},\ \ell = 1,...,m.$$

Let us set

$$\mathcal{R}_\ell(\varepsilon, f) = \inf_{\lambda \in \Lambda_\ell(\varepsilon)} \mathcal{R}_\varepsilon(\widehat{f}^\lambda, f), \tag{6.39}$$

where $\widehat{f}^\lambda$ is the linear estimate with weight vector λ.

Observe that every weight vector $\lambda \in \Lambda_\ell(\varepsilon)$ is a linear combination of the weight vectors of projection estimates $\widehat{f}^k \in P_\ell(\varepsilon)$. It follows that the risk $\mathcal{R}^{\rm LA}_{P_\ell(\varepsilon)}(\varepsilon, f)$ of "ideal linear aggregation" of the estimates from $P_\ell(\varepsilon)$ is, for every f and $\varepsilon > 0$, at most $\mathcal{R}_\ell(\varepsilon, f)$. Applying (6.14), we get

$$\mathcal{R}^2_\varepsilon(\widetilde{f}^\ell_\varepsilon, f) \le \left(\mathcal{R}^{\rm LA}_{P_\ell(\varepsilon)}(\delta_1(\varepsilon), f)\right)^2 + \delta_2^2(\varepsilon)K_\ell(\varepsilon) \le \mathcal{R}^2_\ell(\delta_1(\varepsilon), f) + \delta_2^2(\varepsilon)K_\ell(\varepsilon). \tag{6.40}$$

Recalling how the resulting estimate $\widetilde{f}_\varepsilon$ is obtained from the estimates $\widetilde{f}^\ell_\varepsilon$ and applying the same arguments as those used to get (6.14), we conclude that

$$\mathcal{R}^2_\varepsilon(\widetilde{f}_\varepsilon, f) \le \min_{\ell=1,...,m}\left[\mathcal{R}^2_\ell(\delta_1(\varepsilon), f) + \delta_2^2(\varepsilon)K_\ell(\varepsilon)\right] + \delta_3^2(\varepsilon)m,$$

whence also

$$\mathcal{R}_\varepsilon(\widetilde{f}_\varepsilon, f) \le \min_{\ell=1,...,m}\left[\mathcal{R}_\ell(\delta_1(\varepsilon), f) + \delta_2(\varepsilon)\sqrt{K_\ell(\varepsilon)}\right] + \delta_3(\varepsilon)\sqrt{m}. \tag{6.41}$$

Accuracy analysis, II. We are ready to demonstrate that the estimation method we have built satisfies (!). Let us fix $f \in H$, $\| f \| \leq 1$, and $\varepsilon \in (0,1)$.

1^0. It is clear that the optimization program $(P_{f,\varepsilon})$ specifying the best, for noise intensity ε, linear estimate of f with weights from Λ is solvable; let $\lambda = \lambda(f,\varepsilon)$ be an optimal solution to this problem. The corresponding squared risk is

$$\Phi^2(f,\varepsilon) = \sum_{j=1}^{\infty}(1-\lambda_j)^2 f_j^2 + \varepsilon^2 \sum_{j=1}^{\infty}\lambda_j^2, \tag{6.42}$$

and

$$1 \geq \lambda_1 \geq \lambda_2 \geq ...; \ \lambda_j \to 0, \ j \to \infty \tag{6.43}$$

by the definition of Λ.

2^0 Let n be the largest of integers i such that $\lambda_i \geq \rho(\varepsilon)$ (if no $i \geq 1$ with this property exists, $n = 0$). Note that (6.42), (6.43) imply that

$$\Phi(f,\varepsilon) \geq \varepsilon n^{1/2}\rho(\varepsilon). \tag{6.44}$$

On the other hand, it is clear that $\Phi(f,\varepsilon) \leq 1$ (since the value of the objective in $(P_{f,\varepsilon})$ is $\| f \| \leq 1$ already at the trivial feasible solution $\lambda = 0$). We conclude that

$$n \leq \frac{1}{\varepsilon^2\rho^2(\varepsilon)} < N_1(\varepsilon).$$

Let $\ell_* \equiv \ell_*(f,\varepsilon)$ be the largest of values $\ell = 1,...,m$ such that $n < N_\ell(\varepsilon)$.

3^0. Let us build weight vector $\widehat{\lambda} \in \Lambda_{\ell_*}(\varepsilon)$ as follows:

- If $j \geq N_{\ell_*}(\varepsilon)$, then $\widehat{\lambda}_j = 0$;
- If $j < N_{\ell_*}(\varepsilon)$, then there exists the largest $i = i(j)$ such that $k_i(\varepsilon) \leq j$, and we set
$$\widehat{\lambda}_j = \lambda_{k_i(\varepsilon)}, \ i = i(j).$$

Note that by construction $\widehat{\lambda} \in \Lambda_{\ell_*}(\varepsilon)$.

Let

$$R^2 = \sum_{j=1}^{\infty}(1-\widehat{\lambda}_j)^2 f_j^2 + \delta_1^2(\varepsilon)\sum_{j=1}^{\infty}\widehat{\lambda}_j^2 \tag{6.45}$$

be the squared risk of recovering f by the linear estimate with the weight vector $\widehat{\lambda}$, the intensity of noise being $\delta_1(\varepsilon)$. Our local goal is to verify that

$$R^2 \leq (1+\Theta_m(\varepsilon))^2\Phi^2(f,\varepsilon), \tag{6.46}$$

$\Theta_m(\varepsilon) \geq 0$ being an *independent of* f and converging to 0 as $\varepsilon \to +0$ function. The cases of $f = 0$ and/or $\lambda = 0$ are trivial; assume that $f \neq 0$, $\lambda \neq 0$, and let us bound from above the ratio

$$\begin{array}{rcl} \theta^2 & = & \frac{R^2}{\Phi^2(f,\varepsilon)} \leq \max\{\theta_d^2, \theta_s^2\}, \\ \theta_d^2 & = & \frac{\sum\limits_j (1-\widehat{\lambda}_j)^2 f_j^2}{\sum\limits_j (1-\lambda_j)^2 f_j^2}, \\ \theta_s^2 & = & \frac{\delta_1^2(\varepsilon)}{\varepsilon^2}\frac{\sum\limits_j \widehat{\lambda}_j^2}{\sum\limits_j \lambda_j^2} \end{array} \tag{6.47}$$

By construction, for $j < N_{\ell_*}(\varepsilon)$ we have $0 \le \lambda_j \le \widehat{\lambda}_j \le 1$, while for $j \ge N_{\ell_*}(\varepsilon)$ we have $\lambda_j < \rho(\varepsilon)$ and $\widehat{\lambda}_j = 0$. Thus, we have

$$(1-\widehat{\lambda}_j)^2 \le (1-\rho(\varepsilon))^{-2}(1-\lambda_j)^2$$

for all j, whence

$$\theta_d^2 \le (1-\rho(\varepsilon))^{-2}. \tag{6.48}$$

It remains to bound from above θ_s^2; the first ratio in the expression defining θ_s^2 does not depend on f and tends to 1 as $\varepsilon \to +0$ by (6.38.b), so that we may focus on bounding the ratio

$$\vartheta_s^2 = \frac{\sum\limits_j \widehat{\lambda}_j^2}{\sum\limits_j \lambda_j^2} \le \frac{\sum\limits_{j=1}^N \widehat{\lambda}_j^2}{\sum\limits_{j=1}^N \lambda_j^2}, \quad N = N_{\ell*}(\varepsilon).$$

Note that if the initial N-dimensional segment of $\widehat{\lambda}$ differs from that one of λ (this is the only case we should consider), then the connection between these segments is as follows. We can partition the range $\{1, ..., N\}$ of values of index j in subsequent groups $I_1, ..., I_p$ in such a way that

1. The first $\nu = \nu(\varepsilon) = O\left(\rho^{-1}(\varepsilon)\ln\rho^{-1}(\varepsilon)\right)$ of the groups are singletons: $I_j = \{j\}$, $j \le \nu$, and $\widehat{\lambda}_j = \lambda_j$ for $j \le \nu$;

2. For $\nu < l \le p$, the group $I_l = \{j \mid k_l \le j < k_{l+1}\}$ contains d_l indices, where $d_l = \lfloor (1+\rho(\varepsilon))^l \rfloor$, and $\widehat{\lambda}_j = \lambda_{k_l}$ for $j \in I_l$.

Let

$$\begin{array}{rcl} S^\nu & = & \sum\limits_{j=1}^{\nu} \lambda_j^2 \qquad \left[= \sum\limits_{j=1}^{\nu} \widehat{\lambda}_j^2 \right], \\ S_l & = & \sum\limits_{j \in I_l} \lambda_j^2, \ l = \nu+1, ..., p. \end{array}$$

Note that for $l \ge \nu+2$ we have

$$\lambda_{k_l}^2 \ge \frac{S_{l-1}}{d_{l-1}}$$

(see (6.43)), and therefore

$$\begin{array}{rcl} \sum\limits_{j=1}^N \widehat{\lambda}_j^2 & = & S^\nu + \sum\limits_{l=\nu+1}^{p} d_l \lambda_{k_l}^2 \\ & \le & S^\nu + d_{\nu+1}\lambda_{k_{\nu+1}}^2 + \sum\limits_{l=\nu+2}^{p} d_l d_{l-1}^{-1} S_{l-1} \\ & \le & \left(\max\limits_{l \ge \nu+2}[d_l d_{l-1}^{-1}]\right)\left(S^\nu + \sum\limits_{l=\nu+1}^{p} S_l\right) + d_{\nu+1}\lambda_{k_{\nu+1}}^2 \\ & \le & \left(\max\limits_{l \ge \nu+2}[d_l d_{l-1}^{-1}]\right)\left(\sum\limits_{j=1}^{N} \lambda_j^2\right) + d_{\nu+1}\lambda_{k_{\nu+1}}^2, \end{array}$$

whence

$$\vartheta_s^2 \le \left(\max\limits_{l \ge \nu+2}[d_l d_{l-1}^{-1}]\right) + d_{\nu+1}\lambda_{k_{\nu+1}}^2 / S^\nu. \tag{6.49}$$

When $l \geq \nu + 2$, we have

$$\begin{array}{rcl} \frac{d_l}{d_{l-1}} & = & \frac{\lfloor (1+\rho(\varepsilon))^l \rfloor}{\lfloor (1+\rho(\varepsilon))^{l-1} \rfloor} \\ & \leq & \frac{(1+\rho(\varepsilon))^l}{(1+\rho(\varepsilon))^{l-1}-1} \\ & \leq & (1+\rho(\varepsilon))(1+2\rho(\varepsilon)) \quad [\text{since } (1+\rho(\varepsilon))^\nu \geq \rho^{-1}(\varepsilon)] \end{array} \tag{6.50}$$

Besides this, $S^\nu \geq \nu\lambda^2_{k_{\nu+1}}$ by (6.43), while $d_{\nu+1} \leq (1+\rho(\varepsilon))^{\nu+1}$, so that

$$d_{\nu+1}\lambda^2_{\nu+1}/S^\nu \leq (1+\rho(\varepsilon))^{\nu+1}\nu^{-1} \leq O_m(1)\frac{1}{\ln \rho^{-1}(\varepsilon)} \tag{6.51}$$

(recall that $\nu = \nu(\varepsilon) = O\left(\rho^{-1}(\varepsilon)\ln \rho^{-1}(\varepsilon)\right)$ and $\rho(\varepsilon)$ is small when ε is small).

Combining (6.47), (6.48), (6.49) – (6.51), we come to (6.46).

4^0. With (6.46) at hand, we are nearly done. Indeed, by origin of R we have

$$\mathcal{R}_{\ell_*}(\delta_1(\varepsilon), f) \leq R$$

(see the definition of $\mathcal{R}_\ell$ in "Accuracy analysis I"). Combining this observation, (6.46) and (6.41), we come to the inequality

$$\mathcal{R}_\varepsilon(\widetilde{f}_\varepsilon, f) \leq (1+\Theta_m(\varepsilon))\Phi(f,\varepsilon) + \delta_2(\varepsilon)\sqrt{K_{\ell_*}(\varepsilon)} + \delta_3(\varepsilon)\sqrt{m},$$

whence, by $(6.38.b, c)$,

$$\mathcal{R}_\varepsilon(\widetilde{f}_\varepsilon, f) \leq (1+\Theta_m(\varepsilon))\Phi(f,\varepsilon) + O_m(1)\frac{\varepsilon}{\rho(\varepsilon)}\left[\sqrt{K_{\ell_*}(\varepsilon)} + \sqrt{m}\right]. \tag{6.52}$$

For a given f, there are two possible cases:

(I): $\ell_* < m$;

(II): $\ell_* = m$.

In the case of (I) we have $n \geq N_{\ell_*+1}(\varepsilon)$, whence, by (6.44),

$$\Phi(f,\varepsilon) \geq \varepsilon\rho(\varepsilon)\sqrt{n} \geq \varepsilon\rho(\varepsilon)\sqrt{N_{\ell_*+1}(\varepsilon)} \geq O_m(1)\varepsilon\rho(\varepsilon)\mathrm{Ln}_{\ell_*}(\varepsilon^{-1})$$

(note that due to their origin, $N_\ell(\varepsilon) = O(\mathrm{Ln}^2_{\ell-1}(\varepsilon^{-1}))$ when $\varepsilon \to 0$). Therefore in the case of (I) the ratio of the second right hand side term in (6.52) to the first one does not exceed $O_m(1)$ times the quantity

$$\begin{array}{rcll} \frac{\sqrt{K_{\ell_*}(\varepsilon)}}{\rho^2(\varepsilon)\mathrm{Ln}_{\ell_*}(\varepsilon^{-1})} & \leq & O_m(1)\frac{\rho^{-5/2}(\varepsilon^{-1})\sqrt{\mathrm{Ln}_{\ell_*}(\varepsilon^{-1})}}{\mathrm{Ln}_{\ell_*}(\varepsilon^{-1})} & [\text{we have used } (6.37)] \\ & \leq & O_m(1)\frac{\mathrm{Ln}^{5/12}_m(\varepsilon^{-1})}{\sqrt{\mathrm{Ln}_{\ell_*}(\varepsilon^{-1})}} & [\text{see } (6.33)] \\ & \leq & O_m(1)\mathrm{Ln}^{-1/6}_m(\varepsilon^{-1}) & [\text{since } \ell_* < m] \\ & \equiv & \Omega_m(\varepsilon) \to 0,\ \varepsilon \to +0 & \end{array}$$

Thus, if f is such that (I) is the case, then relation (6.52) implies that

$$\mathcal{R}_\varepsilon(\widetilde{f}_\varepsilon, f) \leq (1+\gamma_m(\varepsilon))\Phi(f,\varepsilon) \tag{6.53}$$

with independent of f function $\gamma_m(\varepsilon) \to 0$, $\varepsilon \to +0$. It remains to consider the case when f is such that (II) is the case. Here the second right hand side term in (6.52) is

$$\begin{array}{rcll} O_m(1)\frac{\varepsilon}{\rho(\varepsilon)}\left[\sqrt{K_{\ell_*}(\varepsilon)} + \sqrt{m}\right] & \leq & O_m(1)\varepsilon\rho^{-3/2}(\varepsilon)\sqrt{\mathrm{Ln}_m(\varepsilon^{-1})} & \\ & \leq & O_m(1)\varepsilon\mathrm{Ln}_m(\varepsilon^{-1}) & [\text{see } (6.33)]. \end{array}$$

Combining the latter relation with (6.53) and (6.52), we come to (6.32). ■

6.4 Approximating the best of given estimates

We have considered two of our three aggregation problems – **C**, where we are interested to mimic the best convex combination of a given family estimates, and **L**, where the goal is to reproduce the best linear combination of the estimates from the family. Now let us address the third problem. Thus, assume we are given a nested family $\mathcal{M} = \{\widehat{f}^j_\varepsilon(\cdot)\}_{j=1,\dots,M(\varepsilon)}^{\varepsilon>0}$ of estimates of signals $f \in H$, H being a separable Hilbert space with an orthonormal basis $\{\phi_i\}$, via observations (6.6). For every $f \in H$ and every $\varepsilon > 0$, let us denote by

$$\mathcal{R}_{\mathcal{M}}(\varepsilon, f) = \min_{j\le M(\varepsilon)} \mathcal{R}_\varepsilon(\widehat{f}^j_\varepsilon, f) \equiv \min_{j\le M(\varepsilon)} \left(\mathcal{E}\left\{\| f - \widehat{f}^j_\varepsilon(y^{f,\varepsilon}) \|^2\right\}\right)^{1/2}$$

the minimal, over the estimates from the family, risk of recovering f, the intensity of noise being ε. We are interested to solve the following

> Aggregation problem **V**. *Given a nested family $\mathcal{M}$ of estimation methods, find an estimation method with the risk, at every $f \in H$, "close" to the risk $\mathcal{R}_{\mathcal{M}}(\varepsilon, f)$ of the best, with respect to f, estimate from the family.*

A solution to the problem can be obtained by straightforward exploiting the aggregation technique from Section 6.2, which now should be used in a "cascade" mode. Namely, without loss of generality we may assume that $M(\varepsilon)$, for every ε, is an integral power of 2:

$$M(\varepsilon) = 2^{\mu(\varepsilon)}$$

and that $\mu(\varepsilon)$ is nonincreasing in $\varepsilon > 0$. What we intend to do is to split a given observation $y^{f,\varepsilon}$ into $\mu(\delta_0(\varepsilon))+1$ independent observations $y^j = y^{f,\delta_j(\varepsilon)}$, $j = 0, 1, \dots, \mu(\delta_0(\varepsilon))$, and to use y^0 to build all $2^{\mu(\delta_0(\varepsilon))}$ of the estimates from the family, let us call them "estimates of generation 0". We partition these estimates into pairs and use the observation y^1 to approximate the closest to f linear combinations of estimates in every of the resulting $2^{\mu(\varepsilon)-1}$ pairs, thus coming to $2^{\mu(\delta_0(\varepsilon))-1}$ estimates of "generation 1". Applying the same construction to estimates of generation 1 with y^2 playing the role of y^1, we get $2^{\mu(\delta_0(\varepsilon))-2}$ estimates of "generation 2", and so on, until a single "estimate of generation $\mu(\delta_0(\varepsilon))$" is built; this estimate is the result of our aggregation routine. The precise description of the routine is as follows:

Setup. We choose somehow a function $\delta(\varepsilon) > \varepsilon$ and set

$$\widehat{\delta}(\varepsilon) = \sqrt{\mu(\delta(\varepsilon))}\frac{\varepsilon\delta(\varepsilon)}{\sqrt{\delta^2(\varepsilon) - \varepsilon^2}}. \tag{6.54}$$

Recovering routine $\widetilde{f}_\varepsilon$. 1) Given observation $y = y^{f,\varepsilon}$ of a signal $f \in H$, we set

$$\widehat{\varepsilon} = \delta(\varepsilon)$$

and split y into $\mu(\widehat{\varepsilon})+1$ independent observations $y^0, y^1, \dots, y^{\mu(\widehat{\varepsilon})}$, the noise intensities being $\delta(\varepsilon)$ for y^0 and $\widehat{\delta}(\varepsilon)$ for every one of the remaining y's.

Note that

$$\frac{1}{\delta^2(\varepsilon)}+\frac{\mu(\widehat{\varepsilon})}{\widehat{\delta}^2(\varepsilon)}=\frac{1}{\delta^2(\varepsilon)}+\frac{\delta^2(\varepsilon)-\varepsilon^2}{\varepsilon^2\delta^2(\varepsilon)}=\frac{1}{\varepsilon^2},$$

so that the required splitting is possible.

2) We use y^0 to build $2^{\mu(\widehat{\varepsilon})}$ vectors $f_0^j \equiv f_0^j(y^0) = \widehat{f}^j_{\widehat{\varepsilon}}(y^0) \in H$ – "estimates of generation 0".

3) For $\nu = 1, ..., \mu(\widehat{\varepsilon})$, we perform the following operations.

Given $2M_\nu = 2^{\mu(\widehat{\varepsilon})-\nu+1}$ "estimates of generation $\nu-1$" – vectors $f^j_{\nu-1} = f^j_{\nu-1}(y^0, ..., y^{\nu-1}) \in H$ – partition them into M_ν pairs P^ν_ℓ, $\ell = 1, ..., M_\nu$. For every pair $P^\nu_\ell = \{f^{\nu-1}_\ell, g^{\nu-1}_\ell\}$, we build an orthonormal basis $\{h^{\ell,\nu}_\kappa\}_{\kappa=1,2}$ in the linear span of the vectors from the pair and use the observation y^ν to build estimates

$$z^{\nu,\ell}_\kappa = (f, h^{\nu,\ell}_\kappa) + \widehat{\delta}(\varepsilon)\xi^{\nu,\ell}_\kappa, \kappa = 1, 2$$

with independent of each other and of $y^0, ..., y^{\nu-1}$ $\mathcal{N}(0, 1)$ random noises $\xi^{\nu,\ell}_\kappa$, $\kappa = 1, 2$.

We set

$$f^\ell_\nu = z^{\nu,\ell}_1 h^{\ell,\nu}_1 + z^{\nu,\ell}_2 h^{\ell,\nu}_2.$$

After all M_ν pairs P^ν_ℓ are processed, $M_\nu = 2^{\mu(\widehat{\varepsilon})-\nu}$ estimates f^ℓ_ν of "generation ν" are built, and we either pass to the next step (if $\nu < \mu(\widehat{\varepsilon})$), increasing ν by one, or terminate (if $\nu = \mu(\widehat{\varepsilon})$), the single estimate of generation $\mu(\widehat{\varepsilon})$ being the result $\widetilde{f}_\varepsilon(y)$ of our aggregation routine.

Exactly the same reasoning which led us to (6.14) demonstrates that for every $f \in H$ and for every $\nu = 1, ..., \mu(\widehat{\varepsilon})$ and every $\ell = 1, ..., M_\nu$ it holds

$$\begin{array}{rcl}\mathcal{E}\left\{\| f - f^\ell_\nu(y^0, ..., y^\nu) \|^2\right\} & \le & \mathcal{E}\left\{\min\limits_{g\in \mathrm{Lin}\{P^\nu_\ell\}} \| f - g \|^2\right\} + 2\widehat{\delta}^2(\varepsilon)\\ & \le & \min\limits_{g\in P^\nu_\ell}\mathcal{E}\left\{\| f - g \|^2\right\} + 2\widehat{\delta}^2(\varepsilon),\end{array}$$

while

$$\mathcal{E}\left\{\| f - f^j_0(y^0) \|^2\right\} \le \mathcal{R}_{\delta(\varepsilon)}(\widehat{f}^j_\varepsilon, f),\ j = 1, ..., M(\delta(\varepsilon)).$$

Combining these observations, we come to

$$\mathcal{E}\left\{\| f - \widetilde{f}_\varepsilon \|^2\right\} \le \min_{j=1,...,M(\delta(\varepsilon))} \mathcal{R}_{\delta(\varepsilon)}(\widehat{f}^j_{\delta(\varepsilon)}, f) + 2\mu(\delta(\varepsilon))\widehat{\delta}^2(\varepsilon).$$

Recalling the origin of $\widehat{\delta}(\varepsilon)$, we come to the following

Proposition 6.4.1 *Let $\mathcal{M} = \{f^j_\varepsilon\}_{\substack{\varepsilon>0\\ j=1,...,M(\varepsilon)}}$ be a nested family of estimation methods, $M(\varepsilon)$ being nonincreasing in $\varepsilon > 0$. For every function $\delta(\varepsilon) > \varepsilon$, the risk of the associated with $\delta(\cdot)$, according to the above construction, aggregated estimation method $\{\widetilde{f}_\varepsilon\}_{\varepsilon>0}$ satisfies the relation*

$$\mathcal{R}_\varepsilon(\widetilde{f}_\varepsilon, f) \le \mathcal{R}_{\mathcal{M}}(\delta(\varepsilon), f) + O(1)\frac{\varepsilon\delta(\varepsilon)}{\sqrt{\delta^2(\varepsilon)-\varepsilon^2}}\ln M(\varepsilon) \quad \forall(f \in H, \varepsilon > 0). \qquad (6.55)$$

In particular, we get a sufficient condition for "asymptotically efficient", in the minimax sense, aggregation:

Corollary 6.4.1 *Let $F \subset H$ be a family of signals, and let $\mathcal{M}$ be the same nested family of estimation methods as in Proposition 6.4.1. Assume that*

I. *The "minimax risk"*

$$\mathcal{R}_{\mathcal{M}}(\varepsilon, F) = \sup_{f\in F} \min_{j=1,\dots,M(\varepsilon)} \left(\mathcal{E}\left\{\| f - \widehat{f}_\varepsilon^j \|^2\right\}\right)^{1/2}$$

associated with $F, \mathcal{M}$ is a "well-behaved" function of ε as $\varepsilon \to 0$: whenever a function $\delta(\varepsilon)$ is such that $\delta(\varepsilon)/\varepsilon \to 1$ as $\varepsilon \to +0$, one has

$$\mathcal{R}_{\mathcal{M}}(\delta(\varepsilon), F) \le (1 + o(1))\mathcal{R}_{\mathcal{M}}(\varepsilon, F),\ \varepsilon \to +0;$$

II. *The risk $\mathcal{R}_{\mathcal{M}}(\varepsilon, F)$ satisfies the relation*

$$\varepsilon \ln M(\varepsilon) = o(1)\mathcal{R}_{\mathcal{M}}(\varepsilon, F),\ \varepsilon \to +0$$

*(**II** in fact is an upper bound on the rate at which the number $M(\varepsilon)$ of estimates to be aggregated can grow as $\varepsilon \to +0$).*

Under these assumptions, the estimation methods from the family $\mathcal{M}$ restricted on the class of signals F admit "asymptotically efficient aggregation": there exists an estimation method $\{\widetilde{f}_\varepsilon\}_{\varepsilon>0}$ such that

$$\mathcal{R}_\varepsilon(\varepsilon, F) \le (1 + o(1))\mathcal{R}_{\mathcal{M}}(\varepsilon, F),\ \varepsilon \to +0.$$

To get the asymptotically efficient aggregation mentioned in the Corollary, it suffices to implement the above construction with $\delta(\varepsilon)/\varepsilon$ approaching 1 as $\varepsilon \to +0$ so slowly that

$$\frac{\varepsilon\delta(\varepsilon)}{\sqrt{\delta^2(\varepsilon) - \varepsilon^2}} \ln M(\varepsilon) = o(1)\mathcal{R}_{\mathcal{M}}(\varepsilon, F),\ \varepsilon \to +0;$$

the possibility of such a choice of $\delta(\cdot)$ is guaranteed by assumption **II.**

Chapter 7

Estimating functionals, I

From now on we switch from the problem of estimating a nonparametric regression function to the problem of estimating functional of such a function.

7.1 The problem

We continue to work within the bounds of the L_2-theory and Gaussian white noise model of observations. Geometrical setting of the generic problem we are interested in is as follows:

> We are given
>
> - a real separable Hilbert space H with inner product $(\cdot,\cdot)$ and an orthonormal basis $\{\phi_i\}_{i=1}^{\infty}$,
> - a set $\Sigma \subset H$,
> - a real-valued functional F defined in a neighbourhood of Σ.
>
> A "signal" $f \in \Sigma$ is observed in Gaussian white noise of intensity ε, i.e., we are given a sequence of observations
>
> $$y^{f,\varepsilon} = \left\{ y_i^{f,\varepsilon} \equiv (f, \phi_i) + \varepsilon \xi_i \right\}, \qquad (7.1)$$
>
> $\{\xi_i\}_{i=1}^{\infty}$ being a collection of independent $\mathcal{N}(0,1)$ random variables ("the noise"), and our goal is to estimate via these observations the value $F(f)$ of F at f.
>
> As always, we will be interested in asymptotic, $\varepsilon \to 0$, results.

Recall that the model (7.1) is the geometric form of the standard model where signals f are functions from $L_2[0,1]$, and observation is the "functional observation"

$$y_f(x) = \int_0^x f(s)ds + \varepsilon W(x), \qquad (7.2)$$

$W(x)$ being the standard Wiener process; in this "functional language", interesting examples of functionals F are the Gateau functionals

$$F(f) = \int\limits_0^1 G(x, f(x))dx \qquad (7.3)$$

or

$$F(f) = \int_0^1 \dots \int_0^1 G(x_1, \dots, x_k, f(x_1), \dots, f(x_k)) dx_1 \dots dx_k. \tag{7.4}$$

In this chapter we focus on the case of a smooth functional F. As we shall see, if the parameters of smoothness of F "fit" the geometry of Σ, then $F(f)$, $f \in \Sigma$, can be estimated with "parametric convergence rate" $O(\varepsilon)$, and, moreover, we can build *asymptotically efficient*, uniformly on Σ, estimates.

7.1.1 Lower bounds and asymptotical efficiency

In order to understand what "asymptotical efficiency" should mean, the first step is to find out what are limits of performance of an estimate. The answer can be easily guessed: if $F(f) = (f, \psi)$ is a continuous linear functional, so that

$$\psi = \sum_{i=1}^{\infty} \psi_i \phi_i, \ \{\psi_i = (\psi, \phi_i)\}_{i=1}^{\infty} \in \ell^2,$$

then seemingly the best way to estimate $F(f)$ is to use the "plug-in" estimate

$$\widehat{F}(y) = \sum_{i=1}^{\infty} \psi_i y_i^{f,\varepsilon} = (f, \psi) + \varepsilon \sum_{i=1}^{\infty} \psi_i \xi_i$$

(the series in the right hand side converges in the mean square sense, so that the estimate makes sense); the estimate is unbiased, and its variance clearly is $\varepsilon^2 \parallel \psi \parallel^2$. Now, if F is Freĉhet differentiable in a neighbourhood of a signal $f \in \Sigma$, then we have all reasons to expect that locally it is basically the same – to estimate F or the linearized functional $\bar{F}(g) = \bar{F}(f) + (F'(f), g - f)$, so that the variance of an optimal estimate in this neighbourhood should be close to $\varepsilon^2 \parallel F'(f) \parallel^2$. Our intuition turns out to be true:

Theorem 7.1.1 [13] *Let $\bar{f} \in \Sigma$ and F be a functional defined on Σ. Assume that*

(i) *Σ is convex, and F is Gateau differentiable "along Σ" in a neighbourhood U of $\bar{f}$ in Σ: for every $f \in U$, there exists a vector $F'(f) \in H$ such that*

$$\lim_{t \to +0} \frac{F(f + t(g - f)) - F(f)}{t} = (F'(f), g - f) \quad \forall g \in \Sigma,$$

and assume that every one of the functions $\psi_g(t) = (F'(\bar{f} + t(g - \bar{f})), g - \bar{f})$, $g \in \Sigma$, is continuous in a neighbourhood of the origin of the ray $\{t \geq 0\}$

(ii) *The "tangent cone" of Σ at $\bar{f}$ – the set*

$$T = \{h \in H \mid \exists t > 0 : \bar{f} + th \in \Sigma\}$$

– is dense in a half-space $H_+ = \{h \in H \mid (\psi, h) \geq 0\}$ associated with certain $\psi \neq 0$.

Then the local, at $\bar{f}$, squared minimax risk of estimating $F(f)$, $f \in \Sigma$, via observations (7.1) *is at least $\varepsilon^2(1 + o(1)) \parallel F'(\bar{f}) \parallel^2$:*

$$\lim_{\delta \to +0} \liminf_{\varepsilon \to +0} \inf_{\widehat{F} \in \mathcal{F}} \sup_{f \in \Sigma, \|f - \bar{f}\| \leq \delta} \mathcal{E}\left\{\varepsilon^{-2}\left[\widehat{F}(y^{f,\varepsilon}) - F(f)\right]^2\right\} \geq \parallel F'(\bar{f}) \parallel^2, \tag{7.5}$$

where $\mathcal{F}$ is the family of all possible estimates (i.e., real-valued Borel functions $\widehat{F}(y)$ on the space $\mathbf{R}^\infty$ of real sequences[1]) *and $\mathcal{E}$ is the expectation with respect to the noises $\{\xi_i\}$.*

In other words, for every fixed $\delta > 0$ the squared minimax risk of estimating $F(f)$ in a δ-neighbourhood (in Σ) of $\bar{f}$ is at least $\varepsilon^2\left(\| F'(\bar{f}) \|^2 + o(1)\right)$, $\varepsilon \to +0$.

Proof. Let $d = F'(\bar{F})$; there is nothing to prove is $d = 0$, so that we may assume that $d \neq 0$. By (ii), either d, or $-d$ is a limit of a sequence $\{h_i \in T\}$; for the sake of definiteness, assume that $d = \lim_{i\to\infty} h_i$ (the alternative case is completely similar).

Let us fix positive δ.

1^0 Let $\kappa \in (0, 1/4)$. Since d is a limiting point of the set T, there exists a unit vector $h \in T$ such that $(h, d) \geq \| d \| (1-\kappa)$. By definition of T, there exists a segment $\Delta = [0, r]$, with $0 < r < \delta$ such that $f_t = \bar{f} + th \in \Sigma$ for all $t \in \Delta$. Taking into account (i) and decreasing, if necessary, the value of r, we may assume that the function

$$\alpha(t) = F(f_t)$$

satisfies the condition

$$(1-2\kappa) \| d \| \leq \alpha'(t) \leq (1+2\kappa) \| d \|, \ t \in \Delta, \tag{7.6}$$

whence the inverse function $\alpha^{-1}(s)$ satisfies the relation

$$|\alpha^{-1}(s) - \alpha^{-1}(s')| \leq \frac{1}{(1-2\kappa) \| d \|} |s - s'|, \ \alpha(0) \leq s, s' \leq \alpha(r). \tag{7.7}$$

2^0. Now let us fix $\varepsilon > 0$, let $\widehat{F}$ be an arbitrary estimate from $\mathcal{F}$, and let

$$\rho^2 = \sup_{t\in\Delta} \mathcal{E}\left\{\left[\widehat{F}(y^{f_t,\varepsilon}) - F(f_t)\right]^2\right\}$$

be the squared minimax risk of estimating the value of F on the segment $S = \{f_t\}_{t\in\Delta}$ of signals. We claim that

$$\rho^2 \geq \left(\frac{r\varepsilon(1-2\kappa) \| d \|}{r + 2\varepsilon}\right)^2, \tag{7.8}$$

Postponing for a while the justification of our claim, let us derive from (7.8) the required lower bound. Indeed, since $\widehat{F}$ is an arbitrary estimate and by construction segment S is contained in Σ and in the δ-neighbourhood of $\widehat{f}$, we get

$$\inf_{\widehat{F}\in\mathcal{F}} \sup_{f\in\Sigma, \|f-\bar{f}\|\leq\delta} \mathcal{E}\left\{\varepsilon^{-2}\left[\widehat{F}(y^{f,\varepsilon}) - F(f)\right]^2\right\} \geq \varepsilon^{-2}\rho^2 = \left(\frac{r(1-2\kappa) \| d \|}{r+2\varepsilon}\right)^2,$$

whence

$$\liminf_{\varepsilon\to+0} \inf_{\widehat{F}\in\mathcal{F}} \sup_{f\in\Sigma, \|f-\bar{f}\|\leq\delta} \mathcal{E}\left\{\varepsilon^{-2}\left[\widehat{F}(y^{f,\varepsilon}) - F(f)\right]^2\right\} \geq (1-2\kappa)^2 \| d \|^2 .$$

The resulting inequality is valid for all $\delta, \kappa > 0$, and (7.5) follows.

[1] As always, $\mathbf{R}^\infty$ is equipped with metric defining the Tikhonov topology

3^0. It remains to verify (7.8). Assume, on contrary, that (7.8) is wrong: there exists $\widehat{F} \in \mathcal{F}$ and $\varepsilon > 0$ such that

$$\sup_{t\in\Delta} \mathcal{E}\left\{\left[\widehat{F}(y^{f_t,\varepsilon}) - F(f_t)\right]^2\right\} < \left(\frac{r\varepsilon(1-2\kappa)\parallel d \parallel}{r+2\varepsilon}\right)^2. \qquad (7.9)$$

3^0.1) Since $F(f_t)$, $0 \le t \le r$, takes its values in the segment $[\alpha(0), \alpha(r)]$, we may assume that $\widehat{F}$ takes its values in the latter segment; indeed, if it is not the case, we may pass from $\widehat{F}$ to the "truncated" estimate

$$\widetilde{F}(y) = \begin{cases} \alpha(0), & \widehat{F}(y) < \alpha(0) \\ \widehat{F}(y), & \alpha(0) \le \widehat{F}(y) \le \alpha(r) \\ \alpha(r), & \widehat{F}(y) > \alpha(r) \end{cases};$$

when replacing $\widehat{F}$ with $\widetilde{F}$, we may only decrease the left hand side in (7.9), and the truncated estimate takes its values in $[\alpha(0), \alpha(r)]$.

3^0.2) Thus, we may assume that $\alpha(0) \le \widehat{F}(\cdot) \le \alpha(r)$. Now let us set

$$\widehat{t}(y) = \alpha^{-1}(\widehat{F}(y)).$$

Combining (7.9) and (7.7), we conclude that

$$\forall t \in \Delta = [0, r]: \qquad \mathcal{E}\left\{\left[\widehat{t}(y^{f_t,\varepsilon}) - t\right]^2\right\} < \frac{r^2\varepsilon^2}{(r+2\varepsilon)^2}. \qquad (7.10)$$

3^0.3) Without loss of generality we may assume that $\bar{f} = 0$; changing, if necessary, our orthonormal basis in H, we may assume also that h is the first basic orth ϕ_1 (recall that our observations have the same structure in every orthonormal basis). Then $f_t = t\phi_1$, and (7.10) says the following:

> (*) There exists possibility to recover parameter $t \in [0, r]$ from observations
> $$y_1 = t + \varepsilon\xi_1, y_2 = \varepsilon\xi_2, y_3 = \varepsilon\xi_3, ... \qquad (7.11)$$
> with independent $\mathcal{N}(0,1)$ random noises $\xi_1, \xi_2, ...$ in such a way that the variance of the recovering error, for every $t \in [0, r]$, is $< \frac{r^2\varepsilon^2}{(r+2\varepsilon)^2}$.

Since observations $y_2, y_3, ...$ impart no information on t, (*) simply says that

> (**) Given that $t \in [0, r]$, there exists possibility to recover the mean t of $\mathcal{N}(t, \varepsilon^2)$ random variable from a single realization of this variable with the variance of the error, uniformly in $t \in [0, r]$, less than $\frac{r^2\varepsilon^2}{(r+2\varepsilon)^2}$.
>
> Formal reasoning corresponding to our "impart no information" arguments is as follows: passing from the estimate $\widehat{t}(y)$ to the estimate
> $$\widetilde{t}(y_1) = \mathcal{E}_{\xi_2,\xi_3,...}\left\{\widehat{t}(y_1, \xi_2, \xi_3, ...)\right\},$$
> we may only improve the variance of recovering t from observations (7.11) and get an estimate which depends on y_1 only.

It remains to note that (**) is forbidden by the Kramer-Rao inequality. To be self-contained, let us reproduce the corresponding reasoning for the simplest case we are interested in.

Let $\widehat{t}(t+\varepsilon\xi)$ be an estimate of $t \in [0,r]$ via noisy observation $t+\xi$ of t, $\xi \sim \mathcal{N}(0,1)$, and let

$$\begin{array}{rrcl} (a) & \delta(t) & = & \mathcal{E}_\xi \left\{ t - \widehat{t}(t+\varepsilon\xi) \right\} = \int (t - \widehat{t}(s)) p(t,s) ds, \\ & & & p(t,s) = (\varepsilon\sqrt{2\pi})^{-1} \exp\{-(s-t)^2/(2\varepsilon^2)\}, \\ (b) & \gamma^2(t) & = & \mathcal{E}_\xi \left\{ \left[\widehat{t}(t+\varepsilon\xi) - t \right]^2 \right) = \int (\widehat{t}(s) - t)^2 p(t,s) ds \end{array} \quad (7.12)$$

be the expectation and the variance of the estimation error; we are interested to bound from below the quantity

$$\gamma^2 \equiv \sup_{0 \le t \le r} \gamma^2(t).$$

In this bounding, we may assume that the estimate $\widehat{t}$ takes its values in $[0,r]$ (cf. the above "truncation" reasoning). When $\widehat{t}$ is bounded, the bias $\delta(t)$ is continuously differentiable in $[0,r]$, and from (7.12.a) we get

$$\begin{array}{rcll} \delta'(t) & = & 1 - \int (t - \widehat{t}(s)) p'_t(t,s) ds & \\ & = & 1 - \int \left[(t - \widehat{t}(s)) \sqrt{p(t,s)} \right] \left[\frac{p'_t(t,s)}{\sqrt{p(t,s)}} \right] ds & \\ & \ge & 1 - \left(\int (t - \widehat{t}(s))^2 p(t,s) ds \right)^{1/2} \left(\int \frac{(p'_t(t,s))^2}{p(t,s)} ds \right)^{1/2} & \text{[Cauchy's inequality]} \\ & = & 1 - \varepsilon^{-1} \left(\int (t - \widehat{t}(s))^2 p(t,s) ds \right)^{1/2} & \text{[direct computation]} \\ & = & 1 - \varepsilon^{-1} \gamma(t) & \\ & \ge & 1 - \varepsilon^{-1} \gamma & \end{array}$$

Integrating the resulting inequality from $t = 0$ to $t = r$ and taking into account that $|\delta(t)| \le \gamma(t) \le \gamma$, we get

$$2\gamma \ge \delta(r) - \delta(0) \ge r(1 - \varepsilon^{-1}\gamma),$$

whence

$$\gamma^2 \ge \left(\frac{r\varepsilon}{r + 2\varepsilon} \right)^2$$

so that (**) indeed is impossible. ■

The lower bound on local minimax risk of estimating smooth functionals stated by Theorem 7.1.1 motivates the following definition of an *asymptotically efficient* estimation method:

Definition 7.1.1 *Let $\Sigma \subset H$ be a convex family of signals and $F : \Sigma \to \mathbf{R}$ be a functional such that for every $f \in \Sigma$ there exists a vector $F'(f) \in H$:*

$$\lim_{t \to +0} \frac{F(f + t(g-f)) - F(f)}{t} = (F'(f), g - f) \quad \forall f, g \in \Sigma.$$

Assume also that linear combinations of elements of Σ are dense in H, so that $F'(f)$ is uniquely defined by F, f. An estimation method $\{\widehat{F}_\varepsilon(\cdot) \in \mathcal{F}\}_{\varepsilon>0}$ is called asymptotically efficient on Σ, if

$$\limsup_{\varepsilon\to+0} \sup_{f\in\Sigma} \left[\varepsilon^{-2}\mathcal{E}\left\{\left[\widehat{F}(y^{f,\varepsilon}) - F(f)\right]^2\right\} - \| F'(f) \|^2\right] \leq 0. \tag{7.13}$$

E.g., we have seen that a continuous linear functional $F(f) = (f, \psi)$ admits asymptotically efficient, on the entire H, estimation method. Such a functional is a simplest – linear – polynomial on H. We shall see in a while that a polynomial of a degree > 1 also can be estimated in an asymptotically efficient, on every bounded subset of H, fashion, provided that the polynomial is a *Hilbert-Schmidt one*. On the other hand, it turns out that already the function $F(f) =\| f \|^2$ cannot be estimated in an asymptotically efficient fashion on the entire unit ball. Thus, in order to be able to build asymptotically efficient estimates of "smooth", but not "very smooth" functionals, we should restrict the class of signals Σ to be "not too massive", similarly to what we did when recovering the signals themselves. A very convenient way to control the "massiveness" of Σ is to impose restrictions on the *Kolmogorov diameters* $d_k(\Sigma)$:

Definition 7.1.2 *Let $\Sigma \subset H$ and m be a positive integer. We say that the m-dimensional Kolmogorov diameter $d_m(\Sigma)$ of Σ is $\leq \delta$, if there exists an m-dimensional linear subspace $H_m \subset H$ such that*

$$\forall f \in \Sigma : \quad \mathrm{dist}(f, H_m) \equiv \min_{f'\in H_m} \| f - f' \| \leq \delta. \qquad ^{2)}$$

In what follows, we impose on Σ restrictions like

$$d_m(\Sigma) \leq Lm^{-\beta}, \ m \geq m_0 \qquad [\beta > 0], \tag{7.14}$$

i.e., say at which rate the "non-parametric" set of signals Σ can be approximated by "m-parametric" sets – by the projections of Σ on appropriately chosen m-dimensional subspaces of H. E.g., if Σ is an ellipsoid

$$\Sigma = \{f \in H \mid \sum_{i=1}^{\infty} \frac{(f, \phi_i)^2}{a_i^2} \leq L^2\} \qquad [a_1 \geq a_2 \geq ..., a_i \to 0, \ i \to \infty], \tag{7.15}$$

then one clearly has

$$d_m(\Sigma) \leq La_{m+1}, \ m = 1, 2, ...$$

In particular, the Kolmogorov diameters of the "periodic part" of a Sobolev ball $\mathbf{S}_1^{k,2}(L)$ (same as the diameters of the ball itself) decrease as m^{-k} (cf. (6.2)):

$$d_m\left(\mathbf{S}_1^{k,2}(L)\right) \leq c_k L m^{-k}, \ k = m+1, m+2, ...$$

Thus, (7.14) in typical applications is an a priori restriction on the smoothness of signals we deal with.

[2)] The "canonical" definition of the Kolmogorov diameters deals with affine rather than linear subspaces of H; note, however, that if there is an affine m-dimensional subspace H' of H such that $\mathrm{dist}(f, H') \leq \delta \ \forall f \in \Sigma$, there exists $(m+1)$-dimensional linear subspace of H with the same property; thus, "up to shift by 1 in the dimension" (absolutely unimportant in what follows), we may speak about approximation by linear, rather than affine, subspaces.

The goal. In what follows we focus on the questions (a) *what should be the relations between the "degree of smoothness" of a functional F to be estimated and the "asymptotical width" of Σ (i.e., the value of β in* (7.15)*) in order for F to admit an asymptotically efficient, on Σ, estimation method*, and (b) *how to build an asymptotically efficient estimation method, provided that it exists.*

To get a kind of preliminary orientation, let us start with the simplest case of a once continuously differentiable functional.

7.2 The case of once continuously differentiable functional

Consider the problem of estimating a functional F on a set of signals Σ and assume that

A.1. Σ is a bounded subset of H, and the Kolmogorov diameters of Σ satisfy (7.14) with certain a priori known β, L.

For the sake of definiteness, assume that Σ is contained in the unit ball

$$\mathcal{O} = \{f \mid \| f \| \leq 1\}$$

of H.

A.2. The functional F to be estimated is defined in the ball

$$\mathcal{O}_{2\rho} = \{f \mid \| f \| < 1 + 2\rho\} \qquad [\rho > 0],$$

is continuously Fréchet differentiable in $\mathcal{O}_{2\rho}$, and its derivative $F'(\cdot)$ is Hölder continuous in $\mathcal{O}_{2\rho}$ with exponent $\gamma > 0$ and constant L:

$$\forall f, g \in \mathcal{O}_{2\rho}: \quad \| F'(f) - F'(g) \| \leq L \| f - g \|^{\gamma}. \tag{7.16}$$

E.g., the Gateau functional (7.3) satisfies **A.2**, provided that the integrand $G(x,t)$ is continuously differentiable in t for almost all $x \in [0,1]$, is measurable in x for every t and

$$\begin{array}{rcl} G(\cdot,0) & \in & L_1[0,1], \\ G'_t(x,0) & \in & L_2[0,1], \\ \| G'_t(\cdot,\tau) - G'_t(\cdot,\tau') \|_\infty & \leq & C \max\left[|\tau-\tau'|, |\tau-\tau'|^{\gamma}\right] \ \forall \tau,\tau' \in \mathbf{R} \end{array}$$

Similarly, the Gateau functional (7.4) satisfies **A.2**, provided that

$$G(x_1,...,x_k,t_1,...,t_k) = G(\bar{x},\bar{t})$$

is continuously differentiable in $\bar{t}$ for almost all $\bar{x}$, is measurable in $\bar{x}$ for all $\bar{t}$ and

$$\begin{array}{rcl} G(\cdot,0) & \in & L_1([0,1]^k), \\ G'_{\bar{t}}(\cdot,0) & \in & L_2([0,1]^k), \\ \| G'_{\bar{t}}(\cdot,\bar{\tau}) - G'_{\bar{t}}(\cdot,\bar{\tau}') \|_\infty & \leq & C \max\left[|\bar{\tau}-\bar{\tau}'|, |\bar{\tau}-\bar{\tau}'|^{\gamma}\right] \ \forall \bar{\tau},\bar{\tau}' \in \mathbf{R}^k. \end{array}$$

We are about to establish the following result:

Theorem 7.2.1 *Assume that* **A.1**, **A.2** *are satisfied and that the parameters* β *and* γ *are linked by the inequality*

$$\gamma > \frac{1}{2\beta}. \tag{7.17}$$

Then F *admits asymptotically efficient on* Σ *estimation method.*

Proof. We build explicitly the corresponding estimation method.

The idea of the construction is quite transparent. Given noise intensity $\varepsilon > 0$, we choose an appropriate $m = m(\varepsilon)$, find an m-dimensional linear subspace H_m such that $\mathrm{dist}(f, H_m) \leq Lm^{-\beta}$, and build the associated projection estimate $\widehat{f}_m$ of f. After $\widehat{f}_m$ is built, we approximate F by the first order Taylor expansion of F at $\widehat{f}_m$:

$$F(f) \approx F(\widehat{f}_m) + (F'(\widehat{f}_m), f - f^m),$$

f^m being the projection of f onto H_m, and estimate the linear part of this expansion as a linear functional – just substituting, instead of $f - f^m$, the observation of this vector. A nice feature of this scheme is that the noises affecting the observation of $f - f^m$ are independent of those affecting the estimate $\widehat{f}_m$, which allows for easy evaluation of the risk.

The construction is as follows.

1^0. We first choose the "order" $m = m(\varepsilon)$ of the estimate $\widehat{f}_m$ from a very natural desire to get an optimal in order nonparametric estimate of $f \in \Sigma$. To understand what is this order, we use the quite familiar to us reasoning as follows. For a given m, we build an m-dimensional subspace H_m in H such that the norm of the projection $f^\perp$ of $f \in \Sigma$ on the orthogonal complement to H_m (this norm is nothing but $\mathrm{dist}(f, H_m)$) is guaranteed to be $\leq Lm^{-\beta}$, build an orthonormal basis $h_1, ..., h_m$ in H_m and define $\widehat{f}_m$ as

$$\widehat{f}_m = \sum_{i=1}^{m} z_i h_i,$$

where $z_i = (f, h_i) + \varepsilon\eta_i$, $\{\eta_i\}_{i=1}^m$ are independent $\mathcal{N}(0,1)$ random variables, are the estimates of the projections of f onto h_i given by observation $y^{f,\varepsilon}$. The squared risk $\mathcal{E}\left\{\| f - \widehat{f}_m \|^2\right\}$ clearly can be bounded as

$$\mathcal{E}\left\{\| f - \widehat{f}_m \|^2\right\} \leq m\varepsilon^2 + \| f^\perp \|^2 \leq m\varepsilon^2 + L^2 m^{-2\beta}, \tag{7.18}$$

and to get an optimal in order estimate, we should balance the stochastic term $m\varepsilon^2$ and the deterministic term $L^2 m^{-2\beta}$, i.e., to set

$$m = m(\varepsilon) = \lfloor \varepsilon^{-\frac{2}{2\beta+1}} \rfloor \quad ^{3)}. \tag{7.19}$$

After our choice of $m(\varepsilon)$ is specified, we may assume – just to save notation – that H_m is simply the linear span of the first basic orths $\phi_1, ..., \phi_m$ of the basis where the

[3] In order to avoid messy expressions, in what follows we do not opimize the choice of parameters with respect to the constant L involved in **A.1**, **A.2**.

observations (7.1) are given. Indeed, we are working now with fixed ε (and therefore – with fixed $H_{m(\varepsilon)}$) and are in our right to use whatever orthonormal basis we want, updating the observations (without any change in their structure) accordingly.

2^0. Let $f \in \Sigma$ be the observed signal, f^m be its projection on the subspace spanned by the first $m = m(\varepsilon)$ basic orths ϕ_i, and $\widehat{f}_m$ be the corresponding projection estimate of f:

$$\widehat{f}_m = f^m + \varepsilon \sum_{i=1}^{m} \eta_i \phi_i.$$

In order to implement the outlined approximation-based scheme, we should ensure that the "preliminary estimate" we use belongs to the domain of the functional F, which is not the case for some realizations of $\widehat{f}_m$. This is, however, a minor difficulty: since $f \in \Sigma \subset \mathcal{O}$, we only improve the quality $\| \cdot - f \|$ of our estimate by projecting $\widehat{f}_m$ on $\mathcal{O}_\rho$ – by passing from $\widehat{f}_m$ to the estimate

$$\widetilde{f}_m = \begin{cases} \widehat{f}_m, & \| \widehat{f}_m \| \le 1+\rho \\ (1+\rho) \| \widehat{f}_m \|^{-1} \widehat{f}_m, & \| \widehat{f}_m \| > 1+\rho \end{cases}.$$

The estimate $\widetilde{f}_m$ is the one we actually use in the above approximation scheme. Important for us properties of the estimate can be summarized as follows:

(a) For a given $f \in \Sigma$, the estimate $\widetilde{f}_m$ depends on the collection $\xi^m = \{\xi_i\}_{i=1}^m$ of the observations noises in (7.1) and does not depend on the sequence $\xi_{m+1}^\infty = \{\xi_i\}_{i=m+1}^\infty$ of the "remaining" noises;

(b) We have (from now on all C's stand for different positive quantities depending only on F and Σ and independent of ε and of a particular choice of $f \in \Sigma$):

$$\begin{array}{lrcl} (a) & \| \widetilde{f}_m \| & \le & 1+\rho \\ (b) & \| \widetilde{f}_m - f^m \| & \le & \| \widehat{f}_m - f^m \| \\ (c) & \| f - f^m \| & \le & C m^{-\beta}(\varepsilon) \le C\varepsilon^{\frac{2\beta}{2\beta+1}} \\ (d) & \mathcal{E}\left\{ \| f^m - \widetilde{f}_m \|^2 \right\} & = & \mathcal{E}_{\xi^m}\left\{ \| f^m - \widetilde{f}_m \|^2 \right\} \le C\varepsilon^2 m(\varepsilon) \\ & & \le & C\varepsilon^{\frac{4\beta}{2\beta+1}} \end{array} \tag{7.20}$$

Note that (d) is given by (7.19) and the fact that $\| \widetilde{f}_m - f^m \| \le \| \widehat{f}_m - f^m \|$.

3^0. The estimate of $F(f)$ we arrive at is

$$\widehat{F}_\varepsilon \equiv \widehat{F}_\varepsilon(y^{f,\varepsilon}) = F(\widetilde{f}_m) + \left(F'(\widetilde{f}_m), \sum_{i=m+1}^{\infty} y_i \phi_i \right), \; y_i = y_i^{f,\varepsilon} \tag{7.21}$$

(of course, $\widetilde{f}_m$ depends on observations $y_i^{f,\varepsilon}$, $i = 1, ..., m$, and $m = m(\varepsilon)$; to save notation, we omit explicit indication of these dependencies).

Accuracy analysis. To evaluate the accuracy of the estimate we have built, let

$$\begin{array}{rcl} R & = & F(f) - F(f^m) - (F'(f^m), f - f^m), \\ \zeta & = & \varepsilon \sum\limits_{i=m+1}^{\infty} \eta_i [F'(\widetilde{f}_m)]_i \qquad [\text{for } g \in H,\, g_i = (g, \phi_i)] \end{array}$$

so that

$$\begin{array}{rcl} F(f)-\widehat{F}_\varepsilon &=& [F(f^m)+(F'(f^m),f-f^m)+R]-\left[F(\widetilde{f}_m)+(F'(\widetilde{f}_m),f-f^m)+\zeta\right] \\ &=& R+\left\{F(f^m)-F(\widetilde{f}_m)\right\}_1+\left\{(F'(f^m)-F'(\widetilde{f}_m),f-f^m)\right\}_2-\zeta \end{array} \tag{7.22}$$

Observe that R is deterministic, $\{\ \}_1$, $\{\ \}_2$ depend only on ξ^m, while the conditional expectations, ξ^m being fixed, of ζ and ζ^2 are, respectively, 0 and $\varepsilon^2 \sum\limits_{i=m+1}^{\infty}[F'(\widetilde{f}_m)]_i^2$. Consequently,

$$\begin{array}{rcl} \mathcal{E}\left\{\left[F(f)-\widehat{F}_\varepsilon\right]^2\right\} &=& \mathcal{E}\left\{\left[R+F(f^m)-F(\widetilde{f}_m)+(F'(f^m)-F'(\widetilde{f}_m),f-f^m)\right]^2\right\} \\ && +\varepsilon^2\mathcal{E}\left\{\sum\limits_{i=m+1}^{\infty}[F'(\widetilde{f}_m)]_i^2\right\}. \end{array} \tag{7.23}$$

We claim that the following facts hold true:

A)

$$|R| \le o(1)\varepsilon \tag{7.24}$$

From now on, $o(1)$'s stand for deterministic functions of ε (independent of a particular choice of $f \in \Sigma$) tending to 0 as $\varepsilon \to +0$.

B)

$$\mathcal{E}\left\{[F(f^m)-F(\widetilde{f}_m)]^2\right\} \le \varepsilon^2\sum_{i=1}^{m}[F'(f)]_i^2+\varepsilon^2 o(1) \tag{7.25}$$

C)

$$\mathcal{E}\left\{[(F'(f^m)-F'(\widetilde{f}_m),f-f^m)]^2\right\} \le \varepsilon^2 o(1) \tag{7.26}$$

D)

$$\varepsilon^2\mathcal{E}\Big\{\sum_{i=m+1}^{\infty}[F'(\widetilde{f}_m)]_i^2\Big\} \le \varepsilon^2\sum_{i=m+1}^{\infty}[F'(f)]_i^2+\varepsilon^2 o(1) \tag{7.27}$$

Note that (7.23) combined with A) – D) clearly implies that

$$\mathcal{E}\left\{\left[F(f)-\widehat{F}_\varepsilon\right]^2\right\} \le \varepsilon^2 \parallel F'(f) \parallel^2 +\varepsilon^2 o(1) \quad \forall f\in\Sigma,$$

i.e., that the estimate we have built is asymptotically efficient on Σ. Thus, all we need is to verify A) – D)

Verifying A) We have

$$\begin{array}{rcll} |R| &\le& C\parallel f-f^m\parallel^{1+\gamma} & [\text{by } \mathbf{A.2}] \\ &\le& C[m(\varepsilon)]^{-\beta(1+\gamma)} & [\text{by } \mathbf{A.1}] \\ &\le& C\varepsilon^{\frac{2\beta(1+\gamma)}{2\beta+1}} & [\text{by } (7.19)] \\ &=& \varepsilon o(1) & [\text{by } (7.17)] \end{array}$$

as required in (7.24).

Verifying B) We have

$$\begin{aligned} F(f^m) - F(\widetilde{f}_m) &= \Big[F(f^m) + (F'(f^m), \widetilde{f}_m - f^m) - F(\widetilde{f}_m)\Big] + \Big[(F'(f^m), f^m - \widehat{f}_m)\Big] \\ &\quad + \Big[(F'(f^m), \widehat{f}_m - \widetilde{f}_m)\Big], \end{aligned} \tag{7.28}$$

and in order to verify (7.25) it suffices to demonstrate that

$$\begin{array}{lrcl} (a) & \mathcal{E}\left\{\Big[F(f^m) + (F'(f^m), \widetilde{f}_m - f^m) - F(\widetilde{f}_m)\Big]^2\right\} & \leq & \varepsilon^2 o(1) \\ (b) & \mathcal{E}\left\{\Big[(F'(f^m), f^m - \widehat{f}_m)\Big]^2\right\} & \leq & \varepsilon^2 \sum\limits_{i=1}^{m} [F'(f)]_i^2 + \varepsilon^2 o(1) \\ (c) & \mathcal{E}\left\{\Big[(F'(f^m), \widehat{f}_m - \widetilde{f}_m)\Big]^2\right\} & \leq & \varepsilon^2 o(1) \end{array} \tag{7.29}$$

(7.29.a): We have

$$\begin{array}{rll} & \mathcal{E}\left\{\Big[F(f^m) + (F'(f^m), \widetilde{f}_m - f^m) - F(\widetilde{f}_m)\Big]^2\right\} & \\ \leq & \mathcal{E}\left\{C \parallel f^m - \widetilde{f}_m \parallel^{2(1+\gamma)}\right\} & \\ & & [\text{by } \mathbf{A.2}] \\ \leq & \mathcal{E}\left\{C \parallel f^m - \widehat{f}_m \parallel^{2(1+\gamma)}\right\} & \\ & & [\text{by } (7.20.b)] \\ \leq & \mathcal{E}\left\{C \left[\varepsilon^{2(1+\gamma)} \sum\limits_{i=1}^{m} \xi_i^2\right]^{1+\gamma}\right\} & \qquad\qquad \square \\ & & [\text{since } \widehat{f}_m - f^m = \varepsilon \sum\limits_{i=1}^{m} \xi_i] \\ \leq & C[\varepsilon^2 m(\varepsilon)]^{1+\gamma} & \\ \leq & C\varepsilon^{\frac{4\beta(1+\gamma)}{2\beta+1}} & \\ & & [\text{by } (7.19)] \\ \leq & \varepsilon^2 o(1) & \\ & & [\text{by } (7.17)] \end{array}$$

(7.29.b): We have

$$\begin{array}{rl}
& \mathcal{E}\left\{\left[(F'(f^m), f^m - \widehat{f}_m)\right]^2\right\} \\
= & \mathcal{E}\left\{\varepsilon^2\left[\sum\limits_{i=1}^m [F'(f^m)]_i\xi_i\right]^2\right\} \\
& \qquad\qquad [\text{since } \widehat{f}_m - f^m = \varepsilon\sum\limits_{i=1}^m \xi_i] \\
= & \varepsilon^2\sum\limits_{i=1}^m [F'(f^m)]_i^2 \\
= & \varepsilon^2\sum\limits_{i=1}^m \left[[F'(f)]_i + \delta_i\right]^2 \quad [\delta_i = [F'(f^m)]_i - [F'(f)]_i] \\
\leq & \varepsilon^2(1+\theta)\sum\limits_{i=1}^m [F'(f)]_i^2 + \varepsilon^2(1+\theta^{-1})\sum\limits_{i=1}^m \delta_i^2 \;\; \forall\theta > 0 \\
& \qquad [\text{since } (a+b)^2 \leq (1+\theta)a^2 + (1+\theta^{-1})b^2] \\
\leq & \varepsilon^2(1+\theta)\sum\limits_{i=1}^m [F'(f)]_i^2 \\
& +C\varepsilon^2(1+\theta^{-1}) \parallel f - f^m \parallel^{2\gamma} \\
& \qquad\qquad [\text{by } \mathbf{A.2}] \\
\leq & \varepsilon^2(1+\theta)\sum\limits_{i=1}^m [F'(f)]_i^2 + C\varepsilon^2(1+\theta^{-1})\varepsilon^{\frac{4\beta\gamma}{2\beta+1}} \\
& \qquad\qquad [\text{by } (7.20.c)] \\
\leq & \varepsilon^2(1+o(1))\sum\limits_{i=1}^m [F'(f)]_i^2 + \varepsilon^2 o(1) \\
& \qquad\qquad [\text{set } \theta = \varepsilon^{\frac{2\beta\gamma}{2\beta+1}}] \\
\leq & \varepsilon^2\sum\limits_{i=1}^m [F'(f)]_i^2 + \varepsilon^2 o(1) \\
& \qquad\qquad [\text{by } \mathbf{A.2}]
\end{array}$$

□

(7.29.c): Observe first that for every $q \geq 1$ one has

$$\mathcal{E}\left\{\parallel \sum_{i=1}^m \xi_i\phi_i \parallel^{2q}\right\} \leq C(q)m^q. \tag{7.30}$$

Consequently, for every $q \geq 1$ and every $\theta \geq 1$ it holds

$$\begin{array}{rl}
& \mathcal{E}\left\{\parallel \widetilde{f}_m - \widehat{f}_m \parallel^q\right\} \\
\leq & \left(\mathcal{E}\left\{\parallel \widetilde{f}_m - \widehat{f}_m \parallel^{2q}\right\}\right)^{1/2}\left(\mathrm{Prob}\{\widehat{f}_m \neq \widetilde{f}_m\}\right)^{1/2} \\
\leq & 2^q\left(\mathcal{E}\left\{\parallel \widehat{f}_m - f^m \parallel^{2q}\right\}\right)^{1/2}\left(\mathrm{Prob}\{\parallel \widehat{f}_m - f^m \parallel > \rho\}\right)^{1/2} \\
& \qquad [\text{since } \parallel \widetilde{f}_m - f^m \parallel \leq \parallel \widehat{f}_m - f^m \parallel] \\
\leq & C_1(q)[\varepsilon^2 m]^{q/2}\left(\mathrm{Prob}\{\parallel \sum\limits_{i=1}^m \xi_i\phi_i \parallel \geq \rho/\varepsilon\}\right)^{1/2} \\
& \qquad\qquad [\text{by } (7.30)] \\
\leq & C(q,\theta)[\varepsilon^2 m]^{q/2+\theta} \\
& \qquad [\text{since } \mathcal{E}\left\{\parallel \sum\limits_{i=1}^m \xi_i\phi_i \parallel^{4\theta}\right\} \leq C(4\theta)m^{2\theta} \text{ by } (7.30) \\
& \qquad \text{and therefore } \mathrm{Prob}\{\parallel \sum\limits_{i=1}^m \xi_i\phi_i \parallel > \rho/\varepsilon\} \leq \bar{C}(\theta)[\varepsilon^2 m]^{2\theta}]
\end{array}$$

Thus, we get

$$\forall q,\theta \geq 1: \quad \mathcal{E}\left\{\parallel \widetilde{f}_m - \widehat{f}_m \parallel^q\right\} \leq C(q,\theta)[\varepsilon^2 m(\varepsilon)]^{q/2+\theta}. \tag{7.31}$$

We now have

$$\begin{aligned}
& \mathcal{E}\left\{\left[(F'(f^m),\widehat{f}_m-\widetilde{f}_m)\right]^2\right\} \\
\leq\ & C\mathcal{E}\left\{\parallel \widehat{f}_m-\widetilde{f}_m \parallel^2\right\} && \text{[by } \mathbf{A.2}\text{]} \\
\leq\ & C(\theta)[\varepsilon^2 m(\varepsilon)]^{1+\theta} \quad \forall\theta\geq 1 && \text{[by (7.31)]} \qquad \square \\
\leq\ & C(\theta)\varepsilon^{\frac{4\beta(1+\theta)}{2\beta+1}} \quad \forall\theta\geq 1 && \text{[by (7.19)]} \\
\leq\ & \varepsilon^2 o(1) && \text{[choose } \theta \text{ appropriately]}
\end{aligned}$$

Verifying C) We have

$$\begin{aligned}
& \mathcal{E}\left\{[(F'(f^m)-F'(\widetilde{f}_m),f-f^m)]^2\right\} \\
\leq\ & \mathcal{E}\left\{\parallel F'(f^m)-F'(\widetilde{f}_m)\parallel^2 \parallel f-f^m\parallel^2\right\} \\
\leq\ & \mathcal{E}\left\{C\parallel f^m-\widetilde{f}_m\parallel^{2\gamma} m^{-2\beta}(\varepsilon)\right\} && \text{[by } \mathbf{A.2} \text{ and } (7.20.c)\text{]} \qquad \square \\
\leq\ & C\varepsilon^{\frac{4\beta\gamma}{2\beta+1}}m^{-2\beta}(\varepsilon) && \text{[by } (7.20.d) \text{ and since } \gamma\leq 1\text{]} \\
\leq\ & C\varepsilon^2\varepsilon^{\frac{4\beta\gamma-2}{2\beta+1}} && \text{[by (7.19)]} \\
=\ & \varepsilon^2 o(1) && \text{[by (7.17)]}
\end{aligned}$$

Verifying D) We have

$$\begin{aligned}
& \varepsilon^2\mathcal{E}\Big\{\sum_{i=m+1}^{\infty}[F'(\widetilde{f}_m)]_i^2\Big\} \\
\leq\ & \varepsilon^2\mathcal{E}\left\{(1+\theta)\sum_{i=m+1}^{\infty}[F'(f)]_i^2+(1+\theta^{-1})\sum_{i=m+1}^{\infty}\delta_i^2\right\} \ \forall\theta>0 \\
& \qquad\qquad [\delta_i=[F'(f)]_i-[F'(\widetilde{f}_m)]_i, \\
& \qquad\qquad \text{cf. verificaton of } (7.29.b)] \\
\leq\ & \varepsilon^2(1+\theta)\sum_{i=m+1}^{\infty}[F'(f)]_i^2 \\
& +(1+\theta^{-1})\varepsilon^2\mathcal{E}\left\{\parallel F'(f)-F'(\widetilde{f}_m)\parallel^2\right\} \\
\leq\ & \varepsilon^2(1+\theta)\sum_{i=m+1}^{\infty}[F'(f)]_i^2+(1+\theta^{-1})\varepsilon^2\mathcal{E}\left\{\parallel f-\widetilde{f}_m\parallel^{2\gamma}\right\} \qquad \square \\
\leq\ & \varepsilon^2(1+\theta)\sum_{i=m+1}^{\infty}[F'(f)]_i^2+C\varepsilon^2(1+\theta^{-1})\varepsilon^{\frac{4\beta\gamma}{2\beta+1}} \\
& \qquad\qquad \text{[by } (7.20.c,d) \text{ and since } \gamma\leq 1\text{]} \\
\leq\ & \varepsilon^2(1+o(1))\sum_{i=m+1}^{\infty}[F'(f)]_i^2+\varepsilon^2 o(1) \\
& \qquad\qquad \text{[set } \theta=\varepsilon^{\frac{2\beta\gamma}{2\beta+1}}\text{]} \\
=\ & \varepsilon^2\sum_{i=m+1}^{\infty}[F'(f)]_i^2+\varepsilon^2 o(1) \\
& \qquad\qquad \text{[by } \mathbf{A.2}\text{]}
\end{aligned}$$

The proof of Theorem 7.2.1 is completed. ■

7.2.1 Whether condition (7.2.2) is sharp?

We have seen that if the “asymptotical width of Σ” β (see **A.1**) and the “degree of smoothness of F” γ are “properly linked”, namely, $\gamma > \frac{1}{2\beta}$ (see (7.17)), then F admits an asymptotically efficient on Σ estimation method. A natural question is whether the condition (7.17) is "definitive" (i.e., if it is violated, then it may happen that F admits no asymptotically efficient estimation on Σ), or it is an "artifact" coming from the particular estimation method we were dealing with. It turns out that (7.17) indeed is "definitive":

Theorem 7.2.2 *Let $\beta > 0$, $\gamma \in (0,1]$ be such that*

$$\gamma < \frac{1}{2\beta}. \tag{7.32}$$

Then there exist $\Sigma \subset \mathcal{O}$ satisfying **A.1** *and a functional $F : H \to \mathbf{R}$ satisfying* **A.2** *on the entire space H such that F does not admit asymptotically efficient estimation on Σ.*

Proof. Let us set

$$\Sigma = \{f \in H \mid \sum_{i=1}^{\infty} i^{2\beta}(f,\phi_i)^2 \le 1\}, \tag{7.33}$$

so that Σ clearly satisfies **A.1** (one can choose as H_m the linear span of the first m basic orths $\phi_1, ..., \phi_m$).

We are about to build a functional F which satisfies **A.2** and does not admit asymptotically efficient estimation on Σ.

The idea. Assume we are given a noise intensity $\varepsilon > 0$. Let us choose somehow $k = k(\varepsilon)$ and $K = K(\varepsilon) = 2^{k(\varepsilon)}$ distinct elements $f_0, ..., f_{K-1} \in \Sigma$ such that for appropriately chosen $\rho = \rho(\varepsilon)$ it holds:

(i) $\parallel f_i \parallel = 8\rho \ \forall i$;

(ii) $\parallel f_i - f_j \parallel > 2\rho$ whenever $i \neq j$.

Let $\Psi(f)$ be a once for ever fixed smooth function on H which is equal to 1 at the point $f = 0$ and is zero outside the unit ball, e.g.,

$$\Psi(f) = \psi(\parallel f \parallel^2), \tag{7.34}$$

where ψ is a C^∞ function on the axis which is 1 at the origin and vanishes outside $[-1,1]$. Given an arbitrary collection $\omega = \{\omega_i \in \{-1;1\}\}_{i=0}^{K-1}$, let us associate with it the functional

$$\Psi_\omega(f) = \sum_{i=0}^{K-1} \omega_i \Psi_i(f), \quad \Psi_i(f) = \rho^{1+\gamma}\Psi(\rho^{-1}(f - f_i)). \tag{7.35}$$

The structure of the functional is very transparent: every f_i is associated with the term $\omega_i \Psi_i(f)$ in Ψ; this term vanishes outside the centered at f_i ball of radius ρ and is equal to $\omega_i \rho^{1+\gamma}$ at the center f_i of this ball. Due to the origin of ρ, the supports of distinct terms have no points in common, so that

$$\Psi_\omega(f_i) = \omega_i \rho^{1+\gamma} \ i = 0, ..., K-1. \tag{7.36}$$

Besides this, from the fact that the supports of distinct terms in Ψ_ω are mutually disjoint it is immediately seen that Ψ_ω is C^∞ on H and

$$\| \Psi'_\omega(f) - \Psi'_\omega(g) \| \le C \| f - g \|^\gamma \ \ \forall f, g \in H \tag{7.37}$$

with C depending on γ only.

We are about to demonstrate that with properly chosen $k(\varepsilon), \rho(\varepsilon)$, at least one of the 2^K functionals $\Psi_\omega(\cdot)$ corresponding to all 2^K collections of $\omega_i = \pm 1$ is "difficult to evaluate" already on the set $\mathcal{F}_\varepsilon = \{f_0, ..., f_{K-1}\}$, provided that the intensity of noises in (7.1) is ε. Namely, there exists a functional Ψ in the family such that no estimate $\widehat{F}_\varepsilon$ is able to recover its values on $\mathcal{F}_\varepsilon$ with squared risk $\le \varepsilon^{2-\delta}$, $\delta > 0$ being chosen appropriately. After this central fact will be established, we shall combine the "difficult to estimate" functionals corresponding to different values of the noise intensity ε in a single functional which is impossible to evaluate in an asymptotically efficient (even in an order-efficient) way.

In order to prove that there exists a "difficult to estimate" functional of the type Ψ_ω, assume, on contrary, that all these functionals are easy to estimate. Note that we can "encode" a signal $f \in \mathcal{F}_\varepsilon = \{f_0, ..., f_{K-1}\}$ by the values of $k = \log_2 K$ functionals from our family, namely, as follows. Let I_ℓ, $\ell = 1, ..., k$, be the set of indices $i = 0, ..., K-1 = 2^k - 1$ such that the ℓ-th binary digit in the binary representation of i is 1, and let $\Psi^\ell(\cdot)$ be the functional $\Psi_\omega(\cdot)$ corresponding to the following choice of ω:

$$\omega_i = \begin{cases} 1, & i \in I_\ell \\ -1, & i \notin I_\ell \end{cases}$$

In other words, the value of the functional $\Psi^\ell(\cdot)$ at f_i "says" what is the ℓ-th binary digit of i: if it is 1, then $\Psi^\ell(f_i) = \rho^{1+\gamma}$, and if it is 0, then $\Psi^\ell(f_i) = -\rho^{1+\gamma}$. It follows that *the collection of values of k functionals $\Psi^1, \Psi^2, ..., \Psi^k$ at every $f \in \mathcal{F}_\varepsilon$ allows to identify f.*

Now, if all k functionals Ψ^ℓ, $\ell = 1, ..., k$, are "easy to estimate" via observations (7.1), we can use their "good" estimates in order to recover a signal f (known to belong to $\mathcal{F}_\varepsilon$) from observations (7.1), since the collection of values of our functionals at $f \in \mathcal{F}_\varepsilon$ identifies f. On the other hand, we know from the Fano inequality what in fact are our abilities to recover signals from $\mathcal{F}_\varepsilon$ from observations (7.1); if these "actual abilities" are weaker than those offered by the outlined recovering routine, we may be sure that the "starting point" in developing this routine – the assumption that every one of the functionals Ψ^ℓ, $\ell = 1, ..., k$, is easy to estimate on $\mathcal{F}_\varepsilon$ – is false, so that one of these functionals is difficult to estimate, and this is exactly what we need.

The implementation of the above plan is as follows.

1^0. Let us fix β' such that

$$\begin{array}{lrcl} (a) & \beta' & < & \beta \\ (b) & 2\beta\gamma & < & 1 + 2\beta' - 2\beta \end{array} \tag{7.38}$$

(this is possible since $2\beta\gamma < 1$).

2^0. Let us fix $\varepsilon > 0$ and set

$$k = k(\varepsilon) = \lfloor \varepsilon^{-\frac{2}{1+2\beta'}} \rfloor. \tag{7.39}$$

Figure 7.1: Three functionals “encoding” $8 = 2^3$ signals.

In the sequel we assume that ε is so small that $k(\varepsilon) \geq 7$.

3^0. The set Σ given by (7.33) contains the centered at the origin k-dimensional disk of the radius $r = k^{-\beta}$. Since m-dimensional unit sphere contains a set of 2^m points with pairwise distances at least $1/4$[4)], we conclude that for

$$\rho = \rho(\varepsilon) = \frac{1}{8} k^{-\beta}(\varepsilon) \tag{7.40}$$

there exist $K = 2^k$ signals $f_i \in \Sigma$ satisfying conditions (i) and (ii) from the previous item. Let Ψ^ℓ, $\ell = 1, ..., k$, be the functionals associated with $\mathcal{F}_\varepsilon = \{f_0, ..., f_{K-1}\}$ by the construction from the previous item. Let

$$\delta_k(\varepsilon) = \max_{\ell=1,...,k} \inf_{\widehat{F}_\varepsilon} \max_{i=0,...,K-1} \mathcal{E}\left\{ \left[\widehat{F}_\varepsilon(y^{f_i,\varepsilon}) - \Psi^\ell(f_i) \right]^2 \right\}$$

Our central auxiliary results is as follows:

Lemma 7.2.1 *For all small enough values of ε one has*

$$\delta_k(\varepsilon) \geq \frac{1}{128} \rho^{2+2\gamma}(\varepsilon) \geq C \varepsilon^{\frac{4\beta(1+\gamma)}{2\beta'+1}} \tag{7.41}$$

with positive $C > 0$.

Proof. Assume that (7.41) does not hold, so that

$$\delta_k(\varepsilon) < \frac{1}{128} \rho^{2+2\gamma}(\varepsilon) \tag{7.42}$$

Let $\widehat{F}_\varepsilon^\ell$, $\ell = 1, ..., k$ be estimates such that

$$\mathcal{E}\left\{ \left[\widehat{F}_\varepsilon^\ell(y^{f_i,\varepsilon}) - \Psi^\ell(f_i) \right]^2 \right\} \leq 2\delta_k(\varepsilon), \; \ell = 1, ..., k, i = 0, ..., K-1. \tag{7.43}$$

Let

$$m = \lfloor 10 \ln k \rfloor.$$

Consider $K = 2^k$ hypotheses $\mathcal{H}_i$ on the distribution of a sample Y of m observations $y^{(1)}, ..., y^{(m)}$; hypotheses $\mathcal{H}_i$ states that Y is a sample of m independent observations (7.1) associated with the signal f_i. Let us look at the following procedure for distinguishing between these hypotheses:

[4] To see this, note that if $X = \{x_i\}_{i=1}^N$ is the maximal subset of the unit sphere in $\mathbf{R}^m$ such that the pairwise distances between the points of the set are $> 1/4$, then N "spherical hats" $\{x \in \mathbf{R}^m \mid \parallel x \parallel = 1, \parallel x - x_i \parallel \leq \frac{1}{4}\}$ cover the entire sphere. On the other hand, the ratio of the "area" of such a hat and the one of the sphere is

$$\frac{\int\limits_0^{2\arcsin(1/8)} \sin^{m-2}(s)ds}{2\int\limits_0^{\pi/2} \sin^{m-2}(s)ds} \leq= \frac{\int\limits_0^{\sin(2\arcsin(1/8))} t^{m-2}(1-t^2)^{-1/2}dt}{\int\limits_0^1 t^{m-2}(1-t^2)^{-1/2}dt}$$
$$\leq \frac{(m-1)\sin^{m-2}(2\arcsin(1/8))}{\cos(2\arcsin(1/8))} \leq 2^{-m}, m \geq 7,$$

so that $N \geq 2^m$ for $m \geq 7$.

> Given $Y = \{y^{(1)}, ..., y^{(m)}\}$, we for every $\ell = 1, ..., k$ build m reals $F_{\ell j} = \widehat{F}^{\ell}_{\varepsilon}(y^{(j)})$. If more than one half of these reals are positive, we set $b_\ell = 1$, otherwise we set $b_\ell = 0$. After b_1,..., b_k are built, we treat them as the binary digits of (uniquely defined) integer i, $0 \leq i \leq 2^k - 1$ and claim that Y is given by the hypotheses $\mathcal{H}_i$.

Let us evaluate the probability θ to reject a particular hypotheses $\mathcal{H}_i$ when it is true. If for every $\ell = 1, ..., k$ in the sequence $\{F_{\ell j}\}_{j=1}^m$ more than one half of the entries are of the same sign as $\Psi^\ell(f_i)$, then b_ℓ will be exactly the ℓth binary digit $b_\ell(i)$ of i, and the hypotheses $\mathcal{H}_i$ will be accepted. Thus, if $\mathcal{H}_i$ is not accepted, it means that there exists ℓ such that among the entries of the sequence $\{F_{\ell j}\}_{j=1}^m$ at least one half is of the sign opposite to that one of $\Psi^\ell(f_i)$. The probability that it is the case for a particular value of ℓ is at most the probability that in a sequence of m independent identically distributed random variables $\zeta_j = \widehat{F}^{\ell}_{\varepsilon}(y^{(j)}) - \Psi^\ell(f_i)$ at least one half of the elements is in absolute value $\geq \rho^{1+\gamma}(\varepsilon)$. On the other hand, by (7.43) we have

$$\mathcal{E}\left\{\zeta_j^2\right\} \leq 2\delta_k(\varepsilon),$$

whence

$$\mathrm{Prob}\left\{|\zeta_j| \geq \rho^{1+\gamma}(\varepsilon)\right\} \leq \frac{\sqrt{2\delta_k(\varepsilon)}}{\rho^{1+\gamma}(\varepsilon)} < \frac{1}{8}$$

(see (7.42)), so that

$$\mathrm{Prob}_{\mathcal{H}_i}\left\{b_\ell \neq b_\ell(i)\right\} \leq \sum_{m/2 \leq j \leq m} \mathrm{C}_m^j (1/8)^j (7/8)^{m-j} \leq 2^{-m}.$$

It follows that

$$\mathrm{Prob}_{\mathcal{H}_i}\left\{\exists \ell \leq k : b_\ell \neq b_\ell(i)\right\} \leq k 2^{-m} \leq \frac{1}{4}$$

(we have taken into account the origin of m). Thus, for every $i = 0, ..., K-1$ the probability to reject the hypotheses $\mathcal{H}_i$ when it is true is at most 1/4. On the other hand, the pairwise Kullback distances between the distributions of $y^{(j)}$ associated with hypotheses $\mathcal{H}_0, ..., \mathcal{H}_{K-1}$ clearly do not exceed

$$\mathcal{K} = \frac{1}{2\varepsilon^2} \max_{i,j=0,...,K-1} \parallel f_i - f_j \parallel^2 \leq \frac{128 k^{-2\beta}}{\varepsilon^2}$$

(we have taken into account property (i) from the previous item). Applying the Fano inequality (1.27) and recalling that $m \leq 10 \ln k(\varepsilon)$ and $K = 2^{k(\varepsilon)}$, we get

$$\frac{1280 k^{-2\beta}(\varepsilon) \ln k(\varepsilon)}{\varepsilon^2} \geq \frac{1}{4} \ln(2^{k(\varepsilon)} - 1) - \ln 2,$$

In view of (7.39) and (7.38.a), the concluding inequality fails to be true for all small enough $\varepsilon > 0$. □

4^0. We have seen that for all small enough values of $\varepsilon > 0$ there exist functionals $\Psi^{(\varepsilon)}$ with the following properties:

A) $\Psi^{(\varepsilon)}$ is continuously differentiable on the entire H, and the derivative of the functional is Hölder continuous with exponent γ and constant independent of ε;

B) $\Psi^{(\varepsilon)}$ is zero outside the $\rho(\varepsilon)$-neighbourhood $U_\varepsilon = \{f \in H \mid 7\rho(\varepsilon) \leq \| f \| \leq 9\rho(\varepsilon)\}$ of the sphere $\{f \mid \| f \| = 8\rho(\varepsilon)\}$, where

$$\rho(\varepsilon) = \frac{1}{8}\left(\lfloor \varepsilon^{-\frac{2}{2\beta'+1}} \rfloor\right)^{-\beta};$$

C) There exists $\mathcal{F}_\varepsilon \subset \Sigma \cap U_\varepsilon$ such that

$$\inf_{\widehat{F}_\varepsilon} \sup_{f\in\mathcal{F}_\varepsilon} \mathcal{E}\left\{\left[\widehat{F}_\varepsilon(y^{f,\varepsilon}) - \Psi^{(\varepsilon)}(f)\right]^2\right\} \geq C\varepsilon^{\frac{4\beta(1+\gamma)}{2\beta'+1}}$$

with some positive C independent of ε.

Note that property C) clearly is preserved under arbitrary modification of $\Psi^{(\varepsilon)}$ which does not vary the functional in U_ε.

Now let us choose a decreasing sequence of positive reals ε_i which converges to 0 so fast that the "outer boundary" of $U_{\varepsilon_{i+1}}$ is inside the "inner boundary" of U_{ε_i} (see B)), and let us set

$$\Psi(f) = \sum_{i=1}^{\infty} \Psi^{(\varepsilon_i)}(f);$$

note that Ψ is well-defined, since at every point f at most one of the terms of the right hand side series differs from 0. Moreover, from A) combined with the fact that $\{U_{\varepsilon_i}\}$ are mutually disjoint it follows that Ψ satisfies **A.2**. We claim that the functional Ψ cannot be evaluated ε^2-consistently on Σ, which is immediate: since Ψ coincides with $\Psi^{(\varepsilon_i)}$ in U_{ε_i}, from C) and the remark accompanying this statement it follows that

$$\inf_{\widehat{F}_{\varepsilon_i}} \sup_{f\in\Sigma} \varepsilon_i^{-2}\mathcal{E}\left\{\left[\widehat{F}_{\varepsilon_i}(y^{f,\varepsilon_i}) - \Psi(f)\right]^2\right\} \geq C\varepsilon_i^{2\frac{2\beta-2\beta'-1+2\beta\gamma}{2\beta'+1}} \to \infty,\ i \to \infty$$

(see (7.38.b)), as claimed. ∎

7.3 Increasing smoothness of F

As we have seen, the sharp link between the "asymptotical width" β of the set of signals Σ and the "degree of smoothness" γ of the functional F we intend to estimate in an asymptotically efficient on Σ fashion is given by the inequality $\gamma > \frac{1}{2\beta}$. It follows that the "wider" is Σ (the less is β in **A.1**), the more smooth should be F. Note that the outlined tradeoff is possible in a restricted range of values of β only: since $\gamma \leq 1$, the "width" parameter β should be $> 1/2$. If we are interested to work with "wider" signal sets – those satisfying **A.1** with $\beta \leq 1/2$ – we should impose stronger requirements on the degree of smoothness of F and switch from the estimates based on the first-order approximation of F to those based on higher-order approximations. The general scheme of the associated estimates is quite transparent: in order to estimate the value of a $k \geq 1$ times differentiable functional F via observations (7.1) of the argument f, we choose somehow $m = m(\varepsilon)$, ε being the noise intensity, build an orthonormal basis where signals from Σ can be approximated as tight as possible by their order m projection estimates $\widehat{f}_m$ and write

$$F(f) \approx \sum_{\ell=0}^{k} \frac{1}{\ell!} D^\ell F(\widehat{f}_m)[f - f^m]_\ell, \tag{7.44}$$

where f^m is the projection of f on the linear span of the first m basic orths,

$$D^\ell F(f)[h_1,...,h_\ell] = \left.\frac{\partial^\ell}{\partial t_1 \partial t_2 ... \partial t_\ell}\right|_{t=0} F(f + t_1 h_1 + ... + t_\ell h_\ell)$$

is the value of ℓ-th differential of F taken at f along the set of directions $h_1,...,h_\ell$, and

$$D^\ell F(f)[h]_\ell = D^\ell F(f)[h,...,h]$$

is the ℓ-th derivative of F taken at f in a direction h. In order to estimate $F(f)$, we use the observations of the first m coordinates of f in our basis to build $\widehat{f}_m$ and therefore – to build the polynomials of $f - f^m$ in the right hand side of (7.44). After these polynomials are built, we use the observations of the remaining coordinates of f (i.e., those of the coordinates of $f - f^m$) in order to estimate the right hand side in (7.44). Note that the estimate we dealt with in the previous section is given by the outlined construction as applied with $k = 1$.

As we shall see in a while, passing from first-order local approximations of F to higher-order approximations allows to get “sharp” tradeoff between the “asymptotical width” of Σ and the degree of smoothness of F in the entire range $\beta > 0$ of the values of the width parameter. However, the implementation of this scheme heavily depends on whether $k \leq 2$ or $k \geq 3$; in the second case, a completely new curious phenomenon occurs. We postpone the case of $k \geq 3$ till the next chapter, and are about to complete the current one with considering the case of $k = 2$ (which is quite similar to the case of $k = 1$ we are already acquainted with).

7.3.1 The case of twice continuously differentiable functional

We are about to replace the assumption **A.2** with

A.3. The functional F to be estimated is defined in the ball

$$\mathcal{O}_{2\rho} = \{f \mid \parallel f \parallel < 1 + 2\rho\} \qquad [\rho > 0],$$

is twice continuously Fréchet differentiable in $\mathcal{O}_{2\rho}$, and its second derivative $F''(\cdot)$ (which is a symmetric bounded linear operator on H) is Hölder continuous in $\mathcal{O}_{2\rho}$ with exponent $\gamma > 0$ and constant L:

$$\forall f, g \in \mathcal{O}_{2\rho} : \quad \parallel F''(f) - F''(g) \parallel \leq L \parallel f - g \parallel^\gamma; \tag{7.45}$$

here for a bounded linear operator A on H $\parallel A \parallel$ is the operator norm of A:

$$\parallel A \parallel = \sup\{\parallel Ah \parallel \mid \parallel h \parallel \leq 1\}.$$

Note that the Gateau functional (7.3) with twice differentiable in f integrand $G(x,t)$ does *not* satisfy **A.3**, except the case when $G(x,t)$ is quadratic in t for almost all x and this integrand defines a continuous quadratic form on H (to this end $G(\cdot,0)$ should belong to $L_1[0,1]$, $G'_t(\cdot,0)$ should belong to $L_2[0,1]$ and $G''_{ff}(\cdot,0)$ should belong to $L_\infty[0,1]$). Similarly, in order to satisfy **A.3**, the Gateau functional (7.4) should have quadratic with respect to every t_i integrand

$G(x_1, ..., x_k, t_1, ..., t_k)$, the coefficient at t_i^2 depending on the x-variables only; an interesting example of this type is a "homogeneous Gateau polynomial"

$$F(f) = \int_0^1 ... \int_0^1 G(x_1, ..., x_k) f(x_1) ... f(x_k) dx_1 ... dx_k \tag{7.46}$$

with square summable kernel $G(x_1, ..., x_k)$.

We are about to prove the following extension of Theorem 7.2.1:

Theorem 7.3.1 *Let assumptions* **A.1**, **A.3** *be satisfied, and let*

$$\gamma > \frac{1}{2\beta} - 1. \tag{7.47}$$

Then F admits an asymptotically efficient on Σ estimation method.

Proof. Let us build the estimation method as follows.

Setup. Given noise intensity $\varepsilon < 0.1$, we set

$$\begin{array}{rcl} m = m(\varepsilon) & = & \lfloor \varepsilon^{-\frac{2}{2\beta+1}} \rfloor, \\ M = M(\varepsilon) & = & \lfloor \frac{1}{\varepsilon^2 \ln(1/\varepsilon)} \rfloor; \end{array} \tag{7.48}$$

note that $M > 2m$, provided that ε is small enough (as it is assumed in the sequel).

According to **A.1**, we may find m-dimensional and $(M-m)$-dimensional subspaces H_m, H_{M-m} in H in such a way that

$$\text{dist}(f, H_m) \le C m^{-\beta}, \text{dist}(f, H_{M-m}) \le C M^{-\beta} \quad \forall f \in \Sigma$$

(as above, C's stand for positive quantities depending on the data in **A.1**, **A.3** only and independent of ε and of a particular choice of $f \in \Sigma$). It follows that we may choose an orthonormal basis in H in such a way that H_m is the linear span of the first m vectors of the basis, while $H_m + H_{M-m}$ is contained in the linear span of the first M vectors from the basis; without loss of generality, we may assume that this basis is our original basis $\{\phi_i\}_{i=1}^{\infty}$. Denoting by f^ℓ the projection of $f \in H$ on the linear span of the first ℓ vectors of the basis, we therefore get

$$\| f - f^\ell \| \le C\ell^{-\beta}, \qquad \ell = m \text{ and } \ell = M. \tag{7.49}$$

Now, by **A.1** the closure of Σ is a compact set, and since by **A.3** F' is Lipschitz continuous on $\text{cl}\,\Sigma$, the image of $\text{cl}\,\Sigma$ under the mapping $f \mapsto F'(f)$ also is a compact set. Consequently, the quantities

$$\| f - f^N \|, \| F'(f) - [F'(f)]^N \|$$

converge to 0 as $N \to \infty$ uniformly in $f \in \Sigma$. Since by **A.3** both F and F' are Lipschitz continuous on Σ, there exists $N = N(\varepsilon) > M(\varepsilon)$ such that

$$\forall f \in \Sigma : \quad \| F(f) - F(f^N) \| \le \varepsilon^4, \ \| F'(f) - [F'(f^N)]^N \| \le \varepsilon^4. \tag{7.50}$$

The estimate $\widehat{F}_\varepsilon$ of F via observations (7.1) is as follows.

1) We use the observations $y_i^{f,\varepsilon}$, $i \le m(\varepsilon)$, to build the projection estimate

$$\widehat{f}_m = \sum_{i=1}^m y_i^{f,\varepsilon}\phi_i = f^m + \varepsilon\sum_{i=1}^m \xi_i\phi_i \tag{7.51}$$

and then "correct" it to get the estimate

$$\widetilde{f}_m = \begin{cases} \widehat{f}_m, & \| \widehat{f}_m \| \le 1+\rho \\ (1+\rho) \| \widehat{f} \|_m^{-1} \widehat{f}_m, & \| \widehat{f}_m \| > 1+\rho \end{cases},$$

exactly as in the construction used to prove Theorem 5.3.1; in particular, we ensure (7.20) and (7.31).

2) In what follows, $f \in \Sigma$, y stands for the observation $y^{f,\varepsilon}$ and ξ is the corresponding sequence of noises. For a pair of nonnegative integers p, q with $p \le q$ we set

$$\begin{array}{rcl} f_p^q &=& \sum\limits_{i=p}^q (f,\phi_i)\phi_i, \\ \xi_p^q &=& \sum\limits_{i=p}^q \xi_i\phi_i, \\ y_p^q &=& \sum\limits_{i=p}^q y_i\phi_i = f_p^q + \varepsilon\sum\limits_{i=p}^q \xi_i\phi_i = f_p^q + \varepsilon\xi_p^q; \end{array}$$

We write f_1^q, y_1^q, ξ_1^q simply as f^q, y^q, ξ^q.

Our estimate is

$$\begin{array}{c} \widehat{F}_\varepsilon = F(\widetilde{f}_m) + (F'(\widetilde{f}_m), y_{m+1}^N) + \frac{1}{2}(F''(\widetilde{f}_m)y_{m+1}^M, y_{m+1}^N + y_{M+1}^N) - \frac{\varepsilon^2}{2}\sum\limits_{i=1}^M F''_{ii}(\widetilde{f}_m), \\ {[m = m(\varepsilon), M = M(\varepsilon),\ \text{see (7.48)};\ \ N = N(\varepsilon),\ \text{see (7.50)}]} \end{array} \tag{7.52}$$

where $F''_{ij}(\widetilde{f}_m)$ are the entries of the matrix of the operator $F''(\widetilde{f}_m)$ in the basis $\{\phi_i\}$.

The origin of the estimate is as follows. It is more convenient to think that we are estimating $F(f^N)$ rather than $F(f)$ – these two quantities, in view of (7.50), differ from each other by no more than ε^4, while the rate of convergence we are interested to get is $O(\varepsilon)$; at the same time, when estimating $F(f^N)$, we should not bother about convergence of infinite series. Now, we have

$$\begin{array}{rcl} F(f^N) &\approx& F(f^M) + (F'(f^M), f_{M+1}^N), \\ F(f^M) &\approx& F(f^m) + (F'(f^m), f_{m+1}^M) + \frac{1}{2}(F''(f^m)f_{m+1}^M, f_{m+1}^M), \\ F'(f^M) &\approx& F'(f^m) + F''(f^m)f_{m+1}^M; \end{array}$$

combining these approximations, we come to

$$\begin{array}{rcl} F(f^N) &\approx& \left[F(f^m) + \left(F'(f^m), f_{m+1}^M\right) + \frac{1}{2}\left(F''(f^m)f_{m+1}^M, f_{m+1}^M\right)\right] \\ && + \left(F'(f^m) + F''(f^m)f_{m+1}^M, f_{M+1}^N\right) \\ &=& F(f^m) + \left(F'(f^m), f_{m+1}^N\right) + \frac{1}{2}\left(F''(f^m)f_{m+1}^M, f_{m+1}^N + f_{M+1}^N\right). \end{array} \tag{7.53}$$

Our concluding step is to replace in the resulting approximation the value and the derivatives of F at f^m with the value and derivatives at $\widetilde{f}_m$ and the vectors f_p^q with their observations y_p^q. We should, however, take care of suppressing the ε^2-terms in the bias resulting from this substitution. There are two sources of ε^2-terms in the bias:

1) When replacing $F(f^m)$ with $F(\widetilde{f}_m)$, the resulting error is, approximately,

$$(F'(f^m), \widehat{f}_m - f_m) + \frac{1}{2}(F''(f^m)(\widehat{f}_m - f^m), \widehat{f}_m - f^m)$$

(recall that $\widetilde{f}_m$ and $\widehat{f}_m$ coincide with probability close to 1); a good approximation to the expectation of this error is

$$\frac{\varepsilon^2}{2}\sum_{i=1}^{m} F''_{ii}(f^m) \approx \frac{\varepsilon^2}{2}\sum_{i=1}^{m} F''_{ii}(\widetilde{f}^m);$$

2) When replacing f_p^q with y_p^q, the ε^2-terms in the expectation of the resulting error are the same as in the expectation of

$$\frac{1}{2}\left(F''(f^m)\sum_{i=m+1}^{M}\xi_i\phi_i, \sum_{i=m+1}^{N}\xi_i\phi_i + \sum_{i=M+1}^{N}\xi_i\phi_i\right),$$

i.e., their sum is

$$\frac{\varepsilon^2}{2}\sum_{i=m+1}^{M} F''_{ii}(f^m) \approx \frac{\varepsilon^2}{2}\sum_{i=m+1}^{M} F''_{ii}(\widetilde{f}_m)$$

Thus, a natural way to convert approximation (7.53) into an estimate of $F(f^N)$ is to plug in the right hand side $\widetilde{f}_m$ instead of f^m and y_p^q instead of f_p^q, subtracting simultaneously the principal term of the bias, which is $\frac{\varepsilon^2}{2}\sum\limits_{i=1}^{M} F''_{ii}(\widetilde{f}_m)$; the resulting estimate is exactly (7.52).

Accuracy analysis. Note that for small enough values of ε, all $f \in \Sigma$ and all realizations of observation noise the points $\widetilde{f}_m + f_{m+1}^N$ and $\widetilde{f}_m$ belong to $\mathcal{O}_{2\rho}$. Indeed, the latter point, by construction, belongs to $\mathcal{O}_\rho$, while $\| f_{m+1}^N \| \leq Cm^{-\beta}(\varepsilon)$ by (7.49), so that $\| f_{m+1}^N < \rho$, provided that ε is small.

Setting

$$G = F(\widetilde{f}_m) + (F'(\widetilde{f}_m), f_{m+1}^N) + \frac{1}{2}(F''(\widetilde{f}_m)f_{m+1}^N, f_{m+1}^N), \tag{7.54}$$

we have

$$\begin{aligned}\widehat{F}_\varepsilon - F(f^N) \;&=\; \underbrace{G - F(\widetilde{f}_m + f_{m+1}^N)}_{A} + \underbrace{\widehat{F}_\varepsilon + \frac{\varepsilon^2}{2}\sum_{i=1}^{m} F''_{ii}(\widetilde{f}_m) - G}_{B} \\ &\quad + \underbrace{F(\widetilde{f}_m + f_{m+1}^N) - F(f^N) - \frac{\varepsilon^2}{2}\sum_{i=1}^{m} F''_{ii}(\widetilde{f}_m)}_{D}\end{aligned} \tag{7.55}$$

As it was already explained, A is well-defined for small enough values of ε, and in the sequel we assume that ε meets this requirement.

We should prove that for all $f \in \Sigma$ we have

$$\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - F(f)\right]^2\right\} \leq \varepsilon^2 \parallel F'(f) \parallel^2 + \varepsilon^2 o(1);$$

from now on, all $o(1)$ stand for deterministic functions of ε independent of $f \in \Sigma$ and converging to 0 as $\varepsilon \to +0$. In view of (7.50), in fact we should verify that

$$\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - F(f^N)\right]^2\right\} \leq \varepsilon^2 \parallel F'(f) \parallel^2 + \varepsilon^2 o(1),$$

or, which is the same in view of (7.55), that

$$\mathcal{E}\left\{(A+B+C)^2\right\} \leq \varepsilon^2 \parallel F'(f) \parallel^2 + \varepsilon^2 o(1) \tag{7.56}$$

We claim that

A)

$$|A| \leq \varepsilon o(1) \tag{7.57}$$

B)

$$\begin{array}{rrcl} (a) & |\mathcal{E}\{B|\xi^m\}| & \leq & o(1)\varepsilon \\ (b) & \mathcal{E}\{B^2|\xi^m\} & \leq & \varepsilon^2 \sum\limits_{i=m+1}^{N} [F'(\widetilde{f}_m)]_i^2 + \varepsilon^2 o(1) \quad [\text{for } g \in H,\ g_i = (g, \phi_i)] \end{array} \tag{7.58}$$

here $\mathcal{E}\{\cdot|\xi^m\}$ is the conditional expectation, the noises $\xi^m = (\xi_1, ..., \xi_m)$ being fixed;

C)

$$\mathcal{E}\left\{D^2\right\} \leq \varepsilon^2 \sum_{i=1}^{m} [F'(f^N)]_i^2 + \varepsilon^2 o(1) \tag{7.59}$$

D)

$$\mathcal{E}\left\{\sum_{i=m+1}^{N} [F'(\widetilde{f}_m)]_i^2\right\} \leq \sum_{i=m+1}^{N} [F'(f^N)]_i^2 + o(1). \tag{7.60}$$

Let us check that A) – D) imply (7.56). Indeed, we have

$$\begin{aligned}
& \mathcal{E}\left\{(A+B+D)^2\right\} \\
\leq\ & (1+o(1))\mathcal{E}\left\{(B+D)^2\right\} + \varepsilon^2 o(1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by A)]} \\
\leq\ & (1+o(1))\left[\mathcal{E}\left\{B^2\right\} + 2\mathcal{E}_{\xi^m}\left\{D\mathcal{E}\left\{B|\xi^m\right\}\right\} + \mathcal{E}\left\{D^2\right\}\right] + \varepsilon^2 o(1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad \text{[since } D \text{ depends on } \xi^m \text{ only]} \\
\leq\ & (1+o(1))\left\{\varepsilon^2\left[\mathcal{E}_{\xi^m}\left\{\sum_{i=m+1}^{N}[F'(\widetilde{f}_m)]_i^2\right\} + o(1)\right] + o(1)\varepsilon\mathcal{E}_{\xi^m}\left\{|D|\right\} + \mathcal{E}\left\{D^2\right\}\right\} \\
& +\varepsilon^2 o(1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by B)]} \\
\leq\ & (1+o(1))\left\{\varepsilon^2\mathcal{E}_{\xi^m}\left\{\sum_{i=m+1}^{N}[F'(\widetilde{f}_m)]_i^2\right\} + o(1)\varepsilon\sqrt{\mathcal{E}_{\xi^m}\left\{D^2\right\}} + \mathcal{E}\left\{D^2\right\}\right\} + \varepsilon^2 o(1) \\
\leq\ & (1+o(1))\left\{\varepsilon^2\mathcal{E}_{\xi^m}\left\{\sum_{i=m+1}^{N}[F'(\widetilde{f}_m)]_i^2\right\} + \varepsilon^2 o(1) + \varepsilon^2\sum_{i=1}^{m}[F'(f^N)]_i^2\right\} + \varepsilon^2 o(1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by C)]} \\
\leq\ & \varepsilon^2\sum_{i=1}^{N}[F'(f^N)]_i^2 + \varepsilon^2 o(1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by D)]} \\
=\ & \varepsilon^2 \parallel F'(f) \parallel^2 + \varepsilon^2 o(1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by (7.50)]}
\end{aligned}$$

It remains to verify A) – D)

Verifying A) We have

$$\begin{aligned}
|A| &= |G - F(\widetilde{f}_m + f_{m+1}^N)| \\
&= \left|F(\widetilde{f}_m) + (F'(\widetilde{f}_m), f_{m+1}^N) + \tfrac{1}{2}(F''(\widetilde{f}_m)f_{m+1}^N, f_{m+1}^N) - F(\widetilde{f}_m + f_{m+1}^N)\right| \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[origin of } G\text{]} \\
&\leq C \parallel f_{m+1}^N \parallel^{2+\gamma} \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by } \mathbf{A.3}\text{]} \qquad \square \\
&\leq Cm^{-\beta(2+\gamma)}(\varepsilon) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by (7.49)]} \\
&\leq C\varepsilon^{\frac{2\beta(2+\gamma)}{2\beta+1}} \\
&\leq \varepsilon o(1) \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{[by (7.47)]}
\end{aligned}$$

Verifying B) We have

$$\begin{array}{rcl}
B & = & \widehat{F}_\varepsilon + \frac{\varepsilon^2}{2}\sum\limits_{i=1}^{m} F''_{ii}(\widetilde{f}_m) - G \\
& = & F(\widetilde{f}_m) + (F'(\widetilde{f}_m), y^N_{m+1}) + \frac{1}{2}(F''(\widetilde{f}_m)y^M_{m+1}, y^N_{M+1} + y^N_{m+1}) - \frac{\varepsilon^2}{2}\sum\limits_{i=1}^{M} F''_{ii}(\widetilde{f}_m) \\
& & -F(\widetilde{f}_m) - (F'(\widetilde{f}_m), f^N_{m+1}) - \frac{1}{2}(F''(\widetilde{f}_m)f^N_{m+1}, f^N_{m+1}) \\
& = & \underbrace{\varepsilon(F'(\widetilde{f}_m), \xi^N_{m+1})}_{B_1} + \underbrace{\frac{\varepsilon^2}{2}\left[(F''(\widetilde{f}_m)\xi^M_{m+1}, \xi^M_{m+1}) - \sum\limits_{i=m+1}^{M} F''_{ii}(\widetilde{f}_m)\right]}_{B_2} \\
& & -\underbrace{\frac{1}{2}(F''(\widetilde{f}_m)f^N_{M+1}, f^N_{M+1})}_{B_3} + \underbrace{\frac{\varepsilon}{2}(F''(\widetilde{f}_m)f^M_{m+1}, \xi^N_{M+1})}_{B_4} \\
& & +\underbrace{\frac{\varepsilon}{2}(F''(\widetilde{f}_m)f^M_{m+1}, \xi^N_{m+1})}_{B_5} + \underbrace{\frac{\varepsilon}{2}(F''(\widetilde{f}_m)\xi^M_{m+1}, f^N_{M+1})}_{B_6} \\
& & +\underbrace{\frac{\varepsilon}{2}(F''(\widetilde{f}_m)\xi^M_{m+1}, f^N_{m+1})}_{B_7} + \underbrace{\varepsilon^2(F''(\widetilde{f}_m)\xi^M_{m+1}, \xi^N_{M+1})}_{B_8}
\end{array} \tag{7.61}$$

(7.58.a): Among the terms $B_1 - B_8$ in (7.61), the only one with nonzero conditional, ξ^m fixed, expectation is B_3, so that

$$\begin{array}{rcll}
|\mathcal{E}\{B|\xi^m\}| & = & \left|(F''(\widetilde{f}_m)f^N_{M+1}, f^N_{M+1})\right| & \\
& \leq & C \parallel f^N_{M+1} \parallel^2 & [\text{by } \mathbf{A.3}] \\
& \leq & CM^{-2\beta}(\varepsilon) & [\text{by } (7.49)] \\
& \leq & C\left(\varepsilon\sqrt{\ln(1/\varepsilon)}\right)^{4\beta} & [\text{by } (7.48)] \\
& = & \varepsilon o(1) & [\text{since } \beta > 1/4 \text{ by } (7.47)]
\end{array} \qquad \square$$

(7.58.b): It suffices to demonstrate that

$$\mathcal{E}\left\{B_1^2|\xi^m\right\} = \varepsilon^2 \sum_{i=m+1}^{N} [F'(\widetilde{f}_m)]_i^2 \tag{7.62}$$

(which is evident) and that

$$\mathcal{E}\left\{B_\ell^2|\xi^m\right\} \leq \varepsilon^2 o(1), \ \ell = 2, 3, ..., 8. \tag{7.63}$$

(7.63) for $\ell = 2$: We have

$$\begin{array}{rcl}
\mathcal{E}\{B_2^2|\xi^m\} & = & \frac{\varepsilon^4}{4}\mathcal{E}\left\{\left[(F''(\widetilde{f}_m)\xi^M_{m+1}, \xi^M_{m+1}) - \sum\limits_{i=m+1}^{M} F''_{ii}(\widetilde{f}_m)\right]^2 |\xi^m\right\} \\
& = & \frac{\varepsilon^4}{4}\mathcal{E}\left\{\left[\sum\limits_{i,j=m+1}^{M} F''_{ij}(\widetilde{f}_m)(\xi_i\xi_j - \delta_{ij})\right]^2 |\xi^m\right\} \\
& = & \frac{\varepsilon^4}{4}\sum\limits_{i,j=m+1}^{M}[F''_{ij}(\widetilde{f}_m)]^2(2-\delta_{ij})\mathcal{E}\left\{(\xi_i\xi_j - \delta_{ij})^2\right\} \\
& \leq & C\varepsilon^4 M \\
& & [\text{since} \parallel F''(\widetilde{f}_m) \parallel \leq C \text{ by } \mathbf{A.3}\text{, whence } \sum\limits_j [F''_{ij}(\widetilde{f}_m)]^2 \leq C \ \forall i] \\
& \leq & \varepsilon^2 o(1) \\
& & [\text{by } (7.48)]
\end{array} \qquad \square$$

(7.63) for $\ell = 3$: We have

$$\begin{array}{rcll}
\mathcal{E}\left\{B_3^2|\xi^m\right\} & \le & C \parallel f^N_{M+1} \parallel^4 & \text{[by } \mathbf{A.3}\text{]}\\
& \le & CM^{-4\beta}(\varepsilon) & \text{[by (7.49)]}\\
& \le & C\left(\varepsilon^2\ln(1/\varepsilon)\right)^{4\beta} & \text{[by (7.49)]}\\
& \le & \varepsilon^2 o(1) & \text{[since } 4\beta > 1 \text{ due to (7.47) combined with } \gamma \le 1\text{]}
\end{array}$$

□

(7.63) for $\ell = 4$: We have

$$\begin{array}{rcll}
\mathcal{E}\left\{B_4^2|\xi^m\right\} & \le & \frac{\varepsilon^2}{4} \parallel F''(\widetilde{f}_m) f^M_{m+1} \parallel^2 & \\
& \le & C\frac{\varepsilon^2}{4} \parallel f^M_{m+1} \parallel^2 & \text{[by } \mathbf{A.3}\text{]}\\
& = & \varepsilon^2 o(1) & \text{[by (7.49)]}
\end{array}$$

□

(7.63) for $\ell = 5, 6, 7$: completely similar to the case of $\ell = 4$.

(7.63) for $\ell = 8$: We have

$$\begin{array}{rcll}
\mathcal{E}\left\{B_8^2|\xi^m\right\} & \le & \varepsilon^4 \mathcal{E}\left\{\parallel F''(\widetilde{f}_m)\xi^M_{m+1} \parallel^2 |\xi^m\right\} & \text{[since } \xi^M_{m+1} \text{ is independent of } \xi^N_{M+1}\text{]}\\
& \le & C\varepsilon^4 \mathcal{E}\left\{\parallel \xi^M_{m+1} \parallel^2\right\} & \text{[by } \mathbf{A.3}\text{]}\\
& = & C\varepsilon^4 M & \\
& \le & \varepsilon^2 o(1) & \text{[by (7.48)]}
\end{array}$$

□

B) is proved.

Verifying C) We have

$$\begin{array}{rl}
& D\\
= & F(\widetilde{f}_m + f^N_{m+1}) - F(f^N) - \frac{\varepsilon^2}{2}\sum\limits_{i=1}^m F''_{ii}(\widetilde{f}_m)\\
= & \underbrace{(F'(f^N), \widehat{f}_m - f^m)}_{D_1} + \underbrace{(F'(f^N), \widetilde{f}_m - \widehat{f}_m)}_{D_2}\\
& + \underbrace{F(\widetilde{f}_m + f^N_{m+1}) - F(f^N) - (F'(f^N), \widetilde{f}_m - f^m) - \frac{1}{2}\left(F''(f^N)(\widetilde{f}_m - f^m), \widetilde{f}_m - f^m\right)}_{D_3}\\
& + \underbrace{\frac{1}{2}\left[\left(F''(f^N)(\widetilde{f}_m - f^m), \widetilde{f}_m - f^m\right) - \varepsilon^2\sum\limits_{i=1}^m F''_{ii}(\widetilde{f}_m)\right]}_{D_4}
\end{array} \tag{7.64}$$

To establish C), it suffices to verify that

$$\mathcal{E}\left\{D_1^2\right\} = \varepsilon^2\sum_{i=1}^M [F'(f^N)]_i^2$$

(which is evident) and that

$$\mathcal{E}\left\{D_\ell^2\right\} \le \varepsilon^2 o(1),\ \ \ell = 2, 3, 4. \tag{7.65}$$

(7.65) for $\ell = 2$: We have

$$\begin{array}{rcll}
\mathcal{E}\left\{D_2^2\right\} & \le & C\mathcal{E}\left\{\parallel \widetilde{f}_m - \widehat{f}_m \parallel^2\right\} & \text{[by } \mathbf{A.3}\text{]}\\
& \le & C(\theta)(\varepsilon^2 m(\varepsilon))^{1+\theta}\ \forall\theta \ge 1 & \text{[by (7.31)]}\\
& \le & \varepsilon^2 o(1) & \text{[choose } \theta \text{ appropriately]}
\end{array}$$

□

<u>(7.65) for $\ell = 3$:</u> We have

$$\begin{array}{rl} & |F(\widetilde{f}_m + f^N_{m+1}) - F(f^N) - \left(F(f^N), \widetilde{f}_m - f^m\right) \\ & -\frac{1}{2}\left(F''(f^N)(\widetilde{f}_m - f^m), \widetilde{f}_m - f^m\right)| \\ \leq & C \parallel \widetilde{f}_m - f^m \parallel^{2+\gamma} \end{array}$$

by **A.3**, whence

$$\begin{array}{rcll} \mathcal{E}\{D_3^2\} & \leq & C\mathcal{E}\left\{\parallel \widetilde{f}_m - f^m \parallel^{2(2+\gamma)}\right\} & \\ & \leq & C\mathcal{E}\left\{\parallel \widehat{f}_m - f^m \parallel^{2(2+\gamma)}\right\} & [\text{by } (7.20.b] \\ & \leq & C[\varepsilon^2 m(\varepsilon)]^{2+\gamma} & [\text{by } (7.30)] \\ & \leq & C\varepsilon^{\frac{4\beta(2+\gamma)}{2\beta+1}} & [\text{by } (7.48)] \\ & \leq & \varepsilon^2 o(1) & [\text{by } (7.47)] \end{array} \qquad \square$$

<u>(7.65) for $\ell = 4$:</u> We have

$$\begin{array}{rcl} 2D_4 & = & \left(F''(f^N)(\widetilde{f}_m - f^m), \widetilde{f}_m - f^m\right) - \varepsilon^2 \sum\limits_{i=1}^m F''_{ii}(\widetilde{f}_m) \\ & = & \underbrace{\left(F''(f^N)(\widehat{f}_m - f^m)\right), \widehat{f}_m - f^m) - \varepsilon^2 \sum_{i=1}^m F''_{ii}(\widetilde{f}_m)}_{D_{4,1}} \\ & & + \underbrace{\left(F''(f^N)(\widetilde{f}_m - f^m), \widetilde{f}_m - f^m\right) - \left(F''(f^N)(\widehat{f}_m - f^m), \widehat{f}_m - f^m\right)}_{D_{4,2}} \\ & & + \underbrace{\varepsilon^2 \sum_{i=1}^m (F''_{ii}(f^N) - F''_{ii}(\widetilde{f}_m))}_{D_{4,3}} \end{array}$$

and in order to establish (7.65) for $\ell = 4$ it suffices to verify that

$$\mathcal{E}\left\{D^2_{4,\kappa}\right\} \leq \varepsilon^2 o(1), \ \kappa = 1, 2, 3. \tag{7.66}$$

<u>(7.66) for $\kappa = 1$:</u> We have

$$\begin{array}{rcl} \mathcal{E}\left\{D^2_{4,1}\right\} & = & \varepsilon^4 \mathcal{E}\left\{\left[\sum\limits_{i,j=1}^m F''_{ij}(f^N)(\xi_i\xi_j - \delta_{ij})\right]^2\right\} \\ & = & \varepsilon^4 \sum\limits_{i,j=1}^m \left[F''_{ij}(f^N)\right]^2 (2 - \delta_{ij}) \mathcal{E}\left\{(\xi_i\xi_j - \delta_{ij})^2\right\} \\ & \leq & C\varepsilon^4 m(\varepsilon) \\ \multicolumn{3}{l}{[\text{since } \parallel F''(f^N) \parallel \leq C \text{ by } \mathbf{A.3}, \text{ whence } \sum\limits_j [F''_{ij}(f^N)]^2 \leq C \ \forall i]} \\ & \leq & \varepsilon^2 o(1) \\ & & \qquad\qquad [\text{by } (7.48)] \end{array} \qquad \square$$

(7.66) for $\kappa = 2$: We have

$$\begin{aligned}
& \mathcal{E}\left\{D_{4,2}^2\right\} \\
= \;& \mathcal{E}\Big\{\Big| \Big(F''(f^N)(\widetilde{f}_m - f^m), \widetilde{f}_m - f^m\Big) \\
& - \Big(F''(f^N)(\widehat{f}_m - f^m), \widehat{f}_m - f^m\Big)\Big|^2\Big\} \\
\leq \;& C\mathcal{E}\left\{\| \widehat{f}_m - \widetilde{f}_m \|^2 + \| \widehat{f}_m - \widetilde{f}_m \|^4\right\} && [\text{by } \mathbf{A.3}] \\
\leq \;& C(\theta)\left[[\varepsilon^2 m(\varepsilon)]^{1+\theta} + [\varepsilon^2 m(\varepsilon)]^{2+\theta}\right] \ \forall \theta \geq 1 && [\text{by } (7.31)] \\
\leq \;& C(\theta)\varepsilon^{\frac{4\beta(1+\theta)}{2\beta+1}} && [\text{by } (7.48)] \\
\leq \;& \varepsilon^2 o(1) && [\text{choose } \theta \text{ appropriately}]
\end{aligned}$$

□

(7.66) for $\kappa = 3$: We have

$$\begin{aligned}
\mathcal{E}\left\{D_{4,3}^2\right\} \;\leq\;& C\varepsilon^4\mathcal{E}\left\{\left[m \| F''(f^N) - F''(\widetilde{f}_m) \|\right]^2\right\} \\
\leq\;& C\varepsilon^4\mathcal{E}\left\{m^2 \| f^N - \widetilde{f}_m \|^{2\gamma}\right\} && [\text{by } \mathbf{A.3}] \\
\leq\;& C[\varepsilon^2 m]^2\mathcal{E}\left\{\| f - \widetilde{f}_m \|^{2\gamma}\right\} && \square \\
\leq\;& C[\varepsilon^2 m]^2\varepsilon^{\frac{4\beta\gamma}{2\beta+1}} && [\text{by } (7.20.c, d)] \\
\leq\;& C\varepsilon^{\frac{4\beta(2+\gamma)}{2\beta+1}} && [\text{by } (7.48)] \\
=\;& \varepsilon^2 o(1) && [\text{by } (7.47)]
\end{aligned}$$

C) is proved.

Verifying D) We have

$$\begin{aligned}
& \mathcal{E}\left\{\sum_{i=m+1}^{N}[F'(\widetilde{f}_m)]_i^2\right\} \\
\leq\;& \mathcal{E}\left\{(1+\theta)\sum_{i=m+1}^{N}[F'(f^N)]_i^2 + (1+\theta^{-1})\sum_{i=m+1}^{N}\delta_i^2\right\} \ \forall\theta > 0 \\
& [\delta_i = [F'(f^N) - F'(\widetilde{f}_m)]_i] \\
\leq\;& (1+\theta)\sum_{i=m+1}^{N}[F'(f^N)]_i^2 + (1+\theta^{-1})\mathcal{E}\left\{\| F'(f^N) - F'(\widetilde{f}_m) \|^2\right\} \\
\leq\;& (1+\theta)\sum_{i=m+1}^{N}[F'(f^N)]_i^2 + (1+\theta^{-1})C\mathcal{E}\left\{\| f^N - \widetilde{f}_m \|^2\right\} && \square \\
& [\text{by } \mathbf{A.3}] \\
\leq\;& (1+\theta)\sum_{i=m+1}^{N}[F'(f^N)]_i^2 + C(1+\theta^{-1})\varepsilon^{\frac{4\beta}{2\beta+1}} \\
& [\text{by } (7.20.c, d)] \\
\leq\;& \sum_{i=m+1}^{N}[F'(f^N)]_i^2 + o(1) \\
& [\text{set } \theta = \varepsilon^{\frac{2\beta}{2\beta+1}}]
\end{aligned}$$

The proof of Theorem 7.3.1 is completed. ■

7.3.2 Concluding remarks

Sharpness of (7.3.4). Relation (7.47) establishes "sharp" link between the asymptotical width of Σ (i.e., the parameter β) and the degree of smoothness γ of a functional satisfying **A.3** (cf. Section 7.2.1). A construction completely similar to the one used to prove Theorem 7.2.2 yields the following result:

Theorem 7.3.2 *Let $\gamma \in (0,1]$ and $\beta > 0$ be such that*

$$\gamma < \frac{1}{2\beta} - 1.$$

Then there exist a set $\Sigma \subset H$ and a functional $F : H \to \mathbf{R}$ satisfying **A.1**, **A.3** *such that F does not admit asymptotically efficient (even efficient in order) on Σ estimation method.*

The case of quadratic functional. Let $F(f) = (Af, f)$, where A is a bounded symmetric operator on H, and let Σ satisfy **A.1**. Consider the estimator resulting from (7.52) by letting $N \to \infty$ and replacing $\widetilde{f}_m$ with $\widehat{f}_m$. Tracing the proof of Theorem 7.3.1, one can see that in the case in question the squared risk of estimating $F(f)$ can be bounded from above as

$$\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - F(f)\right]^2\right\} \le \varepsilon^2 \parallel F'(f) \parallel^2 + C(\varepsilon^4 M + M^{-4\beta}) + \varepsilon^2 o(1) \qquad (7.67)$$

(C is independent of ε and $f \in \Sigma$); here $M > m(\varepsilon)$ is a "free design parameter" of the estimate[5]. Assuming that $\beta \le 1/4$ (the case of $\beta > 1/4$ is covered by Theorem 7.3.1) and setting

$$M = \lfloor \varepsilon^{-\frac{4\beta}{4\beta+1}} \rfloor,$$

we get an estimate of a *quadratic* functional with the squared risk satisfying the relation

$$\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - F(f)\right]^2\right\} \le \varepsilon^2 \parallel F'(f) \parallel^2 + C\varepsilon^{\frac{16\beta}{4\beta+1}} + \varepsilon^2 o(1)$$

(C is independent of $f \in \Sigma$ and of ε). We see that if the asymptotical width of Σ is $\beta \le \frac{1}{4}$, then a *quadratic* functional F can be estimated at points $f \in \Sigma$ with the squared risk not exceeding $\varepsilon^{\frac{16\beta}{4\beta+1}}$. It turns out (see [14]) that this rate of convergence is unimprovable in the minimax sense, provided that

$$d_k(\Sigma) \ge ck^{-\beta}, \ k = 1, 2, ...$$

and that for some $\kappa > 0$ it holds

$$F(f) \ge \kappa \parallel f \parallel^2 \quad \forall f \in H.$$

[5] In (7.52), $M = M(\varepsilon)$ was controlled according to (7.48); the estimate, however, makes sense for other values of the parameter as well.

Chapter 8

Estimating functionals, II

We proceed with constructing asymptotically efficient, on a given compact set Σ, estimates of a smooth functional F of a nonparametric signal f via observations

$$y^{f,\varepsilon} = \left\{ y_i^{f,\varepsilon} \equiv (f, \phi_i) + \varepsilon \xi_i \right\}, \tag{8.1}$$

of the signal ($\{\phi_i\}$ form an orthonormal basis in the Hilbert space H where the signals live, the "noise" $\{\xi_i\}_{i=1}^{\infty}$ is a collection of independent $\mathcal{N}(0,1)$ random variables).

8.1 Preliminaries: estimating polynomials

We already know that if the Kolmogorov diameters $d_k(\Sigma)$ admit an upper bound

$$d_k(\Sigma) \leq c k^{-\beta}$$

and F is κ times ($\kappa = 1, 2$) continuously differentiable in a neighbourhood of Σ functional with Hölder continuous, with exponent γ, κ-th derivative, then F can be asymptotically efficiently estimated on Σ, provided that the asymptotic width of Σ and the degree of smoothness of F are linked according to

$$\kappa + \gamma > 1 + \frac{1}{2\beta}. \tag{8.2}$$

The "widest" Σ we can handle corresponds to the case of $\kappa = 2$, $\gamma = 1$, where (8.2) requires from β to be $> 1/4$. As we remember, (8.2) is sharp; thus, when interested to deal with wider – $\beta \leq 1/4$ – sets of signals, we should impose stronger smoothness restrictions on F. On the other hand, it was mentioned in Section 7.3.2 that if

$$d_k(\Sigma) = O(k^{-\beta}),\ \beta < 1/4,$$

then some quadratic functionals – e.g., $F(f) = \| f \|^2$ – *cannot* be estimated on Σ with uniform squared risk of order ε^2. Since a quadratic functional is "as smooth as a functional can be", we conclude that merely increasing the number of derivatives F is assumed to possess does not help; we should impose certain *structural restrictions* on these derivatives. In order to understand what these restrictions could be, note that if we are planning to build asymptotically efficient, on a "wide" set Σ, estimators of F and the estimators we intend to construct are based on local approximations of F by its Taylor polynomials, then at least the polynomials involved should admit asymptotically efficient, on Σ, estimation. And since we intend to work with "wider and wider" sets of signals – i.e., with β approaching 0 – the above polynomials should admit asymptotically efficient estimation on the entire space H (or at least on any bounded subset of H). Indeed, if our "structural restrictions" on the derivatives of F are such that ε^2-consistent estimation of, say, the Taylor polynomial of degree 5 of F already imposes a nontrivial restriction on the asymptotical width of Σ, we have no hope to work successfully with β too close to 0.

Now, there is a very natural family of polynomials on H admitting asymptotically efficient estimation on all bounded subsets of H – the *Hilbert-Schmidt* polynomials, and this is the family we will work with.

8.1.1 Hilbert-Schmidt polynomials

Recall that a *homogeneous polynomial of degree* k on H is a function

$$P(f) = \Pi_k[\underbrace{f, ..., f}_{k \text{ times}}],$$

where $\Pi_k[f_1, ..., f_k]$ is a symmetric k-linear continuous form on H. Given an orthonormal basis $\{\phi_i\}$ in H, we may associate with Π_k (and therefore – with $P(\cdot)$) the system of *coefficients* $\{P_\iota = \Pi_k[\phi_{\iota_1}, ..., \phi_{\iota_k}]\}_{\iota \in \mathcal{I}_k}$, where $\mathcal{I}_k$ is the set of all k-dimensional multi-indices $\iota = (\iota_1, ..., \iota_k)$ with positive integer entries. We clearly have

$$P(f) = \lim_{N\to\infty} \sum_{\iota:\iota_p \le N,\ p=1,...,k} P_\iota f_{\iota_1} ... f_{\iota_k} \quad [f_i = (f, \phi_i)].$$

A homogeneous polynomial $P(f)$ of degree k is called a *Hilbert-Schmidt* polynomial, if

$$\| P \|_2 \equiv \sqrt{\sum_{\iota \in \mathcal{I}_k} P_\iota^2} < \infty;$$

$\| P \|_2$ is called the *Hilbert-Schmidt* norm of P. It can be proved that the Hilbert-Schmidt norm is independent of the (orthonormal) basis with respect to which the coefficients of P are taken. A generic example of a Hilbert-Schmidt polynomial is the Gateau polynomial

$$F(f) = \int_0^1 ... \int_0^1 G(x_1, ..., x_k) f(x_1) ... f(x_k) dx_1 ... dx_k$$

on $L_2[0, 1]$ with square-summable kernel G; the Hilbert-Schmidt norm of this polynomial is just the L_2-norm of the kernel.

A non-homogeneous polynomial P of degree $\le k$ is a sum of homogeneous polynomials P^p of degrees 0 (a constant), 1,..., k:

$$P(f) = \sum_{p=0}^{k} P^p(f).$$

P is called a Hilbert-Schmidt polynomial, if its homogeneous components $P^1, ..., P^k$ are so.

8.1.2 Estimating Hilbert-Schmidt polynomials

Let P be a Hilbert-Schmidt polynomial of degree $\leq k$ on H. We are about to demonstrate that such a polynomial admits asymptotically efficient, on every bounded subset of H, estimate. Let us fix an orthonormal basis $\{\phi_i\}$, and let

$$f^N = \sum_{i=1}^{N} (f, \phi_i)\phi_i$$

be the projection of $f \in H$ onto the linear span of the first N basic orths. Let also

$$P_N(f) = P(f^N).$$

We start with building an estimator of P_N via observations (8.1). Note that P_N is a polynomial of N real variables and therefore it can be naturally extended onto the complexification $\mathbf{C}^N$ of $\mathbf{R}^N$. Let ζ^N be a random N-dimensional Gaussian vector with zero mean and unit covariance matrix. For $z \in \mathbf{C}^N$, let

$$\widehat{P}_N(z) = \mathcal{E}\left\{P_N(z + i\varepsilon\zeta^N)\right\}.$$

i being the imaginary unit. Setting

$$y^N = y^N(f,\varepsilon) = \sum_{i=1}^{N} y_i^{f,\varepsilon}\phi_i = f^N + \varepsilon\xi^N, \ \xi^N = \sum_{i=1}^{N} \xi_i\phi_i,$$

consider the estimator

$$\widetilde{P}_N = \widehat{P}(y^N). \tag{8.3}$$

Theorem 8.1.1 *$\widetilde{P}_N$ is an unbiased estimator of P_N:*

$$\mathcal{E}\left\{\widetilde{P}_N(y^N(f,\varepsilon))\right\} = P_N(f) \quad \forall f \in H,$$

with the variance

$$\mathcal{E}\left\{\left[\widetilde{P}_N(y^N(f,\varepsilon)) - P_N(f)\right]^2\right\} = \sum_{p=1}^{k} \frac{\varepsilon^{2p}}{p!} \parallel D^p P_N(f) \parallel_2^2 . \tag{8.4}$$

Proof. Let $\omega^N = \xi^N + i\zeta^N$ (ζ^N is independent of ξ^N Gaussian vector with zero mean and unit covariance matrix). The distribution of ω^N remains invariant under rotations of $\mathbf{C}^N$ (viewed as a $2N$-dimensional real Euclidean space), while P_N is an analytic function on $\mathbf{C}^N$ and is therefore a harmonic function on $\mathbf{C}^N$ (again viewed as a $2N$-dimensional real space). Therefore

$$\begin{aligned}
\mathcal{E}\left\{\widetilde{P}_N(y^N(f,\varepsilon))\right\} &= \mathcal{E}_{\xi^N}\left\{\mathcal{E}_{\zeta^N}\left\{P_N(f^N + \varepsilon\xi^N + i\varepsilon\zeta^N)\right\}\right\} \\
&= \mathcal{E}_{\omega^N}\left\{P_N(f^N + \varepsilon\omega^N)\right\} \\
&= P_N(f),
\end{aligned}$$

the concluding equality being given by the Mean Value Theorem for harmonic functions.

Since $\widetilde{P}_N$ is unbiased, to determine the variance of the estimator at a fixed f we can confine ourselves to the case of $P_N(f) = 0$.

Let ρ^N be a random vector identically distributed like ξ^N, ζ^N and independent of these two vectors, and let $\omega^N = \xi^N + i\zeta^N$, $\lambda^N = \xi^N + i\rho^N$. Since $\widetilde{P}_N$ clearly is real-valued, we have

$$\mathcal{E}\left\{\widetilde{P}_N^2\right\} = \mathcal{E}\left\{P_N(f^N + \varepsilon\omega^N)P_N(f^N + \varepsilon\lambda^N)\right\},$$

whence, expanding P_N in a Taylor series around f^N,

$$\mathcal{E}\left\{\widetilde{P}_N^2\right\} = \mathcal{E}\left\{\sum_{p,q=1}^{k} \frac{1}{p!q!}\left[D^p P_N(f^N)[\varepsilon\omega^N]_p\right]\left[D^q P_N(f^N)[\varepsilon\lambda^N]_q\right]\right\}, \tag{8.5}$$

where $A[h]_p = A[\underbrace{h, ..., h}_{p \text{ times}}]$, $A[h_1, ..., h_p]$ being a p-linear form.

Let $\mathcal{J}_N^p$ be the set of multi-indices $\iota = (\iota_1, ..., \iota_N)$ with nonnegative entries and with $|\iota| \equiv \sum\limits_{j=1}^{N} \iota_j = p$. For $\iota \in \mathcal{J}_N^p$, $z = \sum\limits_{j=1}^{N} z_j\phi_j \in \mathbf{C}^N$ and $p = 1, ..., k$ let

$$\begin{array}{rcl} \iota! & = & \iota_1!...\iota_N!, \\ z^\iota & = & z_1^{\iota_1}...z_N^{\iota_N}, \\ P_\iota^p & = & D^p P_N(f^N)[\underbrace{\phi_1, ..., \phi_1}_{\iota_1}, \underbrace{\phi_2, ..., \phi_2}_{\iota_2}, ..., \underbrace{\phi_N, ..., \phi_N}_{\iota_N}]. \end{array}$$

We have

$$\left[D^p P_N(f^N)[\varepsilon\omega^N]_p\right]\left[D^q P_N(f^N)[\varepsilon\lambda^N]_q\right] = \sum_{\iota \in \mathcal{J}_N^p, \nu \in \mathcal{J}_N^q} \varepsilon^{p+q}\frac{p!q!}{\iota!\nu!}P_\iota^p P_\nu^q (\omega^N)^\iota (\lambda^N)^\nu. \tag{8.6}$$

Observe now that

$$\mathcal{E}\left\{(\omega^N)^\iota(\lambda^N)^\nu\right\} = \prod_{j=1}^{N} \mathcal{E}\left\{\omega_j^{\iota_j}\lambda_j^{\nu_j}\right\} = \prod_{j=1}^{N}\left[\delta_{\iota_j\nu_j}\iota_j!\right]. \tag{8.7}$$

Indeed, all we need to verify is the concluding equality, i.e., the fact that if ξ, ζ, ρ are independent $\mathcal{N}(0,1)$ random variables and r, s are nonnegative integers, then

$$\mathcal{E}\left\{(\xi + i\zeta)^r(\xi + i\rho)^s\right\} = \delta_{rs}r!. \tag{8.8}$$

But $\mathcal{E}_\zeta\left\{(\xi + i\zeta)^r\right\} = H_r(\xi)$ is the r-th Hermite polynomial (see [31], p. 163), and (8.8) is precisely the orthogonality property of these polynomials:

$$\mathcal{E}\left\{H_r(\xi)H_s(\xi)\right\} = \delta_{rs}r!$$

(see [4], p. 133).

Combining (8.5), (8.6) and (8.7), we get

$$\mathcal{E}\left\{\widetilde{P}_N^2\right\} = \sum_{p=1}^{k} \varepsilon^{2p} \sum_{\iota \in \mathcal{J}_N^p} \frac{(P_\iota^p)^2}{\iota!} = \sum_{p=1}^{k} \frac{\varepsilon^{2p}}{p!} \parallel D^p P_N(f) \parallel_2^2,$$

the concluding equality being given by the fact that every P_ι^p occurs exactly $\frac{p!}{\iota!}$ times among the coefficients of the p-linear form $D^p P_N(f)[\cdot, ..., \cdot]$ with respect to the basis $\{\phi_i\}_{i=1}^N$. ■

Remark 8.1.1 A simple modification of the proof of Theorem 8.1.1 yields the following result. Let $G(x)$ be a function on $\mathbf{R}^N$ which can be continued to an entire function $G(z)$ on $\mathbf{C}^N$ such that

$$|G(z)| \leq c \exp\left\{\theta \frac{\| z \|_2^2}{2\varepsilon^2}\right\}$$

with some $\theta \in (0,1)$, $c < \infty$. Assume that an observation $y = x + \varepsilon\xi$ of a point $x \in \mathbf{R}^N$ is given, where the noise ξ is Gaussian with zero mean and identity covariance matrix. Then the estimator

$$\widehat{G}(y) \equiv \mathcal{E}_\zeta G(y + i\varepsilon\zeta),$$

ζ being independent of ξ Gaussian random vector with zero mean and identity covariance matrix, is an unbiased estimator of $G(x)$, $x \in \mathbf{R}^N$, with variance

$$\mathcal{E}\left\{\left[\widehat{G} - G(x)\right]^2\right\} = \sum_{p=1}^{\infty} \frac{\varepsilon^{2p}}{p!} \parallel D^p G(x) \parallel_2^2 \qquad \left[\parallel D^p G(x) \parallel_2^2 \equiv \sum_{\iota \in \mathcal{J}_N^p} \left| \frac{\partial^p f(x)}{\partial x_1^{\iota_1} \ldots \partial x_N^{\iota_N}} \right|^2 \frac{p!}{\iota!}\right]$$

Note that $\widehat{G}$ is the unique unbiased estimator of G in the class of estimators $\Psi(y)$ satisfying the condition

$$\forall y \in \mathbf{R}^N: \ \Psi(y) \leq c_\Psi \exp\left\{\theta_\Psi \frac{\| y \|_2^2}{2\varepsilon^2}\right\} \qquad [c_\Psi < \infty, \theta_\Psi \in (0,1)]$$

Corollary 8.1.1 *Let*

$$P(f) = \sum_{p=0}^{k} P^p(f)$$

be a polynomial on H with Hilbert-Schmidt homogeneous components $P^0, \ldots, P^k$ of the Hilbert-Schmidt norms not exceeding L. Then for every $\varepsilon > 0$ one can choose $N = N(P, \varepsilon) < \infty$ in such a way that for the associated estimator $\widetilde{P}_N$ of $P(f)$ via observations (8.1) *one has*

$$\forall f \in H: \qquad \mathcal{E}\left\{\left[\widetilde{P}_N(y^N(f,\varepsilon)) - P(f)\right]^2\right\} \leq \varepsilon^2 \parallel P'(f) \parallel^2 + c(k) L^2 \varepsilon^4 (1 + \parallel f \parallel^{2k}). \tag{8.9}$$

In particular, the resulting estimator is asymptotically efficient on every bounded subset of H.

Proof. For every positive integer N, for every $p \leq k$ and every $f \in H$ we have

$$\begin{aligned} \left| P^p(f) - P^p(f^N) \right| &\leq \sum_{\substack{\iota_1, \ldots \iota_p: \\ \max_j \iota_j > N}} \left| P^p_{\iota_1, \ldots, \iota_p} f_{\iota_1} \cdots f_{\iota_p} \right| \\ &\leq \sqrt{\sum_{\substack{\iota_1, \ldots \iota_p: \\ \max_j \iota_j > N}} \left(P^p_{\iota_1, \ldots, \iota_p}\right)^2} \parallel f \parallel^p, \end{aligned}$$

whence for every positive δ there exists $N_1(\delta)$ such that

$$|P(f) - P(f^N)| \leq \delta(1 + \parallel f \parallel^k) \qquad \forall f \forall N \geq N_1(\delta).$$

By similar reasons, for every $\delta > 0$ there exists $N_2(\delta)$ such that

$$\| P'(f) - DP_N(f) \| \leq \delta(1 + \| f \|^{k-1}) \qquad \forall f \forall N \geq N_2(\delta).$$

It is also clear that for some $c_1(k)$ (depending only on k) and for all N and all $f \in H$ we have

$$\| D^p P_N(f) \|_2 \leq L c_1(k)(1 + \| f \|^k).$$

Letting $N = \max[N_1(L\varepsilon^2), N_2(L\varepsilon^2)]$, we get (8.9) as a consequence of (8.4). ■

Remark 8.1.2 It is easily seen that if a polynomial P satisfies the premise of Corollary 8.1.1, then the estimators $\widetilde{P}_N$ (see (8.3)) converge in the mean square as $N \to \infty$ to an unbiased estimator $\widetilde{P}$ of $P(\cdot)$, the variance of the estimator being

$$\mathcal{E}\left\{\left[\widetilde{P}(y^{f,\varepsilon}) - P(f)\right]^2\right\} = \sum_{p=1}^{k} \frac{\varepsilon^{2p}}{p!} \| D^p P(f) \|_2^2 .$$

Examples. **I.** A continuous linear form $P(f) = (p, f)$ always is a Hilbert-Schmidt polynomial, and the corresponding unbiased estimator is the standard plug-in estimator

$$\widetilde{P}(y^{f,\varepsilon}) = \sum_{j=1}^{\infty} y_j^{f,\varepsilon} p_j \qquad [p_j = (p, \phi_j)]$$

II. Let $P(f) = (Af, f)$ be a homogeneous continuous quadratic form, and let $[a_{j\ell}]$ be the matrix of the form with respect to the basis $\{\phi_j\}$. The estimator $\widetilde{P}_N$ of $P_N(f) = (Af^N, f^N)$ is

$$\widetilde{P}_N = \sum_{j \neq \ell, j, \ell \leq N} a_{j\ell} y_j^{f,\varepsilon} y_\ell^{f,\varepsilon} + \sum_{j=1}^{N} a_{jj}\left(\left[y_j^{f,\varepsilon}\right]^2 - \varepsilon^2\right),$$

and the variance of this estimator is

$$\mathcal{E}\left\{\left[\widetilde{P}_N - P_N(f)\right]^2\right\} = 4\varepsilon^2 \| (Af^N)^N \|^2 + 2\varepsilon^4 \sum_{j,\ell=1}^{N} a_{j\ell}^2.$$

For N fixed, this is an asymptotically efficient estimator of $P_N(f)$. The trivial plug-in estimator $P_N(y^N(f, \varepsilon))$ also is an asymptotically efficient estimator of $P_N(f)$ (N is fixed), but its risk is greater than the one of $\widetilde{P}_N$ in terms of order of ε^4:

$$\mathcal{E}\left\{\left[P_N(y^N(f,\varepsilon)) - P_N(f)\right]^2\right\} = 4\varepsilon^2 \| (Af^N)^N \|^2 + 2\varepsilon^4 \left(\sum_{j,\ell=1}^{N} a_{j\ell}^2 + \frac{1}{2}\left|\sum_{j=1}^{N} a_{jj}\right|^2\right);$$

when N is large, this difference can be decisive.

If A is a Hilbert-Schmidt operator (i.e., $\sum_{j,\ell} a_{j\ell}^2 < \infty$), then the estimators $\widetilde{P}_N$ converge in the mean square, as $N \to \infty$, to an unbiased asymptotically efficient, on every bounded subset of H, estimator of $P(\cdot)$.

III. Let $P(f) = \sum\limits_{j=1}^{\infty}[(f,\phi_j)]^3$. Then the unbiased estimator $\widetilde{P}_N$ of $P_N(f) = \sum\limits_{j=1}^{N}[(f,\phi_j)]^3$ is

$$\widetilde{P}_N = \sum_{j=1}^{N}\left(\left[y_j^{f,\varepsilon}\right]^2 - 3\varepsilon^2 y_j^{f,\varepsilon}\right),$$

and its variance is

$$\mathcal{E}\left\{\left[P_N(y^N(f,\varepsilon)) - P_N(f)\right]^2\right\} = 9\varepsilon^2\sum_{j=1}^{N} f_j^4 + 18\varepsilon^4\sum_{j=1}^{N} f_j^2 + 6\varepsilon^6 N,\ f_j = (f,\phi_j).$$

8.1.3 Extension

We have built asymptotically efficient, on bounded subsets of H, estimators for Hilbert-Schmidt polynomials. To achieve our final goals, we need to build a "nearly" asymptotically efficient estimate for a "nearly" Hilbert-Schmidt polynomial. Namely, assume that

$$P(f) = \sum_{p=0}^{k} P^p(f)$$

is a polynomial of degree $k \geq 2$ such that

$$\begin{array}{ll} (a) & P^p \text{ are Hilbert-Schmidt polynomials for } p \leq k-1 \text{ with } \| P^p \|_2 \leq L < \infty \\ (b.1) & \| P^k \| \equiv \sup\{|\Pi_k[f_1,...,f_k]| \mid \| f_\ell \| \leq 1,\ \ell = 1,...,k\} \leq L \\ (b.2) & \| P^{k,h} \|_2 \leq L \| h \|, \end{array} \tag{8.10}$$

where $\Pi_k[f_1,...,f_k]$ is the symmetric k-linear form associated with P^k and $P^{k,h}$ is the $(k-1)$-symmetric linear form obtained from $\Pi_k[f_1,...,f_k]$ when the last argument is set to a constant value h. E.g., the quadratic form (Af,f) associated with a bounded symmetric operator A satisfies $(b.1),(b.2)$ with $L = \| A \|$. Another example of a homogeneous polynomial satisfying $(b.1),(b.2)$ is given by a "diagonal" polynomial

$$P^k(f) = \sum_{j=1}^{\infty} c_j f_j^k \quad [f_j = (f,\phi_j)]$$

with bounded sequence of coefficients $\{c_j\}$ or by a continuous "band-type" polynomial

$$P^k(f) = \sum_{\iota\in\mathcal{I}_k^d} c_\iota f_{\iota_1}...f_{\iota_k} \qquad [f_j = (f,\phi_j)],$$

where $\mathcal{I}_k^d$ is the set of multi-indices $\iota = (\iota_1,...,\iota_k)$ such that $\max\limits_{\ell=1,...,k}\iota_\ell - \min\limits_{\ell=1,...,k}\iota_\ell \leq d < \infty$.

Under condition (8.10) we can build an estimator for the polynomial $P_N(f) = P(f^N)$ as follows. Let $M < N$ be a given natural number. Consider the polynomial

$$P_{*,M}(f) = \sum_{p=1}^{k-1} P^p(f) + P^k[f,...,f,f^M] \tag{8.11}$$

Then, by virtue of (8.10.b.1),

$$\begin{array}{rcl} \left|P_{*,M}(f^N) - P_N(f^N)\right| & \leq & L \parallel f^N_{M+1} \parallel \parallel f^N \parallel^{k-1}, \\ f^N_{M+1} & = & \sum\limits_{j=M+1}^{N} (f, \phi_j)\phi_j. \end{array} \tag{8.12}$$

At the same time, the homogeneous polynomial $\bar{P}^k(f) = P^k[f, ..., f, f^M]$ corresponds to the symmetric k-linear form

$$\bar{\Pi}_k(h_1, ..., h_k) = \frac{1}{k}\left(\Pi_k[(h_1)^M, h_2, ..., h_k] + ... + \Pi_k[h_1, ..., h_{k-1}, (h_k)^M]\right),$$

and the coefficients of this form are as follows. Let us partition the set $\mathcal{I}_k$ of multi-indices $\iota = (\iota_1, ..., \iota_k)$ of the coefficients into $M + 1$ groups: the first M groups G_j contain the multi-indices ι with $\min\limits_{\ell=1,...,k} \iota_\ell = j$, $j = 1, ..., M$, and the group G_{M+1} contains the multi-indices ι with $\min\limits_{\ell=1,...,k} \iota_\ell > M$. The coefficients of $\bar{\Pi}_k$ with indices from G_{M+1} are zero, and absolute values of the coefficients with indices from G_j, $j = 1, ..., M$, are less than or equal to the absolute values of the coefficients $\Pi_{k,\iota}$ of Π_k with the same indices. By (8.10.b.2),

$$\sqrt{\sum_{\iota \in G_j} \Pi^2_{k,\iota}} \leq L,$$

whence

$$\parallel \bar{P}_k \parallel_2^2 \leq \sum_{j=1}^{M} \sum_{\iota \in G_j} \Pi^2_{k,\iota} \leq ML^2. \tag{8.13}$$

Associating with the Hilbert-Schmidt polynomial $P_{*,M}(\cdot)$ estimator (8.3), let the latter be denoted by $\widetilde{P}_{M,N}(\cdot)$, and applying Theorem 8.1.1, we get the following result:

Proposition 8.1.1 *Let P be a polynomial of degree $k \geq 2$ satisfying* **(8.10)**. *Then, for every pair of positive integers M, N $(M < N)$, the estimator $\widetilde{P}_{M,N}(\cdot)$ for every $f \in H$ and every $\varepsilon \in (0, 1)$ satisfies the relations*

$$\begin{array}{llcl} (a) & \left|\mathcal{E}\left\{\widetilde{P}_{M,N}(y^N(f, \varepsilon)) - P(f^N)\right\}\right| & \leq & L \parallel f^N_{M+1} \parallel \parallel f^N \parallel^{k-1} \\ (b) & \left(\mathcal{E}\left\{\left[\widetilde{P}_{M,N}(y^N(f, \varepsilon)) - P(f^N)\right]^2\right\}\right)^{1/2} & \leq & \varepsilon \parallel P'_{*,M}(f^N) \parallel \\ & & & +c_1(k)L\varepsilon^2(1+ \parallel f^N \parallel^k) \\ & & & +c_2(k)\varepsilon^k\sqrt{M}L(1+ \parallel f^N \parallel^k) \\ & & & +L \parallel f^N_{M+1} \parallel \parallel f^N \parallel^{k-1}. \end{array} \tag{8.14}$$

8.2 From polynomials to smooth functionals

We are about to extend the techniques for asymptotically efficient estimating Hilbert-Schmidt polynomials to estimating smooth functionals with Hilbert-Schmidt derivatives. As before, we assume that the set of signals Σ satisfies **A.1**, i.e., it is a subset of the unit ball $\mathcal{O}$ of H with Kolmogorov diameters satisfying

$$d_k(\Sigma) \leq Lk^{-\beta} \tag{8.15}$$

As about the functional F to be estimated, we assume that

A.4. F is defined in the ball

$$\mathcal{O}_\rho = \{f \in H \mid \| f \| < 1 + 2\rho\} \quad [\rho > 0]$$

and is $k \geq 3$ times continuously Fréchet differentiable in $\mathcal{O}_\rho$. Moreover,

A.4.1. The derivatives $F^{(j)}(f)$, $f \in \mathcal{O}_\rho$, of order $j \leq k-1$ have bounded Hilbert-Schmidt norms:

$$\sup\left\{\| F^{(j)}(f) \|_2 | \ f \in \mathcal{O}_\rho\right\} \leq L \qquad 1 \leq j \leq k-1; \tag{8.16}$$

A.4.2. The k-th derivative $F^{(k)}(f)$ satisfies the inequality

$$\| F^{(k),g}(f) \|_2 \leq L \| g \| \quad \forall f \in \mathcal{O}_\rho \ \forall g \in H \tag{8.17}$$

(cf. (8.10)), where

$$F^{(k),g}(f)[h_1, ..., h_{k-1}] \equiv D^k F(f)[h_1, ..., h_{k-1}, g].$$

A.4.3. $F^{(k)}(f)$ is Hölder continuous, with exponent $\gamma > 0$, in the usual norm:

$$\| F^{(k)}(f) - F^{(k)}(g) \| \leq L \| f - g \|^\gamma \quad \forall f, g \in \mathcal{O}_\rho. \tag{8.18}$$

Note that **A.2**, **A.3** are nothing but the versions of **A.4** associated with $k = 1, 2$, respectively. In these cases the sharp link between the asymptotical width β of the set Σ and the smoothness parameters of F ensuring possibility for asymptotically efficient, on Σ, estimation of F was given by

$$\gamma > \frac{1}{2\beta} + 1 - k, \tag{8.19}$$

and it would be natural to suppose that the same link works for $k > 2$ as well. It turns out, however, that the "correct tradeoff" between the width of Σ and the smoothness of F under assumption **A.4** is given by

$$\gamma > \frac{1}{2\beta} - k, \ k \geq 3. \tag{8.20}$$

E.g., (8.19) says that to ensure asymptotically efficient estimation of twice continuously differentiable functional with Lipschitz continuous second derivative ($k = 2$, $\gamma = 1$) the asymptotical width of Σ should be $> \frac{1}{4}$, while (8.20) says that in order to ensure the same possibility for three times continuously differentiable functional with Hölder continuous, with *close to* 0 exponent γ, third derivative, it suffices to have $\beta > \frac{1}{6}$. At the same time, common sense says to us that a twice continuously differentiable functional with Lipschitz continuous second order derivative is basically the same as a three times continuously differentiable functional with small Hölder continuity exponent of the third derivative; if so, where the "jump down" $\beta > \frac{1}{4} \mapsto \beta > \frac{1}{6}$ in the condition ensuring possibility of asymptotically efficient estimation comes from?

The answer is that when passing from **A.3** to **A.4**, we do not merely increase the number of derivatives of the functional, but impose a *structural* assumption on the derivatives of order $< k$ – now they should be Hilbert-Schmidt polylinear operators. This structural assumption is exactly what is responsible for the above "jump down". More specifically, imposing on the second derivative of a smooth functional the restriction to be bounded in the Hilbert-Schmidt norm results in a completely new phenomenon – *measure concentration*.

8.2.1 Measure concentration

The phenomenon of measure concentration was discovered by P. Levy; in its rough form, the phenomenon is that a function G with fixed modulus of continuity, say, Lipschitz continuous with constant 1, on a high-dimensional unit Euclidean sphere "almost everywhere is almost constant": there exists a constant $a = a(G)$ such that $\mathrm{Prob}\{x \mid |G(x) - a(G)| > \varepsilon\}$, the probability being taken with respect to the uniform distribution of x on the unit n-dimensional sphere, for every fixed $\varepsilon > 0$ goes to 0 as the dimension $n \to \infty$. In the case we are interested in – the one when G is with Hilbert-Schmidt second-order derivative – this phenomenon can be expressed as follows:

Proposition 8.2.1 *Let G be a twice continuously differentiable in the ball*

$$V_r = \{x \in \mathbf{R}^n \mid \| x \| \leq r\}$$

function, and let $\| G'(0) \| \leq T$ and $\| G''(x) \|_2 \leq T$ for all $x \in V_r$ and some $T < \infty$. For $L : V_r \to \mathbf{R}$, let $M_\rho[L]$ be the average of L taken over the uniform distribution on the sphere of radius ρ centered at the origin, $0 \leq \rho \leq r$. Then

$$M_r\left[(G(x) - G(0))^2\right] \leq (1+\theta)\frac{r^2}{n}\Big(\| G'(0) \|^2 + T^2 r(2+r)\Big) + (2+\theta+\theta^{-1})\frac{r^4T^2}{4n} \quad \forall \theta > 0. \tag{8.21}$$

Remark 8.2.1 Note that if T is fixed and n is large, then (8.21) demonstrates that G in V_r is close, in the mean square sense, to the constant $G(0)$. Thus, Proposition indeed demonstrates a kind of "measure concentration" phenomenon.

Proof. Let $g(x) = (G(x) - G(0))^2$. For $0 < \rho \leq r$, let $Q_\rho(h) = \| h \|^{2-n} - \rho^{2-n}$. For $0 < \delta < \rho$ by Green's formula (Δ is the Laplacian) we have

$$\begin{aligned}\int\limits_{\delta \leq \|h\| \leq \rho} \{g \Delta Q_\rho - Q_\rho \Delta g\}\, dh \;&=\; \int\limits_{\|h\|=R} \left\{ g\frac{\partial Q_\rho}{\partial e} - Q_\rho \frac{\partial g}{\partial e}\right\} dS(h) \\ &\quad + \int\limits_{\|h\|=\delta} \left\{ g\frac{\partial Q_\rho}{\partial e} - Q_\rho \frac{\partial g}{\partial e}\right\} dS(h),\end{aligned} \tag{8.22}$$

where $dS(h)$ is the element of area of the boundary of the strip $\{\delta \leq \| h \| \leq \rho\}$ and e is the outer unit normal to the boundary. Since $\Delta Q_\rho = 0$, the left hand side in (8.22) is equal to

$$-\int\limits_{\delta}^{\rho} s^{n-1}(s^{2-n} - \rho^{2-n})\sigma_n M_s\left[\Delta g\right] ds,$$

where σ_n is the surface area of a unit sphere in $\mathbf{R}^n$. As $\delta \to +0$, the right hand side in (8.22) tends to $(2-n)\sigma_n M_\rho[g]$ (note that $g(0) = 0$). Thus, passing to limit in (8.22) as $\delta \to +0$, we get

$$(n-2)M_\rho[g] = -\int\limits_{0}^{\rho} (s - s^{n-1}\rho^{2-n}) M_s[\Delta g] ds,$$

or, which is the same,

$$\begin{array}{rcl} M_\rho[g] & = & (2n)^{-1}\rho^2 \int\limits_0^\rho \theta_\rho(s) M_s[\Delta g] ds, \\ \theta_\rho(s) & = & \frac{2n}{\rho^2(n-2)}(s - s^{n-1}\rho^{2-n}) \\ & \geq & 0, \\ \int\limits_0^\rho \theta_\rho(s) ds & = & 1. \end{array} \tag{8.23}$$

Now let $\ell(x) = G(x) - G(0)$, so that $g(x) = \ell^2(x)$. We have

$$\frac{1}{2}\Delta g = \ell \Delta \ell + \| \nabla \ell \|^2 . \tag{8.24}$$

Let

$$\begin{array}{rcl} A(\rho) & = & \max\limits_{0 \leq s \leq \rho} M_s[g], \\ B(\rho) & = & \max\limits_{0 \leq s \leq \rho} M_s\left[(\Delta \ell)^2\right], \\ C(\rho) & = & \max\limits_{0 \leq s \leq \rho} M_s\left[\| \nabla \ell \|^2\right]. \end{array}$$

From (8.23) and (8.24) it follows that

$$\begin{array}{rcl} M_\rho[g] & = & \frac{\rho^2}{n} \int\limits_0^\rho \theta_\rho(s) M_s\left[|\ell \Delta_\ell| + \| \nabla \ell \|^2\right] ds \\ & \leq & \frac{\rho^2}{n} \int\limits_0^\rho \theta_\rho(s) \left(M_s^{1/2}[\ell^2] M_s^{1/2}[(\Delta_\ell)^2] + M_s[\| \nabla \ell \|^2] \right) ds \\ & \leq & \frac{\rho^2}{n} \left(A^{1/2}(\rho) B^{1/2}(\rho) + C(\rho) \right) \\ & & [\text{since } \theta_\rho \geq 0 \text{ and } \int\limits_0^\rho \theta_\rho(s) ds = 1] \\ & \leq & \frac{r^2}{n} \left(A^{1/2}(r) B^{1/2}(r) + C(r) \right). \end{array}$$

Since the resulting inequality is valid for all $\rho \leq r$, we get for every $\delta > 0$:

$$\begin{array}{rrcl} & A(r) & \leq & \frac{r^2}{n}\left[\frac{\delta}{2}A(r) + \frac{1}{2\delta}B(r) + C(r)\right] \\ & & & \left[\text{setting } 1 - \frac{r^2\delta}{2n} = \frac{1}{\theta+1}\right] \\ \Rightarrow & A(r) & \leq & \frac{r^2}{n}(1+\theta)C(r) + \frac{r^4}{4n^2}\left(2 + \theta + \theta^{-1}\right)B(r) \quad \forall \theta > 0. \end{array} \tag{8.25}$$

Now, by assumptions of Proposition in V_r we have $\| \nabla \ell \| = \| \nabla G \| \leq \| G'(0) \|_2 + Tr$, whence

$$C(r) = \max_{0 \leq \rho \leq r} M_\rho[\| \nabla \ell \|^2] \leq (\| G'(0) \| + Tr)^2 \leq \| G'(0) \|^2 + T^2 r(2+r),$$

and

$$B(r) = \max_{0 \leq \rho \leq r} M_\rho[(\Delta \ell)^2] = \max_{0 \leq \rho \leq r} M_\rho[(\Delta G)^2] \leq nT^2,$$

the concluding inequality being given by

$$(\Delta G)^2 = \left(\sum_{i=1}^n \frac{\partial^2 G}{\partial x_i^2} \right)^2 \leq n \sum_{i=1}^n \left(\frac{\partial^2 G}{\partial x_i^2} \right)^2 \leq nT^2.$$

In view of these bounds, (8.25) implies (8.21). ∎

8.2.2 The estimate

We are about to prove the following

Theorem 8.2.1 *Let Σ, F satisfy conditions* **A.1**, **A.4** *and let* (8.20) *take place. The F admits asymptotically efficient on Σ estimation method.*

Remark 8.2.2 The link (8.20) between β and γ is sharp (in the same sense as in Theorem 7.2.2). The proof (see [24]) follows the same line of argument as in Theorem 7.2.2, but is more involving, since now we should ensure the Hilbert-Schmidt property of the derivatives.

Proof. We just build the asymptotically efficient estimation method.

Setup. Given noise intensity $\varepsilon < 0.1$, let us set

$$m = m(\varepsilon) = \lfloor \frac{1}{\varepsilon^2 \ln(1/\varepsilon)} \rfloor, \quad M = M(\varepsilon) = \lfloor \frac{1}{\varepsilon^{2(k-1)} \ln(1/\varepsilon)} \rfloor \tag{8.26}$$

(note that $M > m$ for all small enough values of ε, which is assumed from now on). Same as in the proof of Theorem 7.3.1, without loss of generality we may assume that

$$\forall f \in \Sigma : \quad \| f - f^n \| \leq c n^{-\beta}, \quad n = m, M \tag{8.27}$$

and can find $N = N(\varepsilon) > M$ such that

$$\forall f \in \Sigma : \quad |F(f) - F(f^N)| \leq \varepsilon^4, \ \| F'(f) - F'(f^N) \| \leq \varepsilon^4; \tag{8.28}$$

here and in what follows, as usual,

$$\begin{array}{rcl} f_p^q &=& \sum\limits_{i=p}^{q} (f, \phi_i)\phi_i, \\ f^q &=& f_1^q; \qquad\qquad\qquad [f \in H] \\ y_p^q &=& y_p^q(f, \varepsilon) = \sum\limits_{i=p}^{q} y_i^{f,\varepsilon}\phi_i \\ &=& f_p^q + \varepsilon \sum\limits_{i=p}^{q} \xi_i \phi_i, \\ y^q &=& y_1^q; \\ \xi_p^q &=& \{\xi_i\}_{i=p}^q, \\ \xi^q &=& \xi_1^q. \end{array}$$

In view of (8.28), we may focus on estimating the functional $F_N(f) = F(f^N)$, $f \in \Sigma$.

The estimate is as follows. Let

$$\begin{array}{rcl} \widehat{f}_m &=& y^m = f^m + \varepsilon \sum\limits_{i=1}^{m} \xi_i \phi_i, \\ \widetilde{f}_m &=& \begin{cases} \widehat{f}_m, & \| \widehat{f}_m \| \leq 1 + \rho \\ (1+\rho) \| \widehat{f}_m \|^{-1} \widehat{f}_m, & \| \widehat{f}_m \| > 1 + \rho \end{cases}, \end{array} \tag{8.29}$$

and let

$$G(h) = G^{\widetilde{f}_m}(h) \equiv \sum_{\ell=0}^{n} \frac{1}{\ell!} F^{(\ell)}(\widetilde{f}_m)\underbrace{[h^N_{m+1}, ..., h^N_{m+1}]}_{\ell} \tag{8.30}$$

Note that the polynomial G is random – it depends, as on parameters, on the "initial fragments" f^m, ξ^m of the observed signal and the observation noise; usually we skip indicating this dependence in notation.

Since the polynomial $G(h)$ depends on h^N_{m+1} only and clearly satisfies (8.10) with L depending only on the parameters involved into **A.1**, **A.4**, we may apply to this polynomial the construction from Section 8.1.3 with already specified M, N to get an estimator

$$\widehat{F}_\varepsilon = \widehat{F}_\varepsilon^{\widetilde{f}_m}(y^N_{m+1})$$

of $G^{\widetilde{f}_m}(f^N_{m+1})$ via observations $y^N_{m+1} = y^N_{m+1}(f, \varepsilon)$ with conditional (ξ^m being fixed) bias and risk satisfying the relations (see Proposition 8.1.1 and take into account that $G(h)$, as a function of h, depends on the "tail" h^N_{m+1} of h only)

$$\begin{array}{lrcl}
(a) & \left|\mathcal{E}\left\{\widehat{F}_\varepsilon - G(f^N_{m+1})\middle|\xi^m\right\}\right| & \leq & C \parallel f^N_{M+1} \parallel\parallel f^N_{m+1} \parallel^{k-1}, \\
(b) & \left(\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - G(f^N_{m+1})\right]^2\middle|\xi^m\right\}\right)^{1/2} & \leq & \varepsilon \parallel G'_*(f^N_{m+1}) \parallel \\
 & & & +C\left(\varepsilon^2 + \varepsilon^k\sqrt{M} + \parallel f^N_{M+1} \parallel\parallel f^N_{m+1} \parallel^{k-1}\right);
\end{array} \tag{8.31}$$

where

$$G_*(h) = \sum_{\ell=0}^{k-1} \frac{1}{\ell!} F^{(\ell)}(\widetilde{f}_m)[h^N_{m+1}, ..., h^N_{m+1}] + \frac{1}{k!} F^{(k)}(\widetilde{f}_m)[h^N_{m+1}, ..., h^N_{m+1}, h^M_{m+1}]; \tag{8.32}$$

here and in what follows, all C denote positive quantities depending only on the data involved **A.1**, **A.4** and all $o(1)$ are deterministic functions of ε depending on the same "side parameters" as C's and converging to 0 as $\varepsilon \to +0$.

The above $\widehat{F}_\varepsilon$ is our estimate of $F(f)$ via observations (8.1).

Accuracy analysis. We should prove (see (8.28)) that if $f \in \Sigma$, then

$$\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - F_N(f)\right]^2\right\} \leq \varepsilon^2 \parallel F'_N(f) \parallel^2 + \varepsilon^2 o(1),\ \varepsilon \to 0 \tag{8.33}$$

Assume that ε is so small that $\parallel f^N_{m+1} \parallel \leq \rho$ for all $f \in \Sigma$ (this indeed is the case for all small enough values of ε in view of (8.27)). Since F is well-defined in $\mathcal{O}_\rho$ and $\parallel \widetilde{f}_m \parallel \leq 1 + \rho$ by construction, the functional $F(\widetilde{f}_m + f^N_{m+1})$ is well-defined for all $f \in \Sigma$ and all realizations of noises. Representing

$$\begin{array}{rcl}
\widehat{F}_\varepsilon - F_N(f) & = & \underbrace{\widehat{F}_\varepsilon - G^{\widetilde{f}_m}(f^N_{m+1})}_{A} + \underbrace{G^{\widetilde{f}_m}(f^N_{m+1}) - F(\widetilde{f}_m + f^N_{m+1})}_{B} \\
 & & + \underbrace{F(\widetilde{f}_m + f^N_{m+1}) - F(f^N)}_{D},
\end{array}$$

we claim that in order to get (8.33) it suffices to verify that

A)

$$\begin{array}{lrcl}(a) & \mathcal{E}\left\{\left[\widehat{F}_\varepsilon - G^{\widetilde{f}_m}(f^N_{m+1})\right]^2\right\} & \leq & \varepsilon^2 \sum\limits_{i=m+1}^{N} [F'(f^N)]_i^2 + \varepsilon^2 o(1) \\ & & & [\text{for } g \in H,\ g_i = (g, \phi_i)] \\ (b) & \left|\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - G^{\widetilde{f}_m}(f^N_{m+1})\right] \Big| \xi^m\right\}\right| & \leq & \varepsilon o(1)\end{array} \tag{8.34}$$

B)

$$\mathcal{E}\left\{\left[G^{\widetilde{f}_m}(f^N_{m+1}) - F(\widetilde{f}_m + f^N_{m+1})\right]^2\right\} \leq \varepsilon^2 o(1) \tag{8.35}$$

C)

$$\mathcal{E}\left\{\left[F(\widetilde{f}_m + f^N_{m+1}) - F(f^N)\right]^2\right\} \leq \varepsilon^2 \sum_{i=1}^{m} [F'(f^N)]_i^2 + \varepsilon^2 o(1). \tag{8.36}$$

Indeed, assuming that (8.34) – (8.36) take place, we have

$$\begin{array}{rl} & \mathcal{E}\left\{\left[\widehat{F}_\varepsilon - F_N(f)\right]^2\right\} \\ = & \mathcal{E}\left\{[A+B+D]^2\right\} \\ \leq & (1+\theta)\mathcal{E}\left\{[A+D]^2\right\} + (1+\theta^{-1})\mathcal{E}\left\{B^2\right\} \quad [\forall \theta > 0] \\ \leq & (1+\theta)\mathcal{E}\left\{[A+D]^2\right\} + (1+\theta^{-1})\varepsilon^2 o(1) \\ & \hfill [\text{by } (8.35)] \\ \leq & (1+\theta)\left[\mathcal{E}\left\{A^2+D^2\right\} + 2\mathcal{E}_{\xi^m}\left\{D\mathcal{E}\left\{A\Big|\xi^m\right\}\right\}\right] + (1+\theta^{-1})\varepsilon^2 o(1) \\ & \hfill [\text{since } D \text{ depends on } \xi^m \text{ only}] \\ \leq & (1+\theta)\left[\mathcal{E}\left\{A^2+D^2\right\} + \varepsilon o(1)\mathcal{E}\left\{|D|\right\}\right] + (1+\theta^{-1})\varepsilon^2 o(1) \\ & \hfill [\text{by } (8.34.b)] \\ \leq & (1+\theta)\left[\varepsilon^2\left(\| F_N'(f) \|^2 + o(1)\right) + \varepsilon o(1)\mathcal{E}\left\{|D|\right\}\right] + (1+\theta^{-1})\varepsilon^2 o(1) \\ & \hfill [\text{by } (8.34.a),\ (8.36)] \\ \leq & (1+\theta)\varepsilon^2\left(\| F_N'(f) \|^2 + o(1)\right) + (1+\theta^{-1})\varepsilon^2 o(1) \\ & \hfill [\text{by } (8.36)] \\ \leq & \varepsilon^2 \| F_N'(f) \|^2 + \varepsilon^2 o(1) \\ & \hfill [\text{choose appropriate } \theta = \theta(\varepsilon) \to 0, \varepsilon \to +0] \end{array}$$

as required in (8.33).

It remains to verify A) – C)

Verifying A) We have

$$\begin{array}{rll} & \left|\mathcal{E}\left\{\left[\widehat{F}_\varepsilon - G^{\widetilde{f}_m}(f^N_{m+1})\right] \Big| \xi^m\right\}\right| & \\ \leq & C \| f^N_{M+1} \| \| f^N_{m+1} \|^{k-1} & [\text{by } (8.31.a)] \\ \leq & C M^{-\beta} m^{-(k-1)\beta} & [\text{by } (8.27)] \\ \leq & C\varepsilon^{4(k-1)\beta} \left(\ln(1/\varepsilon)\right)^C & [\text{by } (8.26)] \\ \leq & \varepsilon o(1) & [\text{since} \\ & \multicolumn{2}{l}{4\beta(k-1) > 2\frac{k-1}{k+\gamma} \geq 2\frac{k-1}{k+1} \geq 1 \text{ by } (8.20) \text{ and due to } \gamma \leq 1,\ k \geq 3]} \end{array} \tag{8.37}$$

as required in (8.34.b). To prove (8.34.a), observe first that

$$\begin{array}{rl} & \mathcal{E}\left\{\left[\widehat{F}_\varepsilon - G^{\widetilde{f}_m}(f^N_{m+1})\right]^2\right\} \\ \leq & (1+\theta)\mathcal{E}\left\{\varepsilon^2 \parallel G'_*(f^N_{m+1}) \parallel^2\right\} \\ & +(1+\theta^{-1})C\left(\varepsilon^4+\varepsilon^{2k}M+\parallel f^N_{M+1}\parallel^2\parallel f^N_{m+1}\parallel^{2k-2}\right) \quad [\forall\theta>0] \\ & [\text{see } (8.31.b)] \\ \leq & (1+\theta)\mathcal{E}\left\{\varepsilon^2 \parallel G'_*(f^N_{m+1}) \parallel^2\right\} + (1+\theta^{-1})\varepsilon^2 o(1) \\ & [\text{since } \varepsilon^{2k}M \leq C\varepsilon^2/\ln(1/\varepsilon) \text{ by } (8.26) \\ & \text{and } \parallel f^N_{M+1}\parallel^2\parallel f^N_{m+1}\parallel^{2k-2}\leq \varepsilon^2 o(1) \text{ as in } (8.37)] \end{array} \tag{8.38}$$

To complete the proof of (8.34.a), it suffices to show that

$$\mathcal{E}\left\{\parallel G'_*(f^N_{m+1}) \parallel^2\right\} \leq \sum_{i=m+1}^{N}[F'_N(f)]^2_i + o(1); \tag{8.39}$$

given (8.39), we can choose in the resulting estimate of (8.38) $\theta = \theta(\varepsilon)$ so slowly converging to 0 as $\varepsilon \to +0$ that the estimate will imply (8.34.a).

To verify (8.39), note that by (8.32) and in view of **A.4**

$$\parallel G'_*(f^N_{m+1}) - \left([F'(\widetilde{f}_m)]^N_{m+1}\right) \parallel \leq C \parallel f^N_{m+1} \parallel \leq o(1)$$

(the concluding inequality is given by (8.27)), whence

$$\begin{array}{rl} & \left(\mathcal{E}\left\{\parallel G'_*(f^N_{m+1}) \parallel^2\right\}\right)^{1/2} \\ \leq & \sqrt{\sum\limits_{i=m+1}^{N}[F'(f^N)]^2_i} + \left(\mathcal{E}\left\{\parallel F'(\widetilde{f}_m) - F'(f^N) \parallel^2\right\}\right)^{1/2} + o(1) \\ \leq & \sqrt{\sum\limits_{i=m+1}^{N}[F'(f^N)]^2_i} + C\left(\mathcal{E}\left\{\parallel f^N - \widetilde{f}_m \parallel^2\right\}\right)^{1/2} + o(1) \\ & \qquad [\text{since } F' \text{ is Lipschitz continuous on } \mathcal{O}_\rho \text{ by } \mathbf{A.4}] \\ \leq & \sqrt{\sum\limits_{i=m+1}^{N}[F'(f^N)]^2_i} + C\left(\mathcal{E}\left\{\parallel f^N - \widehat{f}_m \parallel^2\right\}\right)^{1/2} + o(1) \\ \leq & \sqrt{\sum\limits_{i=m+1}^{N}[F'(f^N)]^2_i} + C\left(m\varepsilon^2 + \parallel f^N_{m+1} \parallel^2\right)^{1/2} + o(1) \\ = & \sqrt{\sum\limits_{i=m+1}^{N}[F'(f^N)]^2_i} + o(1) \\ & \qquad\qquad [\text{see } (8.26), (8.27)]\ ; \end{array}$$

since F' is bounded in $\mathcal{O}_\rho$, (8.39) follows. A) is proved.

Verifying B) As it was already mentioned, for all small enough values of ε the segment $[\widetilde{f}_m, \widetilde{f}_m + f^N_{m+1}]$ is, for all $f \in \Sigma$ and all realizations of noises, contained in $\mathcal{O}_\rho$. Due to the origin of $G(h) = G^{\widetilde{f}_m}(h)$ and in view of **A.4.3** we have

$$\begin{array}{rll} |F(\widetilde{f}_m + f^N_{m+1}) - G(f^N_{m+1})| & \leq & C \parallel f^N_{m+1} \parallel^{k+\gamma} \\ & \leq & Cm^{-\beta(k+\gamma)} \qquad [\text{by } (8.27)] \\ & \leq & \varepsilon o(1) \qquad [\text{by } (8.26) \text{ and } (8.20)] \end{array}$$ □

Verifying C) We have

$$\begin{array}{rcl}\mathcal{E}\left\{\left[F(\widetilde{f}_m + f^N_{m+1}) - F(f^N)\right]^2\right\} & = & \mathcal{E}\left\{\left[F(\widehat{f}_m + f^N_{m+1}) - F(f^N)\right]^2 \chi_{\varepsilon\|\xi^m\|\le\rho}\right\} \\ & & +\mathcal{E}\left\{\left[F(\widetilde{f}_m + f^N_{m+1}) - F(f^N)\right]^2 \chi_{\varepsilon\|\xi^m\|>\rho}\right\},\end{array}$$

so that to verify C) it suffices to check that

$$\begin{array}{lrcl}(a) & \mathcal{E}\left\{\left[F(\widetilde{f}_m + f^N_{m+1}) - F(f^N)\right]^2 \chi_{\varepsilon\|\xi^m\|>\rho}\right\} & \le & \varepsilon^2 o(1), \\ (b) & \mathcal{E}\left\{\left[F(\widehat{f}_m + f^N_{m+1}) - F(f^N)\right]^2 \chi_{\varepsilon\|\xi^m\|\le\rho}\right\} & \le & \varepsilon^2 \sum\limits_{i=1}^{m}[F'(f^N)]_i^2 + \varepsilon^2 o(1).\end{array} \tag{8.40}$$

Verifying (8.40.a): Since F is bounded in $\mathcal{O}_\rho$, it suffices to prove that

$$\mathrm{Prob}\left\{\varepsilon \parallel \xi^m \parallel > \rho\right\} \le \varepsilon^2 o(1),$$

which is immediately given by Bernstein's inequality; for the sake of completeness, here is the proof:

$$\begin{array}{ll} & \mathrm{Prob}\left\{\varepsilon \parallel \xi^m \parallel > \rho\right\} \\ = & \mathrm{Prob}\left\{\sum\limits_{i=1}^{m} \xi_i^2 > \rho^2\varepsilon^{-2}\right\} \\ = & \mathrm{Prob}\left\{\frac{1}{4}\sum\limits_{i=1}^{m} \xi_i^2 > \frac{1}{4}\rho^2\varepsilon^{-2}\right\} \\ \le & \mathcal{E}\left\{\exp\left\{\sum\limits_{i=1}^{m} \frac{\xi_i^2}{4}\right\}\right\}\exp\left\{-\frac{1}{4}\rho^2\varepsilon^{-2}\right\} \quad \text{[by Tschebyshev's inequality]} \\ = & \left[\mathcal{E}\left\{\exp\left\{\frac{\xi_1^2}{4}\right\}\right\}\right]^m \exp\left\{-\frac{1}{4}\rho^2\varepsilon^{-2}\right\} \\ \le & \exp\left\{Cm - \frac{1}{4}\rho\varepsilon^{-2}\right\} \\ \le & \exp\left\{-\frac{1}{8}\rho^2\varepsilon^{-2}\right\} \quad \forall\varepsilon \le \varepsilon_0 \qquad \text{[by (8.26)]} \\ \le & \varepsilon^2 o(1). \end{array}$$

□

Verifying (8.40.b): this is the central point, and this is the point where the "measure concentration" is exploited. Let

$$g(h) = F(f^N + h) - F(f^N), \quad h \in H_m,$$

where H_m is the linear span of $\phi_1, ..., \phi_m$. By **A.4.1**, this function satisfies the premise of Proposition 8.2.1 with $r = \rho$ and with $T = C$. Denoting by

$$\psi_m(s) = a_m \exp\{-s^2/2\}s^{m-1}$$

the density of the Euclidean norm of m-dimensional random Gaussian vector ξ^m, we have

$$
\begin{array}{rl}
& \mathcal{E}\left\{\left[F(\widehat{f}_m + f^N_{m+1}) - F(f^N)\right]^2 \chi_{\varepsilon\|\xi^m\|\leq\rho}\right\} \\
= & \mathcal{E}\left\{g^2(\varepsilon\xi^m)\chi_{\|\xi^m\|\leq\rho/\varepsilon}\right\} \\
= & \int\limits_0^{\rho/\varepsilon} M_{\varepsilon s}[g^2]\psi_m(s)ds \qquad \text{[averages } M_t[\cdot] \text{ are defined in Proposition 8.2.1]} \\
\leq & \int\limits_0^{\infty}\Big[(1+\theta)\dfrac{\varepsilon^2 s^2}{m} \parallel g'(0) \parallel^2 + (1+\theta)C\dfrac{\varepsilon^3 s^3(1+\varepsilon s)}{m} \\
& +(2+\theta+\theta^{-1})C\frac{s^4\varepsilon^4}{m}\Big]\psi_m(s)ds \quad \forall\theta>0 \\
& \text{[by (8.21)]} \\
\leq & (1+\theta) \parallel g'(0) \parallel^2 \frac{\varepsilon^2}{m}\mathcal{E}\left\{\parallel \xi^m \parallel^2\right\} + (1+\theta)C\mathcal{E}\left\{\frac{\varepsilon^3}{m} \parallel \xi^m \parallel^3 + \frac{\varepsilon^4}{m} \parallel \xi^m \parallel^4\right\} \qquad \blacksquare \\
& +(2+\theta+\theta^{-1})C\mathcal{E}\left\{\frac{\varepsilon^4}{m} \parallel \xi^m \parallel^4\right\} \\
\leq & (1+\theta)\varepsilon^2 \parallel g'(0) \parallel^2 + C(2+\theta+\theta^{-1})\varepsilon^2\left[\varepsilon m^{1/2} + \varepsilon^2 m\right] \\
& \text{[since } \mathcal{E}\left\{\parallel \xi^m \parallel^2\right\} = m,\ \mathcal{E}\left\{\parallel \xi^m \parallel^p\right\} \leq c_p m^{p/2}\text{]} \\
\leq & (1+\theta)\varepsilon^2 \parallel g'(0) \parallel^2 + (2+\theta+\theta^{-1})\varepsilon^2 o(1) \\
& \text{[since } \varepsilon m^{1/2} = o(1) \text{ by (8.26)]} \\
= & (1+\theta)\varepsilon^2 \sum\limits_{i=1}^m [F'(f^N)]_i^2 + (2+\theta+\theta^{-1})\varepsilon^2 o(1) \\
& \text{[the origin of } g\text{]} \\
\leq & \varepsilon^2 \sum\limits_{i=1}^m [F'(f^N)]_i^2 + \varepsilon^2 o(1) \\
& \text{[choose appropriate } \theta=\theta(\varepsilon)\to 0,\ \varepsilon\to+0\text{]}
\end{array}
$$

Bibliography

[1] Barron A. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. on Information Theory*, v. 39 No. 3 (1993).

[2] Besov O.V., V.P. Il'in, and S.M. Nikol'ski. *Integral representations of functions and embedding theorems*, Moscow: Nauka Publishers, 1975 (in Russian)

[3] Birgé L. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, **v.** 65 (1983), 181-237.

[4] Cramer H. *Mathematical Methods of Statistics*, Princeton Univ. Press, Princeton, 1957.

[5] Donoho D., I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* **v.** 81 (1994) No.3, 425-455.

[6] Donoho D., I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **v.** 90 (1995) No. 432, 1200-1224.

[7] Donoho D., I. Johnstone, G. Kerkyacharian, D. Picard. *Wavelet shrinkage: Asymptopia?* (with discussion and reply by the authors). *J. Royal Statist. Soc. Series B* **v.** 57 (1995) No.2, 301-369.

[8] Eubank R. *Spline smoothing and Nonparametric Regression*, Dekker, New York, 1988.

[9] Goldenshluger A., A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods of Statistics*, **v.** 6 (1997) No. 2, 135 – 170.

[10] Goldenshluger A., A. Nemirovski. Adaptive de-noising of signals satisfying differential inequalities. *IEEE Transactions on Information Theory* **v.** 43 (1997).

[11] Golubev Yu. Asymptotic minimax estimation of regression function in additive model. *Problemy peredachi informatsii*, **v.** 28 (1992) No. 2, 3-15. (English transl. in *Problems Inform. Transmission* **v.** 28, 1992.)

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

[13] Ibragimov I.A., R.Z. Khasminski. *Statistical Estimation: Asymptotic Theory*, Springer, 1981.

[14] Ibragimov I., A. Nemirovski, R. Khas'minski. Some problems of nonparametric estimation in Gaussian white noise. *Theory Probab. Appl.* **v.** 31 (1986) No. 3, 391-406.

[15] Juditsky, A. Wavelet estimators: Adapting to unknown smoothness. *Math. Methods of Statistics* **v.** 6 (1997) No. 1, 1-25.

[16] Juditsky A., A. Nemirovski. *Functional aggregation for nonparametric estimation.* Technical report # 993 (March 1996), IRISA, Rennes

[17] Korostelev A., A. Tsybakov. *Minimax theory of image reconstruction. Lecture Notes in Statistics* v. 82, Springer, New York, 1993.

[18] Lepskii O. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications*, **v.** 35 (1990) No. 3, 454-466.

[19] Lepskii O. Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory of Probability and Its Applications*, **v.** 36 (1991) No. 4, 682-697.

[20] Lepskii O., E. Mammen, V. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **v.** 25 (1997) No.3, 929-947.

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

[22] Nemirovski A. On forecast under uncertainty. *Problemy peredachi informatsii*, **v.** 17 (1981) No. 4, 73-83. (English transl. in *Problems Inform. Transmission* **v.** 17, 1981.)

[23] Nemirovski A. On nonparametric estimation of smooth regression functions. *Sov. J. Comput. Syst. Sci.*, **v.** 23 (1985) No. 6, 1-11.

[24] Nemirovski A. On necessary conditions for efficient estimation of functionals of a nonparametric signal in white noise. *Theory Probab. Appl.* **v.** 35 (1990) No. 1, 94-103.

[25] Nemirovski A. On nonparametric estimation of functions satisfying differential inequalities. – In: R. Khasminski, Ed. *Advances in Soviet Mathematics*, v. 12, American Mathematical Society, 1992, 7-43.

[26] Pinsker M., Optimal filtration of square-integrable signals in Gaussian noise. *Problemy peredachi informatsii*, **v.** 16 (1980) No. 2, 120-133. (English transl. in *Problems Inform. Transmission* **v.** 16, 1980.)

[27] Pinsker M., S. Efroimovich. Learning algorithm for nonparametric filtering. *Automation and Remote Control*, **v.** 45 (1984) No. 11, 1434-1440.

[28] Pisier G. Remarques sur un resultat non publie de B. Maurey, - in: *Seminaire d'analyse fonctionelle 1980-1981*, v. 1 – v. 12, Ecole Polytechnique, Palaiseau, 1981.

[29] Prakasa Rao B.L.S. *Nonparametric functional estimation.* Academic Press, Orlando, 1983.

[30] Rosenblatt M. *Stochastic curve estimation.* Institute of Mathematical Statistics, Hayward, California, 1991.

[31] Suetin P.K. *The classical orthogonal polynomials.* Nauka, Moscow, 1976 (in Russian).

[32] Wahba G. *Spline models for observational data.* SIAM, Philadelphia, 1990.