The Annals of Statistics
2000, Vol. 28, No. 3, 681-712

FUNCTIONAL AGGREGATION FOR NONPARAMETRIC REGRESSION

BY ANATOLI JUDITSKY AND ARKADII NEMIROVSKI

Domaine Universitaire and Technion, Israel Institute of Technology

We consider the problem of estimating an unknown function f from N noisy observations on a random grid. In this paper we address the following aggregation problem: given M functions $f_1, \dots, f_M$ find an "aggregated" estimator which approximates f nearly as well as the best convex combination f^* of $f_1, \dots, f_M$. We propose algorithms which provide approximations of f^* with expected L_2 accuracy $O(N^{-1/4} \ln^{1/4} M)$. We show that this approximation rate cannot be significantly improved.

We discuss two specific applications: nonparametric prediction for a dynamic system with output nonlinearity and reconstruction in the Jones–Barron class.

1. Introduction. Consider the following *nonparametric estimation problem*: we are interested in recovering the true regression function $f(x)$ (which is a bounded Borel real-valued function of d real variables), given N observations

$$(x_t, y_t = f(x_t) + e_t), \qquad t = 1, 2, \dots, N, \tag{1}$$

of f; here, x_t are independent random vectors with common probability distribution μ, e_t, independent of each other and of x_t, are real errors such that

$$E\{e_t\} = 0, \qquad E\{e_t^2\} \le \sigma^2 < \infty. \tag{2}$$

Throughout the paper, the quality of an estimator $\hat{f}(\cdot)$ of $f(\cdot)$ is measured by its squared L_2-risk

$$\int (f(x) - \hat{f}(x))^2 \mu(dx)$$

associated with the measure μ. Note that μ is not assumed to be known in advance.

A general approach to this problem can be summarized as follows: one first chooses an approximation method, that is, represent the unknown function as a member of a parametric family; then the parameters of this approximation are estimated in order to obtain $\hat{f}$. The approximation in question is often obtained using the decomposition f in a functional basis; that is, f is represented as a weighted combination of given functions $\{f_1, \dots\}$.

When the functional basis is orthonormal (or "close to orthonormal"), the processing of the estimator is reduced to efficient estimation of corresponding

Received June 1996; revised March 2000.

AMS 1991 *subject classifications*. 62G08, 62L20.

Key words and phrases. Functional aggregation, convex optimization, stochastic approximation.

sequences of coefficients. This approach has been successfully implemented for the trigonometric basis and for wavelet systems (for the actual "state of the art" see [7] and [6], respectively).

Clearly, the quality of such an estimator depends on how well the basis "fits" f. However, we normally do not know in applications which basis better fits the function to be restored. An attractive way to resolve the arising uncertainty is to use part of the observations to build a number of estimators corresponding to several "natural" bases and to use the remaining observations to "aggregate" these estimators, that is, to find nearly the best of their convex combinations.

On the other hand, we can consider the collection of all elements of possible candidate bases as an "overcomplete" system of functions $\{f_1, \dots\}$ and then search for the "best" weighted combination $\sum \lambda^i f_i(x)$ using the data from (1). It can easily be seen that such a problem cannot be successfully solved if we do not impose some restrictions on the set of coefficients λ. The following problem arises if λ's have bounded L_1-norm.

FUNCTIONAL AGGREGATION PROBLEM. Let $\Lambda \in \mathbf{R}^M$ be a convex compact set contained in the standard $\|\cdot\|_1$-ball,

$$\max\{\|\lambda\|_1 \mid \lambda \in \Lambda\} \le 1;$$

let $f_1, \dots, f_M$ be a system of functions, and let f_Λ be the best estimator of f among the estimators which are combinations of $f_1, \dots, f_M$ with coefficients from Λ,

$$f_\Lambda = \sum_{i=1}^{M} \lambda_i^* f_i,$$

with

$$\lambda^* \in \arg\min_{\lambda \in \Lambda} \psi(\lambda) \quad \text{and} \quad \psi(\lambda) \equiv \int \left(f(x) - \sum_{i=1}^{M} \lambda_i f_i(x) \right)^2 \mu(dx).$$

Given Λ, $f_1, \dots, f_M$, a constant $L < \infty$ such that $|f|$, $|f_i| \le L$, and N observations (1), the problem is to find an estimator which is nearly as good as f_Λ.

From now on we refer to the collection $\{\mu, f, f_1, \dots, f_M, \Lambda, L\}$ as the *data* of the aggregation problem.

The study of the functional aggregation problem was initiated by Jones (cf. [13]). In that paper, the properties of the relaxed greedy algorithm approximations have been studied in the case when f is a convex combination of M functions bounded by L and observed at N sites $x_1, \dots, x_N$ without noise. As refined in [2], Jones's result states that in that case the relaxed greedy approximation f_n attains the averaged squared error

$$\frac{1}{N} \sum_{t=1}^{N} \big(f_n(x_t) - f(x_t) \big)^2 \le \frac{L^2}{n}$$

in computational time $O(nMN)$, where the parameter n is a positive integer (this is closely related to the problem of approximating functions from the Jones–Barron class below).

Those results have been developed by Lee, Barlett and Williamson [14] who studied the application of the relaxed greedy algorithm to the functional aggregation problem in the noisy environment. In that paper an aggregation estimator $\hat{f}_N$ was constructed in the case when the response variables y_t are bounded. In particular, it has been shown that the risk of the aggregation estimator satisfies

$$\int(\hat{f}_N(x) - f(x))^2 \mu(dy, dx) - \psi(\lambda^*) = O\left(\sqrt{\frac{\ln M}{N}}\right)$$

(cf. Theorem 2 of [14]) and the computation time of the algorithm is $O\big(N^{3/2}M/(\ln M)^{1/2}\big)$.

The main result of the paper is the following.

THEOREM 1.2. *Let $M > 2$. Assume that we are given in advance an upper bound L on the uniform norm of f, and that all functions $f_1, \dots, f_M$ take their values in $[-L, L]$. Given N observations* (1) *it is possible to find $\lambda_N \in \Lambda$ such that*

$$E\{\psi(\lambda_N)\} - \psi(\lambda^*) \le 8\sqrt{2e \ln M}L(2L + \sigma)N^{-1/2}, \tag{3}$$

E being the expectation with respect to the distribution of observations (1).

If, in addition, Λ is the $\|\cdot\|_1$-ball,

$$\Lambda = \big\{\lambda \in \mathbf{R}^M \mid \|\lambda\|_1 \le 1\big\}, \tag{4}$$

or the simplex

$$\Lambda = \big\{\lambda \in \mathbf{R}^M \mid \|\lambda\|_1 \le 1, \lambda \ge 0\big\}, \tag{5}$$

or the simplex

$$\Lambda = \big\{\lambda \in \mathbf{R}^M \mid \|\lambda\|_1 = 1, \lambda \ge 0\big\}, \tag{6}$$

the above λ_N may be chosen to have no more than $N + 1$ nonzero entries.

The main feature of this result is that the "expected nonoptimality"

$$\nu(N) = E\big\{\psi(\lambda_N)\big\} - \psi(\lambda^*)$$

of our resulting estimator f_{λ_N} (when compared to the best possible combination f_Λ of $f_1, \dots, f_M$ with coefficients from Λ) is basically independent of the number M of functions we are aggregating. Indeed, the right-hand side of (3) is proportional to $\sqrt{\ln M}$. This property is crucial for applications indicated in Examples 1 and 2 below, when we typically aggregate a huge number of functions. From the viewpoint of applications, the bound (3) says that our "aggregation abilities" are only limited by the computational effort to process M functions $f_1, \dots, f_M$ (notice that in Example 2 these functions are estimators themselves). On the other hand, it also means that the amount of data necessary to obtain a reasonable aggregation performance is practically independent of M.

When compared to the result of [14], we establish a direct relation of the aggregation problem with classical optimization techniques (stochastic counterpart and stochastic approximation methods). There is an improvement in the generality of distributions permitted for the response variable $y_t = f(x_t)+e_t$ (an arbitrary error distribution with finite variance is allowed). The most important achievement is that the stochastic approximation algorithm reduces the computation time to $O(NM)$.

Now let us present two specific applications of functional aggregation which deal with "dimensionality reduction" in nonparametric regression estimation. The majority of the known estimators of multivariate regression functions (see [11, 20] and references therein) are aimed to restore smooth signals (f belongs to a Sobolev ball with known or unknown smoothness parameters). It is well known that in this case the rates of convergence degrade rather fast when the dimensionality d of f increases and become exceedingly slow when d approaches $\ln N$. For example, the rate is $O(N^{-1/(2+d)})$ for Lipschitz continuous functions f.

There are basically two ways to overcome the indicated difficulty (known as the "curse of dimensionality"): either to accept that a huge amount of data is necessary or to strengthen restrictions on the function class in order to bound its "effective dimension." There are different ways to achieve this latter goal; what we are about to do is to demonstrate that some of these ways lead naturally to the aggregation problem.

EXAMPLE 1. Restoring functions from the Jones–Barron class. One way to bound the function class in its "effective dimension" has been considered by L. Jones [13] and A. Barron [2]. We can reformulate the main result of [13] and [2] for our purposes as follows: let f be the Fourier transform

$$f(x) = \int_{\mathbf{R}^d} \hat{f}(\omega)\exp\{i\omega^T x\}\,d\omega \tag{7}$$

of a function from the L_1-ball of radius $L < \infty$,

$$\int_{\mathbf{R}^d} |\hat{f}(\omega)|\,d\omega \le L < \infty. \tag{8}$$

Then for any positive integer n and for any probability distribution μ on $\mathbf{R}^d$ there exists an n-tuple $(\omega_1, \dots, \omega_n)$ and coefficients $\lambda_1, \dots, \lambda_n$, $\sum_{k=1}^n |\lambda_k| \le L$, such that the combination

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

satisfies

$$\int |f(x) - f_n(x)|^2 \mu(dx) \le L^2/n. \tag{9}$$

The problem of recovering a function f of the Jones–Barron class from noisy observations (1) has been studied, for instance, in [3] and [4]. Consider the estimator $\hat{f}_N$ of f proposed in [4]. Its construction can be summarized as follows: let Ω be a "fine" grid in the space of frequencies; then the estimator $\hat{f}_N$ is obtained via the exhaustive search over the frequency space Ω for an m-tuple (ω_k), $k = 1, \dots, m$ which minimizes

$$\sum_{i=1}^{N} \left(y_i - \sum_{k=1}^{m} \lambda_k f_k(x_i) \right)^2,$$

where $f_k(x) = L \exp\{i\omega_k^T x\}$, $\lambda_k = \pm 1/m$ and $m = O(\sqrt{N})$. It is shown that the quadratic risk of the estimator $\hat{f}_N$ satisfies

$$\int (\hat{f}_N(x) - f(x))^2 \mu(dy, dx) = O\left(\sqrt{\frac{\ln N}{N}} \right)$$

(cf. Theorem 3 of [3]). Although the theorem states that the quality of the estimator $\hat{f}_N$ is fair, it can be easily verified that the total number of elementary operations required to compute the estimator is $O(N^{\sqrt{N}})$, which is of prohibitive value even for relatively small N.

On the other hand, in order to use the indicated existence theorem we can act as follows: consider the functional system $f_k(x) = L \exp\{i\omega_k^T x\}$, $\omega_k \in \Omega$, and use observations (1) to solve the associated functional aggregation problem with Λ being the $\|\cdot\|_1$-ball $\Lambda = \{\{\lambda_\omega\}_{\omega\in\Omega} \mid \sum_{\omega\in\Omega} |\lambda_\omega| \le 1\}$. Surprisingly enough, this approach, under minor additional assumptions on f, allows recovering f with basically the same quality as that stated for Barron's estimator $\hat{f}_N$.

EXAMPLE 2. Recovering a "structured" regression function Another interesting example where "dimensionality reduction" can be achieved is the case when the function f to be recovered possesses some specific structure enabling expressing f in terms of a reasonably smooth function g of smaller dimension (i.e., depending on less than d variables),

$$f = F(g, \pi), \tag{10}$$

where the mapping F is known in advance, and π is a finite-dimensional vector of parameters. For instance, it often makes sense to assume that

$$f(x) = g(\pi^T x), \tag{11}$$

where π is a $d \times d'$-matrix with $d' < d$. This is the crux of the projection pursuit algorithm developed in [8] (a very good review of these results can be found in [12]), which considers estimators of f in the form

$$\hat{f}_N(x) = \hat{g}(\pi^T x),$$

where π is a unit vector and $\pi^T x$ may be thought of as a projection of x. The term $\hat{g}(\cdot)$ is constant along $\pi^T x = c$ and so is often called a *ridge function*: the estimator at a given point x, $\pi^T x = t$, can be thought of as based on the average over a certain (in general, adaptively chosen) strip $\{x\colon |\pi^T x - t| \le \delta\}$. Other examples are recursive partitioning [15], [5], and related methods (see, e.g., [9] and the discussion therein). These methods are derived from some mixture of statistical and heuristic arguments and sometimes give impressive results in simulations. Their drawback lies in the almost total absence of any theoretical results on their convergence rates. We refer the reader to the above references for additional information.

Now note that whenever we could expect f to be representable, in a simple fashion like (11), via a function g of smaller dimension, it would be very attractive to reduce the nonparametric estimation of f to the similar operation for g. The difficulty in carrying out this approach is that reduction (10) of f to g involves the unknown parameter π, and without this knowledge we are unable to reduce estimation of f to that of g. The parameters in question typically are "unrecoverable," for example, in the case of reduction (11) the problem of consistent estimation of π is ill-posed, because π is not uniquely defined by f (cf. the case when $f \equiv 0$). Here the only way to utilize our structural information concerning f seems to be the following: split all observations into two groups; generate "a fine grid" Π in the space of parameters and use the first portion of observations to build estimators $\hat{g}_p(\cdot)$, $p \in \Pi$. Here $\bar{f}_p$ is the estimator of f which corresponds to the hypothesis that "the parameter vector π in (10) is p." Then we use remaining observations to find an estimator $\hat{f}$ which is nearly as good as the best of the estimators $\hat{g}_p$, $p \in \Pi$. Note that this latter problem is covered by the aggregation (where we look for the best convex combination of the estimators $\hat{g}_p$ rather than for the best of these estimators).

The rest of the paper is organized as follows. The main result of the paper, Theorem 1.1 is established in Section 2. In Section 3 we demonstrate that the "aggregation performance" stated in Theorem 1.1 is the best possible in the worst-case setting.

The proof of Theorem 1.1 given in Section 2 is "constructive"; the resulting estimator f_{λ_N} is given by an explicit algorithm, which we refer to as the *stochastic counterpart*. Then in Section 4 we present another aggregation algorithm of *stochastic approximation* type. This algorithm yields estimators of the same quality as in Theorem 1.1, but is much more efficient computationally than the stochastic counterpart routine. Section 5 is devoted to a particular application: restoring functions from the Jones-Barron class. The concluding Section 6 presents numerical results for the stochastic approximation algorithm as applied to identification of nonlinear dynamic systems. In what follows we use the notation $\|f\|_p = (\int |f(x)|^p \, dx)^{1/p}$ for the L_p-norm, with $\|f\|_\infty = \max_x |f(x)|$. If λ is a vector in $\mathbf{R}^M$ then $\|\lambda\|_p = (\sum_{i=1}^M |\lambda_i|^p)^{1/p}$, $\|\lambda\|_\infty = \max_i |\lambda_i|$. We denote

$$\|f\|_{2,\mu} = \left(\int |f(x)|^2 \mu(dx)\right)^{1/2}$$

the L_2-norm associated with the measure μ. For a real a we denote by $\lfloor a \rfloor$ the largest integer which is less than or equal to a and by $\rfloor a \lfloor$ the smallest integer which is greater than or equal to a.

In the proofs we use the generic notation κ_i for positive absolute constants with unimportant values.

2. Functional aggregation: the stochastic counterpart approach. Recall that we are in the following situation: we are given M functions $f_1, \dots, f_M$ and the constant $L > 0$ such that the true function f and all f_i take values from $[-L, L]$; our goal is to approximate, given N observations (1), the optimal solution to the optimization problem

$$\text{(12)} \qquad \min\{\psi(\lambda) \mid \lambda \in \Lambda\}, \qquad \psi(\lambda) = \int \left(f(x) - \sum_{i=1}^M \lambda_i f_i(x) \right)^2 \mu(dx)$$

associated with a given convex compact set $\Lambda \subset \mathbf{R}^M$ such that $\|\lambda\|_1 \le 1$, $\lambda \in \Lambda$.

We are about to solve (12) via the *stochastic counterpart* approach [19]. To apply this general approach, first note that the objective $\psi(\lambda)$ in (12) is, up to an additive constant, a convex quadratic form,

$$\psi(\lambda) = \psi_0(\lambda) + \int f^2(x)\mu(dx),$$

where

$$\psi_0(\lambda) = \lambda^T A \lambda - b^T \lambda$$

with

$$A_{ij} = \int f_i(x) f_j(x) \mu(dx), \; i, j = 1, \dots, M,$$

$$b_i = 2 \int f(x) f_i(x) \mu(dx), \; i = 1, \dots, M.$$

Let

$$\alpha = \big(\{A_{ij}, 1 \le i \le j \le M\}, \{b_i, i = 1, \dots, M\}\big) \in \mathbf{R}^{M+},$$
$$M^+ = M + M(M+1)/2,$$

be the vector of coefficients of the form ψ_0. It is immediately seen that an observation (x_t, y_t) from sample (1) generates the estimator

$$\alpha_t = \big(\{f_i(x_t) f_j(x_t), 1 \le i \le j \le M\}, \{2(f(x_t) + e_t) f_i(x_t), i = 1, \dots, M\}\big)$$

of the vector α. It is evident that α_t is an unbiased estimator of α,

$$\text{(13)} \qquad E\{\alpha_t\} = \alpha.$$

Then we can act as follows: in order to solve (12), we use the observations (1) to build the estimator

$$\bar{\alpha}_N = \frac{1}{N}\sum_{t=1}^{N}\alpha_t$$

of the vector α. Note that $\bar{\alpha}_N$ is the vector of coefficients of the convex quadratic form

$$\psi_0^N(\lambda) = \frac{1}{N}\sum_{t=1}^{N}\psi_t(\lambda),$$

where

$$\psi_t(\lambda) = \left((f(x_t)+e_t) - \sum_{i=1}^{M}\lambda_i f_i(x_t)\right)^2 - (f(x_t)+e_t)^2.$$

The problem of minimizing the form ψ_0^N over Λ is, using the terminology of [19], the stochastic counterpart of the problem (12), we are interested in. When solving the stochastic counterpart by an appropriate convex programming algorithm, we find a minimizer $\bar{\lambda}_N$ of the quadratic form ψ_0^N on Λ and take this minimizer as an estimator of the solution to (12).

The convergence properties of the outlined stochastic counterpart method are stated in the following.

THEOREM 2.1. *Let $M > 2$. For the stochastic counterpart method, one has*

$$E\{\psi(\bar{\lambda}_N)\} - \psi^* \le 8\sqrt{2e\ln M}L(2L+\sigma)N^{-1/2}, \tag{14}$$

where ψ^* is the optimal value in (12).

PROOF. We first remark that

$$\psi(\bar{\lambda}_N) - \psi^* \le \sup_{\lambda\in\Lambda}|\psi_0(\lambda) - \psi_0^N(\lambda)|. \tag{15}$$

On the other hand, since $\|\lambda\|_1 \le 1$ whenever $\lambda \in \Lambda$, for a quadratic form,

$$\phi(\lambda) \equiv \lambda^T B\lambda - d^T\lambda,$$

one has

$$\sup_{\lambda\in\Lambda}|\phi(\lambda)| \le 2\|\alpha[\phi]\|_\infty,$$

where $\alpha[\phi]$ is the vector of coefficients of the form ϕ. Consequently, (15) implies that

$$E\{\psi(\bar{\lambda}_N) - \psi^*\} \le 2E\{\|\alpha - \bar{\alpha}_N\|_\infty\} \equiv 2N^{-1}E\left\{\left\|\sum_{t=1}^{N}\zeta_t\right\|_\infty\right\}, \tag{16}$$

where $\zeta_t = \alpha - \alpha_t$. Note next that

$$|A_{ij} - f_i(x_t)f_j(x_t)| \le 2L^2$$

and

$$\big|b_i - 2(f(x_t)+e_t)f_i(x_t)\big| \le 4L^2 + 2L|e_t|.$$

Consequently,

$$\|\alpha - \alpha_t\|_\infty^2 \le 4(2L^2 + L|e_t|)^2,$$

which implies

$$E\{\|\alpha_t - \alpha\|_\infty^2\} \le 4(2L^2 + \sigma L)^2. \tag{17}$$

Now let us use the following technical result (for the proof, see [16]):

LEMMA 2.1. *Let $M > 2$, and let $q = 2\ln M$. Then the function*

$$W(z) = \tfrac{1}{2}\|z\|_q^2 \colon \mathbf{R}^M \to \mathbf{R}$$

satisfies for every z, $d \in \mathbf{R}^M$, the relation

$$W(z+d) \le W(z) + d^T \nabla W(z) + c^*(M)\|d\|_\infty^2, \qquad c^*(M) = 4e\ln M. \tag{18}$$

We have, by virtue of Lemma 2.1,

$$W\bigg(\sum_{t=1}^{k+1}\zeta_t\bigg) \le W\bigg(\sum_{t=1}^{k}\zeta_t\bigg) + (\zeta_{k+1})^T\nabla W\bigg(\sum_{t=1}^{k}\zeta_t\bigg) + c^*(M)\|\zeta_{k+1}\|_\infty^2,$$

whence, taking expectation and using (13), (17),

$$E\bigg\{W\bigg(\sum_{t=1}^{k+1}\zeta_t\bigg)\bigg\} \le E\bigg\{W\bigg(\sum_{t=1}^{k}\zeta_t\bigg)\bigg\} + 4c^*(M)(2L^2+\sigma L)^2.$$

It follows that

$$E\bigg\{W\bigg(\sum_{t=1}^{N}\zeta_t\bigg)\bigg\} \le 4Nc^*(M)(2L^2+\sigma L)^2,$$

and since $W(z) \ge \frac{1}{2}\|z\|_\infty^2$, we end up with

$$E\bigg\{\bigg\|\sum_{t=1}^{N}\zeta_t\bigg\|_\infty\bigg\} \le \sqrt{8Nc^*(M)}(2L^2+\sigma L) = 4\sqrt{2eN\ln M}(2L^2+\sigma L).$$

This estimate combined with (16) completes the proof. □

2.1. *Concentration.* A drawback of the aggregated estimator

$$f_N(x) = \sum_{i=1}^{M} \lambda_N^i f_i(x)$$

as given by the stochastic counterpart method is that the estimator, generally speaking, includes with nonzero weights all M basic functions. When M is large, which is the case we are mainly interested in, it might be computationally too expensive to use the estimator. This motivates the following natural question: whether it is possible to replace f_N with another estimator

$$\tilde{f}_N = \sum_{i=1}^{M} \tilde{\lambda}_N^i f_i(x), \qquad \tilde{\lambda}_N \in \Lambda$$

of basically the same quality, but with moderate number of nonvanishing coefficients $\tilde{\lambda}_N^i$.

We are about to show that if Λ is a "simple" set, then we can shrink the aggregated estimator; namely, to make it a combination of order of N or even of $N^{1/2}$ of the basic functions.

"N-CONCENTRATED" AGGREGATION. Let ω be an M-dimensional vector with the entries $\omega_i = \{\pm 1\}$; let

$$\mathbf{R}_\omega^M = \{\lambda \in \mathbf{R}^M \mid \omega_i \lambda_i \geq 0, i = 1, \ldots, M\}$$

be the corresponding orthant and let $\Lambda_\omega = \Lambda \cap \mathbf{R}_\omega^M$. Assume that Λ possesses the following property (where k is an integer parameter):

($\mathscr{A}_k$) *For every ω, the set Λ_ω is cut of the orthant $\mathbf{R}_\omega^M$ with at most k linear inequalities and equations.*

For example, the sets (4)–(6) satisfy ($\mathscr{A}_1$).

PROPOSITION 2.1. *Assume that Λ satisfies $\mathscr{A}_k$. Then the result $\bar{\lambda}_N$ of the SC method with N steps always can be chosen "$k+N$"-concentrated, that is, with no more than $k+N$ nonzero entries.*

PROOF. Let E^N be the orthogonal complement to the set of N M-dimensional vectors

$$(f_1(x_t), \ldots, f_M(x_t))^T,$$

$t = 1, \ldots, N$. E_N clearly is the recessive space of $\psi_0^N(\lambda)$, a translation of the argument of this quadratic form along E^N does not vary the value of the form. Now, let $\hat{\lambda}_N$ be an arbitrary minimizer of ψ_0^N over Λ, and let ω be such that $\hat{\lambda}_N \in \Lambda_\omega$. We can take as $\bar{\lambda}_N$ any point of the set $Q = \Lambda_\omega \cap [\hat{\lambda}_N + E^N]$. Since Q is a polyhedral set which is cut off $\mathbf{R}_\omega^M$ by not more than $k+N$ linear inequalities and equations, all its extreme points have at most $k+N$ nonzero entries, and we can choose as $\bar{\lambda}_N$ any one of these extreme points. □

Note that Theorem 2.1 combined with Proposition 2.1 implies the result of Theorem 1.1.

"$\sqrt{N}$-CONCENTRATED" AGGREGATION. The construction presented here goes back to [18]. We assume that Λ is the $\|\cdot\|_1$-ball (4); however, the following argument can be modified in a straightforward way to handle the cases of simplices (5) and (6).

Consider the following procedure. In order to aggregate the functions $f_1, \dots, f_M$, we first use the SC method; let λ_N be the resulting vector of weights. We set

$$\nu = \sum_{i=1}^{M} |\lambda_N^i|$$

and associate with λ_N the probability distribution π on the $2M+1$-element set $I = \{0, \pm 1, \dots, \pm M\}$ according to the following rule: the probability π_i of the element $i \neq 0$ is 0 if λ_N^i and i have different signs; otherwise it is $|\lambda_N^i|$; the probability of the element $i = 0$ is $\pi_0 = 1 - \nu$. When setting

$$\begin{aligned} &g_i(x) = f_i(x), i = 1, \dots, M,\\ &g_0(x) \equiv 0,\\ &g_i(x) = -g_{-i}(x), i = -1, -2, \dots, -M, \end{aligned}$$

we can represent the aggregated estimator $f_N = \sum_{i=1}^{M} \lambda_N^i f_i$ as

$$f_N = \sum_{i \in I} \pi_i g_i. \tag{19}$$

Now let us draw independently of each other K indices $i_1, \dots, i_K \in I$ according to the probability distribution $\{\pi_i\}_{i\in I}$ and take as a new aggregated estimator $\tilde{f}_N$ of f the function

$$\tilde{f}_N = \frac{1}{K} \sum_{l=1}^{K} g_{i_l}.$$

Note that $\tilde{f}_N$ clearly is of the form $\sum_{i=1}^{M} \tilde{\lambda}_N^i f_i$ with K-concentrated (i.e., with no more than K nonzero entries) weight vector $\tilde{\lambda} \in \Lambda$.

The quality of the estimator $\tilde{f}_N$ is given by the following simple proposition.

PROPOSITION 2.2. *One has*

$$\begin{aligned} E\{\psi(\tilde{\lambda}_N)\} - \psi^* &\le E\{\psi(\lambda_N)\} - \psi^* + K^{-1}L^2\\ &\le 8\sqrt{e \ln M} L(2L + \sigma)N^{-1/2} + K^{-1}L^2. \end{aligned} \tag{20}$$

Here E stands for the expectation with respect to the probability π and the distribution of observations.

PROOF. Let ξ be a random variable taking values in the Hilbert space of μ-square summable functions with the distribution ν as follows: with probability π_i the value of ξ is g_i, $i \in I$. By construction, $\tilde{f}_N(x) = (1/K)\sum_{l=1}^K \xi_l(x)$, where $\xi^K = (\xi_1, \dots, \xi_K)$ is a sequence of independent random variables with the distribution ν. By construction, ξ_i is independent of the observation $z^N = \{x_i, y_i\}_{i=1}^N$ used to compute λ_N. Meanwhile, (19) implies that f_N is the conditional (z^N being fixed) expectation of ξ: $f_N = E_\nu \xi$. Keeping this in mind and using evident notations for expectations, we obtain

$$\begin{aligned}
E\{\psi(\tilde{\lambda}_N)\} &= E_{z^N} E_\nu \left\| f - \frac{1}{K}\sum_{l=1}^K \xi_l \right\|_{2,\mu}^2 = E_\nu E_{z^N} E_\mu \left| f - \frac{1}{K}\sum_{l=1}^K \xi_l \right|^2 \\
&= E_\nu E_{z^N} E_\mu \left| f - E_\nu \xi + E_\nu \xi - \frac{1}{K}\sum_{l=1}^K \xi_l \right|^2 \\
&= E_\nu E_{z^N} E_\mu |f - E_\nu \xi|^2 + E_\nu E_{z^N} E_\mu \left| \frac{1}{K}\sum_{l=1}^K \xi_l - E_\nu \xi \right|^2 \\
&\le E\|f - f_N\|_{2,\mu}^2 + \frac{1}{K} E_{z^N} E_\nu E_\mu |\xi|^2 \le E\{\psi(\lambda_N)\} + K^{-1}L^2
\end{aligned}$$

(the latter inequality is due to the fact that the uniform norm of all f_i is bounded by L, so that the uniform norm of all realizations of ξ is $\le L$), and (20) follows. □

We see that in order to transform f_N into a "well-concentrated" estimator of basically the same quality it suffices to choose K in the above scheme in a way which ensures that, say

$$K^{-1}L^2 \le 8\sqrt{2e \ln M}\, L(2L + \sigma)N^{-1/2},$$

for example, as

$$K = \left\rfloor \frac{\sqrt{N}}{16\sqrt{2e \ln M}} \right\lfloor.$$

The corresponding randomized estimator $\tilde{f}_N$ is $O(\sqrt{N})$-concentrated and possesses the same quality (in fact, worse by a factor of 2) as our original estimator f_N.

3. Lower bound.

We have shown that when aggregating M functions on the basis of N observations (1), expected inaccuracy of aggregation

$$E\{\psi(\lambda_N)\} - \psi^*, \psi(\lambda) = \int \left(f(x) - \sum_{i=1}^N \lambda^i f_i(x) \right)^2 \mu(dx), \qquad \psi^* = \min_{\lambda \in \Lambda} \psi(\lambda)$$

can be made as small as $O(\sqrt{\ln M} N^{-1/2})$, with the constant factor in $O(\cdot)$ depending on the parameters L (a priori upper bound on the uniform norms of f and f_i) and σ [intensity of noise e_t in observations (1)]. A natural question is whether an aggregation with essentially better expected performance is possible. We are about to show that the answer to the latter question is negative in the minimax setting.

THEOREM 3.1. *For an appropriately chosen positive absolute constant κ, for every positive L, σ, integer $M > 2$ and every positive integer N satisfying the inequality*

$$\frac{\sigma^2 \ln M}{L^2} \leq N \leq \kappa \frac{\sigma^2 M \ln M}{L^2} \tag{21}$$

and for every method $\mathscr{B}$ solving the aggregation problem on the basis of N observations (1) *one can point out*:

(i) *M functions $f_1, \dots, f_M$ in $L_2[0,1]$ of the uniform norm not exceeding L*;

(ii) *a function $f \in L_2[0,1]$ which is a convex combination of the functions $f_1, \dots, f_M$, with the following property. Let*

$$f_N^{\mathscr{B}} = \sum_{i=1}^{M} (\lambda_N^{\mathscr{B},f})_i f_i$$

be the result obtained by $\mathscr{B}$ as applied to the aggregation problem given by the data

$$\{\mu, f, f_1, \dots, f_M, \Lambda, L\},$$

where μ is the Lebesgue measure on $[0,1]$ *and Λ is the standard simplex* (6), *and by observations* (1) *with $e_t \sim \mathscr{N}(0, \sigma^2)$. Then*

$$E\{\psi_f(\lambda_N)\} - \psi_f^* \geq \kappa L \sigma \sqrt{\frac{\ln M}{N}}, \tag{22}$$

where

$$\psi_f(\lambda) = \int \left(f(x) - \sum_{i=1}^{M} \lambda_i f_i(x) \right)^2 \mu(dx)$$

and

$$\psi_f^* = \min_{\lambda \in \Lambda} \psi_f(\lambda)$$

(*note that in fact $\psi_f^* = 0$*).

COMMENT. In the case of $L = O(1)\sigma$ the lower bound (22) differs from the upper bound (14) by an absolute constant factor only.

PROOF OF THEOREM 3.1. Let f_k, $k = 1, \dots, M$, be the first M cosines from the standard trigonometric basis in $L_2[0,1]$ multiplied by L,

$$f_k(x) = L \cos(2\pi k x).$$

Given a positive integer p, let us denote by $\mathscr{F}_p$ the set of all convex combinations of the functions $f_1, \dots, f_M$ with the coefficients as follows: $2p$ of the M

coefficients are equal to $(2p)^{-1}$, and other coefficients vanish. It is easily seen that if $p \le \sqrt{M}$, then $\mathscr{F}_p$ contains a subset $\mathscr{F}_p^*$ with the following properties:

(i) Every two distinct functions f, g from $\mathscr{F}_p^*$ have at most p common nonzero Fourier coefficients, so that

$$\frac{L^2}{4p} \le \|f - g\|_2^2 \le \frac{L^2}{2p}, \tag{23}$$

$\|\cdot\|_2$ being the standard norm in $L_2[0, 1]$.

(ii) The cardinality K of $\mathscr{F}_p^*$ satisfies the relation

$$K \ge M^{\kappa_1 p}. \tag{24}$$

Now let

$$\varepsilon(p) = \max_{f \in \mathscr{F}_p^*}\{E\{\psi_f(\lambda_N^{\mathscr{B}, f})\} - \psi_f^*\} = \max_{f \in \mathscr{F}_p^*} E\{\psi_f(\lambda_N^{\mathscr{B}, f})\}.$$

We claim that for any $p \le \sqrt{M}$ one has

$$\varepsilon(p) < \frac{L^2}{64p} \Rightarrow N \ge \kappa_2 L^{-2}\sigma^2 p^2 \ln M. \tag{25}$$

Consider the set of K hypotheses, where the kth hypothesis state that N observations in (1) are generated with the kth element of the set $\mathscr{F}_p^*$. Let us associate with $\mathscr{B}$ the method $\mathscr{B}'$ of distinguishing between K hypotheses which is as follows: given observations, we use $\mathscr{B}$ to solve the aggregation problem; when the corresponding aggregated estimator $f_{\mathscr{B}}$ is obtained we find the closest (in $\|\cdot\|_2$) to $f_{\mathscr{B}}$ element (any one of them in the non-uniqueness case) in $\mathscr{F}_p^*$ and claim that this is the function underlying our observations.

It is immediately seen if any one of our K hypotheses is true, the probability that $\mathscr{B}'$ fails to recognize it properly is at most 1/4. Indeed, assume that the true hypothesis is associated with $f \in \mathscr{F}_p^*$. If $\mathscr{B}'$ fails to say that this is the case, then the estimator $f_{\mathscr{B}}$ is at least at the same $\|\cdot\|_2$-distance from f as from some $g \in \mathscr{F}_p^*$ distinct from f. Taking into account the left inequality in (23), we conclude that then $\psi_f(\lambda_N^{\mathscr{B}, f}) \ge L^2/16p$; from the definition of $\varepsilon(p)$ and the Chebyshev inequality it follows that under the premise of (25) the probability of the event in question is at most 1/4, as claimed.

Now note that the Kullback distance between the distributions of N-observation samples (1) coming from two distinct elements of $\mathscr{F}_p^*$, in view of the right inequality in (23), is at most $N\sigma^{-2}L^2(2p)^{-1}$. Then the Fano inequality implies that the above K hypotheses can be distinguished only if

$$N\sigma^{-2}L^2(2p)^{-1} \ge \kappa_3 \ln K = \kappa_4 p \ln M$$

[we have used (24)], as required in the conclusion of (25). Let us now choose

$$p = \frac{\kappa_5 L}{\sigma}\sqrt{\frac{N}{\ln M}},$$

then by (21) we have $p < \sqrt{M}$. Furthermore, this value of p gives the desired bound (22) due to the left inequality of (23). Now the conclusion of Theorem 3.1 is an immediate consequence of (25). □

4. Functional aggregation: the stochastic approximation approach.

MOTIVATION. In the case when Λ is a "computationally tractable" convex compact set, for example, the $\|\cdot\|_1$-ball (4), or one of the simplices (5), (6), the stochastic counterpart scheme combined with any computationally efficient routine $\mathscr{S}$ for constrained convex optimization yields an implementable algorithm for aggregating functions $f_1, \dots, f_M$. However, in the case of large M (which is the case we actually are interested in) the resulting algorithm is rather costly computationally. Indeed, consider, for the sake of definiteness, the simplest case when Λ is the standard simplex (6). In our presentation of the stochastic counterpart method we spoke about exact identification of the optimal solution to the problem

$$\min\{\psi_0^N(\lambda) \mid \lambda \in \Lambda\}, \tag{26}$$

which is the stochastic counterpart of (12). In fact, of course, it suffices to find an approximate solution λ_N to the latter problem with inaccuracy, in terms of the objective $\psi_0^N(\lambda) - \min \Lambda\psi_0^N$, of order of

$$\varepsilon = \sqrt{N^{-1}\ln M}L(L+\sigma).$$

When replacing the exact solution to the stochastic counterpart with an approximate solution of the latter problem, we vary only the absolute constant factor in the right-hand side of (14).

Now, in the case of large M seemingly the best, from the viewpoint of overall computational complexity (i.e., total number of arithmetic operations) procedure for solving the stochastic counterpart within accuracy ε is the $\|\cdot\|_1$-version of the mirror descent method for large-scale convex minimization; see [17]. The method finds ε-solution to (26) in $O(N)$ iterations; the computational effort at a single iteration is dominated by the necessity to compute the value and the gradient of the objective ψ_0^N at the current iterate. In order to implement the method efficiently, we should first compute N M-dimensional vectors $(f_1(x_t), \dots, f_M(x_t))$, $t = 1, \dots, N$, let the arithmetic cost of this computation be $\mathscr{C}_{\text{est}}$. If $M < N$, it makes sense not to store these N vectors explicitly, but to assemble them into the vector of coefficients of the quadratic form ψ_0^N and to use these coefficients to compute $\psi_0^N(\lambda)$ and $\nabla\psi_0^N(\lambda)$; with this scheme, both the memory requirements and the arithmetic cost of computing ψ_0^N at a point will be $O(M^2)$. In the case of $M > N$ it is better not to assemble the coefficient of ψ_0^N explicitly, but to store the above N M-dimensional vectors and to compute ψ_0^N and its gradient via the representation

$$\psi_0^N(\lambda) = \frac{1}{N}\sum_{t=1}^{N}\left(f(x_t) - \sum_{i=1}^{M}\lambda_i f_i(x_t)\right)^2;$$

here both the memory requirements and the arithmetic cost of computing ψ_0^N at a point are $O(MN)$. Thus, the memory requirements for the stochastic counterpart method, the same as the arithmetic cost of an iteration, are $O(M\ \min\{N, M\})$. Recalling that we should perform $O(N)$ iterations of the method, we end up with the following complexity characteristics of the stochastic counterpart approach:

Stochastic counter approach	
Memory	$O(M\ \min\{N, M\})$
Total number of operations	$C_{\text{est}} + O(MN\ \min\{N, M\})$

We are about to develop another algorithm, based on non-Euclidean stochastic approximation, which yields the aggregation of basically the same quality as the one given by the stochastic counterpart approach, but with significantly less computational effort: the memory required by the SA algorithm is $O(M)$, and the overall arithmetic complexity is $C_{\text{est}} + O(MN)$:

Stochastic approximation approach	
Memory	$O(M)$
Total number operations	$C_{\text{est}} + O(MN)$

Recall that the indicated complexity bounds relate to the case of "simple" Λ, for example, (4), (5), or (6).

4.1. *The idea.* Let $(x_t, y_t = f(x_t) + e_t)$ be an observation from the sample (1). If for some $\lambda \in \mathbf{R}^M$ we denote

$$f_\lambda(x) = \sum_{i=1}^{M} \lambda_i f_i(x)$$

and

$$\psi(\lambda) = \int (f(x) - f_\lambda(x))^2 \mu(dx),$$

then we observe immediately that the vector $\xi_t(\lambda) \in \mathbf{R}^M$ with the entries

$$\xi_t^i(\lambda) = -2(y_t - f_\lambda(x_t)) f_i(x_t), \qquad i = 1, \dots, M, \tag{27}$$

is an unbiased estimate of $\nabla_\psi(\lambda)$, that is,

$$E\{\xi_t^i(\lambda)\} = -2\int (f(x) - f_\lambda(x)) f_i(x)\mu(dx) = \frac{\partial}{\partial \lambda_i}\psi(\lambda). \tag{28}$$

Recall that our objective is to solve the problem (12), and the relation (28) implies that we can compute unbiased estimates of $\nabla\psi(\cdot)$ from the observations (1). The latter fact suggests that we can achieve our objective via stochastic approximation (SA).

Which SA to use? It can be immediately seen that the standard SA,

$$\lambda_{t+1} = \pi_\Lambda[\lambda_t - \gamma_t \xi_t(\lambda_t)], \qquad \pi_\Lambda(\lambda) = \operatorname*{arg\,min}_{\lambda' \in \Lambda} \|\lambda - \lambda'\|_2, \tag{29}$$

does not fit the situation. First, our quadratic objective may be extremely ill-conditioned, and this may dramatically slow down the classical (with the step sizes $\gamma_t = O(t^{-1})$) SA; since we have no way to control the condition number of $\nabla^2\psi$, the classical SA is completely inappropriate for us. There are, however, "robust" versions of the process (29), those with the step sizes $\gamma_t = O(t^{-1/2})$ and the Cesaro averaging

$$\bar{\lambda}_t = \left(\sum_{\tau=1}^{t} \gamma_\tau\right)^{-1} \sum_{\tau=1}^{t} \gamma_\tau \lambda^\tau.$$

These algorithms attain the efficiency

$$E\Big\{\psi(\bar{\lambda}_t) - \min_{\Lambda} \psi(\lambda)\Big\} \le O(t^{-1/2}), \tag{30}$$

which is independent of the condition number of ψ. At the first glance, this revised SA fits the situation better. However, we are still confronted by the following problem: the constant factor in the right-hand side of (30) is proportional to the "$\|\cdot\|_2$-level of noise" $E\{\|\xi_t(\cdot) - \nabla_\psi(\cdot)\|_2^2\}$ in the observations of $\nabla\psi$. In our case this level, as is easily seen, is proportional to the number M of functions we intend to combine. In typical applications (see Section 1) M is very large, it is the cardinality of some multidimensional grid; as a result, the "constant" factor $O(M)$ in the right-hand side of (30) makes the robust versions of the standard SA useless for our purposes.

What seems to meet our needs is the *non-Euclidean robust SA associated with the L_1-norm* [17]. As we shall see in a while, this version of SA yields the efficiency estimate (30) with the constant factor in the right-hand side $O(\cdot)$ proportional to $\sqrt{\ln M}$, which fits our goals incomparably better than the versions of SA discussed above.

4.2. *The algorithm.* The robust SA algorithm, associated with $\|\cdot\|_1$-norm, for solving (12) is as follows. Let (cf. Lemma 2.1)

$$W(z) = \tfrac{1}{2}\|z\|_q^2 : \mathbf{R}^M \to \mathbf{R}, \qquad q = 2 \ln M,$$

$\theta \in (0, 1)$ and R be such that

$$R \ge \max_{\lambda \in \Lambda} \|\lambda\|_1 \tag{31}$$

(one can take $R = 1$).

Consider the following.

ALGORITHM 4.1 ($\|\cdot\|_1$-SA).

Initialization. Set $z^0 = 0$.

Step t for $t \geq 1$. Given z_{t-1}, $\|z_{t-1}\|_q \leq R$, acts as follows:

1. Compute

$$\lambda_t = \nabla W(z_{t-1})$$

and find the $\|\cdot\|_1$-projection ν_t of the vector λ_t onto Λ,

$$\nu_t \in \arg\min_{\lambda \in \Lambda} \|\lambda_t - \lambda\|_1.$$

2. Define vector $\Delta_t \in \mathbf{R}^M$ as follows:

(a) If $\nu_t = \lambda_t$, we set $\Delta_t = 0$.

(b) In the case of $\nu_t \neq \lambda_t$, by construction of ν_t, the interior of the $\|\cdot\|_1$-ball,

$$V_t = \{\lambda | \; \|\lambda - \lambda_t\|_1 \leq \rho_t \equiv \|\lambda_t - \nu_t\|_1\},$$

does not intersect the set Λ, and therefore int V_t and Λ can be separated by a linear form: there exists $a \neq 0$ such that

$$\min_{\lambda \in V_t} a^T \lambda \geq \max_{\lambda \in \Lambda} a^T \lambda. \tag{32}$$

We find such an a and set $\Delta_t = \|a\|_\infty^{-1} a$.

3. Using the observation (x_t, y_t) from the sample (1), compute the vector $\xi_t(\nu_t) \in \mathbf{R}^M$ with components

$$\xi_t^i(\nu_t) = -2(y_t - f_{\nu t}(x_t)) f_i(x_t)$$

[cf. (27)], where $f_{\nu t}(x_t) = \sum_{i=1}^M \nu_t^i f_i(x_t)$. Set

$$\bar{\xi}_t = \xi_t(\nu_t) + \|\xi_t(\nu_t)\|_\infty \Delta_t. \tag{33}$$

4. Put

$$\gamma_t = \frac{\theta^{1/4}}{4\sqrt{2e(1-\theta)\ln M}} \frac{R}{\sigma L} t^{-1/2}, \tag{34}$$

$$w_t = z_{t-1} - \gamma_t \bar{\xi}_t \tag{35}$$

and

$$z_t = \begin{cases} w_t, & \|w_t\|_q \leq R, \\ R w_t \|w_t\|_q^{-1}, & \|w_t\| q > R. \end{cases}$$

5. If $t < N$, go to step $t+1$, otherwise define the result of the algorithm as

$$\bar{\lambda}_N = \left(\sum_{t=K(N)}^{N} \gamma t \right)^{-1} \sum_{t=k}^{N} \gamma_t \nu_t, \qquad K(N) = \lfloor \theta N \rfloor. \tag{36}$$

The rate of convergence of Algorithm 4.1 is described in the following theorem.

THEOREM 4.1. *Let*

$$N \geq N^* \equiv \frac{512 e \ln M}{(1-\theta)\sqrt{\theta}} \frac{R^2 L^2}{\sigma^2}. \tag{37}$$

Then Algorithm 4.1 *with the step sizes* (34) *yields*

$$E\{\psi(\bar{\lambda}_N)\} - \psi^* \leq 24 \frac{\sqrt{2e \ln M}}{\theta^{1/4}\sqrt{1-\theta}} \frac{RL[\sigma + \sigma^{-1}\psi^*]}{\sqrt{N}}.$$

PROOF Let λ^* and $\psi^* = \psi(\lambda^*)$ be an optimal solution and the optimal value of (12), respectively. We start with the following simple observation.

LEMMA 4.1. *One has*

$$\|\bar{\xi}_t\|_\infty \leq 2\|\xi_t(\nu_t)\|_\infty, \tag{38}$$

$$E\{\|\xi_t(\nu_t)\|_\infty^2\} \leq 4L^2(\sigma^2 + E\psi(\nu_t)), \tag{39}$$

$$(\lambda_t - \lambda^*)^T \bar{\xi}_t \geq (\nu_t - \lambda^*)^T \xi_t(\nu_t) \tag{40}$$

and

$$(\lambda - \lambda^*)^T \nabla\psi(\lambda) \geq \psi(\lambda) - \psi^*, \qquad \lambda \in \mathbf{R}^M. \tag{41}$$

PROOF Equation (38) is an immediate consequence of the relation $\|\Delta_t\|_\infty \leq 1$; see Algorithm 4.1.

For $|f_i(\cdot)| \leq L$ and $\lambda \in \mathbf{R}^M$, we can bound

$$\|\xi_t(\lambda)\|_\infty = 2 \max_i |f(x_t) + e_t - f_\lambda(x_t)||f_i(x_t)| \leq 2L|f(x_t) + e_t - f_\lambda(x_t)|,$$

whence

$$\begin{aligned} E\{\|\xi_t(\nu_t)\|_\infty^2\} &\leq 4L^2 E\big[E_{x_t,e_t}\{(f(x_t) + e_t - f_{\nu_t}(x_t))^2\}\big] \\ &= 4L^2 E\{e_t^2\} + 4L^2 E\psi(\nu_t), \end{aligned}$$

and (39) as follows.

Let us prove (40). The inequality certainly holds when $\nu_t = \lambda_t$ since here $\bar{\xi}_t = \xi_t(\lambda_t)$ [cf. steps (1) and (2) of Algorithm 4.1]. Now assume that $\nu_t \neq \lambda_t$. In this case (32) is satisfied with $a = \Delta_t$ and the left-hand side in this relation is

$(\Delta_t)^T\lambda_t - \|\lambda_t - \nu_t\|_1$ due to $\|\Delta_t\|_\infty = 1$. On the other hand, the same left-hand side is equal to $(\Delta_t)^T\nu_t$, since ν_t belongs both to V_t and Λ. We conclude that

$$\max_{\lambda\in\Lambda}(\Delta_t)^T\lambda = (\Delta_t)^T\nu_t = (\Delta_t)^T\lambda_t - \|\lambda_t - \nu_t\|_1.$$

Hence

$$(\Delta_t)^T(\nu_t - \lambda) \ge 0 \qquad \forall\lambda \in \Lambda \tag{42}$$

and

$$(\Delta_t)^T(\lambda_t - \nu_t) = \|\lambda_t - \nu_t\|_1. \tag{43}$$

Setting $\xi = \xi_t(\nu_t),\ d = \nu_t - \lambda^*,\ \delta = \lambda_t - \nu_t$, we have

$$\begin{aligned}(\lambda_t - \lambda^*)^T\bar{\xi}_t &= (d+\delta)^T[\xi + \|\xi\|_\infty\Delta_t] && \text{[see (33)]}\\ &= d^T\xi + \delta^T\xi + \|\xi\|_\infty[d^T\Delta_t + \delta^T\Delta_t]\\ &\ge d^T\xi + \delta^T\xi + \|\xi\|_\infty\|\delta\|_1 && \text{[see (42), (43)]}\\ &\ge d^T\xi,\end{aligned}$$

as required in (40).

(41) follows from the convexity of ψ. □

We return now to the proof of the theorem. We set

$$W_*(z) = W(z) - z^T\lambda^*.$$

Let us track the evolution of the function $W_*(\cdot)$ along the trajectory $\{z_t\}$.

Step 1. We note first that

$$W_*(z_t) \le W_*(w_t). \tag{44}$$

Indeed, if $w_t \ne z_t$ then $w_t = sz_t$ with $s > 1$, and $\|z_t\|_q \ge R$. If for a $r \ge 1$ we denote

$$\zeta'(r) = r^2W(z_t) - r(z_t)^T\lambda^*$$

then $W_*(w_t) = \zeta(s)$. However, we observe that for $r \le 1$,

$$\begin{aligned}\zeta'(r) &= 2rW(z_t) - (z_t)^T\lambda^*\\ &\ge r\|z_t\|_q^2 - \|z_t\|_q\|\lambda^*\|_p \quad [q^{-1} + p^{-1} = 1]\\ &\ge R^2 - R\|\lambda^*\|_1\\ &\ge 0 \qquad \text{[see (31)]},\end{aligned}$$

whence $\zeta(s) > \zeta(1) = W_*(z_t)$ and (44) follows.

Step 2. $W_*(z_t)$ satisfies the following recursive inequality:

$$W_*(z_t) \le W_*(z_{t-1}) - \gamma_t[\xi_t(\nu_t)]^T(\nu_t - \lambda^*) + 4c^*(M)\gamma_t^2\|\xi_t(\nu_t)\|_\infty^2, \tag{45}$$

where $c^*(M) = 4\,e\,\ln(M)$.

Using Lemma 4.1 we obtain for the increment of $W_*(z_t)$,

$$\begin{aligned} W_*(z_t) &\le W_*(w_t) && \text{[see (44)]} \\ &= W_*(z_{t-1} - \gamma_t\bar{\xi}_t) && \text{[(35)]} \\ &\le W_*(z_{t-1}) - \gamma_t(\bar{\xi}_t)^T\nabla W_*(z_{t-1}) + c^*(M)\gamma_t^2\|\bar{\xi}_t\|_\infty^2 && \text{[see Lemma 2.1]} \\ &= W_*(z_{t-1}) - \gamma_t(\bar{\xi}_t)^T(\lambda_t - \lambda^*) + c^*(M)\gamma_t^2\|\bar{\xi}_t\|_\infty^2 && \text{[structure of } W_*\text{]} \\ &\le W_*(z_{t-1}) - \gamma_t(\xi_t(\nu_t))^T(\nu_t - \lambda^*) && \text{[(44), (40)]} \\ &\quad + 4c^*(M)\gamma_t^2\|\xi_t(\nu_t)\|_\infty^2 \end{aligned}$$

where $c^*(M)$ is defined as in (18).

Step 3. We denote $\varepsilon(N) = E\{\psi(\bar{\lambda}_N)\} - \psi^*$. Then we have the bound

$$\varepsilon(N) \le \left(\sum_{t=K}^{N}\gamma_t\right)^{-1}\left(2R^2 + 16L^2c^*(M)\sum_{t=K}^{N}\gamma_t^2(\sigma^2 + E\{\psi(\nu_t)\})\right), \tag{46}$$

where $K = K(N) = \lfloor \theta N \rfloor$.

Let $\nu_t = EW_*(z_t)$. We now take the expectation on both sides of the inequality (45). We first take the conditional expectation over (x_t, y_t), the previous observations being fixed. Then we take the expectation over previous observations, and using the bound (39) we obtain

$$\nu_t \le \nu_{t-1} - \gamma_t E\{(\nu_t - \lambda^*)^T\nabla\psi(\nu_t)\} + 16c^*(M)L^2\gamma_t^2E\{\sigma^2 + \psi(\nu_t)\}.$$

Due to (41) the latter inequality implies

$$\nu_t \le \nu_{t-1} - \gamma_t E\{\psi(\nu_t) - \psi^*\} + 16c^*(M)L^2\gamma_t^2(\sigma^2 + E\psi(\nu_t)), \tag{47}$$

Next we deduce from (47) the recursive inequality

$$\gamma_t E\{\psi(\nu_t)\} - \psi^* \le \nu_{t-1} - \nu_t + 16L^2c^*(M)\gamma_t^2(\sigma^2 + E\{\psi(\nu_t)\}).$$

When summing up over $t = K \equiv K(N), K+1, \dots, N$, we get

$$\begin{aligned} \sum_{t=K}^{N}\gamma_t E\{\psi(\nu_t) - \psi^*\} &\le v_{K-1} - v_N \\ &\quad + 16L^2c^*(M)\sum_{t=K}^{N}\gamma_t^2(\sigma^2 + E\{\psi(\nu_t)\}). \end{aligned} \tag{48}$$

Now note that $\|z_t\|_q \le R$ for all t, thus

$$\begin{aligned} W_*(z_{K-1}) - W_*(z_N) &= \tfrac{1}{2}[\|z_{K-1}\|_q^2 - \|z_N\|_q^2] - (z_{K-1} - z_N)_t\lambda^* \\ &\le \tfrac{1}{2}[\|z_{K-1}\|_q^2 - \|z_N\|_q^2] + \|z_{K-1} - z_N\|_q\|\lambda^*\|_1 \\ &\le \tfrac{1}{2}[\|z_{K-1}\|_q^2 - \|z_N\|_q^2] + R\|z_{K-1} - z_N\|_q \\ &\le 2R^2, \end{aligned}$$

so that $v_{K-1} - v_N \le 2R^2$, and we get from (48),

$$\sum_{t=K}^N \gamma_t E\{\psi(\nu_t) - \psi^*\} \le 2R^2 + 16L^2c^*(M)\sum_{t=K}^N \gamma_t^2(\sigma^2 + E\{\psi(\nu_t)\}). \tag{49}$$

By the Jensen inequality we conclude from (36) that

$$E\{\psi(\bar\lambda_N) - \psi^*\} \le \left(\sum_{t=K}^N \gamma_t\right)^{-1}\left(\sum_{t=K}^N \gamma_t E\{\psi(\nu_t) - \psi^*\}\right)$$

and, when substituting from (49),

$$E\{\psi(\bar\lambda_N) - \psi^*\} \le \left(\sum_{t=K}^N \gamma_t\right)^{-1}\left(2R^2 + 16L^2c^*(M)\sum_{t=K}^N \gamma_t^2(\sigma^2 + E\{\psi(\nu_t)\})\right).$$

Step 4. In order to extract from the bound (46) a reasonable stepsize policy, let us see what happens when γ_t tends steadily to zero, that is,

$$\gamma_t \to 0, \qquad \gamma_{t+1} \le \gamma_t \quad \text{and } \frac{\gamma_{K(n)}}{\gamma_N} \le C < \infty. \tag{50}$$

If we set $\Omega = 16L^2c^*(M)$, $l(N) = N - K(N) + 1$ and

$$\alpha(N) = \frac{\sum_{t=K(N)}^N \gamma_t E\{\psi(\nu_t) - \psi^*\}}{\sum_{t=K(N)}^N \gamma_t}$$

we obtain from (46) the inequality

$$\varepsilon(N) \le \alpha(N) \le \left\{\frac{2R^2}{l(N)_{\gamma_N}} + C\Omega(\sigma^2 + \psi^*)_{\gamma_N}\right\} + C\Omega_{\gamma_N}\alpha(N). \tag{51}$$

We conclude from this inequality that if the gain sequence satisfies (50) and $l(N)_{\gamma_N} \to \infty$ as $N \to \infty$, then both $\varepsilon(N)$ and $\alpha(N)$ converge to 0 as $N \to \infty$. The rate of convergence, at least for large N, is given by the bracketed term in the right-hand side of (51). Since $l(N) \approx (1-\theta)N$, the best rate of convergence of the bracketed term to 0 as $N \to \infty$ is $O(N^{-1/2})$, and the corresponding choice of γ_t is

$$\gamma_t = R\sqrt{\frac{2\theta^{1/2}}{(1-\theta)\Omega(\sigma^2 + \psi^*)}}t^{-1/2}. \tag{52}$$

This choice of γ_t in (52) involves the unknown optimal value ψ^* of problem (12). If we set $\psi^* = 0$ in (52) and substitute for Ω its expression via L and M, we obtain the expression (34) for γ_t.

Step 5. To finish the proof it suffices to note that if the inequality in (37) is satisfied, then the coefficient at $\alpha(N)$ in (51) is $\leq 1/2$. Therefore, (51) implies that

$$\varepsilon(N) \leq \alpha(N) \leq 2\left\{\frac{2R^2}{l(N)_{\gamma_N}} + C\Omega(\sigma^2 + \psi^*)_{\gamma_N}\right\}.$$

When taking into account that $l(N) \geq (1-\theta)N, C = \theta^{-1/2}$, and substituting expressions for Ω and γ_N we obtain the required bound for $\varepsilon(N)$. □

5. Application: restoring functions from the Jones–Barron class. We now apply the result of Section 3 to estimation in the Jones–Barron model, Example 1 of the Introduction.

Class $\mathscr{F}_N^d(L, \gamma, \nu)$. Let L, γ, ν be positive reals such that $\gamma \leq 1$ and d, N be positive integers. We associate with the tuple (L, γ, ν, d, N) the class $\mathscr{F}_n^d(L, \gamma, \nu)$ comprising all functions $f \colon \mathbf{R}^d \to \mathbf{R}$ which are Fourier transforms of finite Borel complex-valued measures on $\mathbf{R}^d$,

$$f(x) = \int \exp(iw^T x)\widehat{F}(dw),$$

such that

$$\int |\widehat{F}(dw)| \leq \frac{L}{\sqrt{2}},$$

$$\int_{|w|>\rho} |\widehat{F}(dw)| \leq \gamma^{-1}\rho^{-\gamma}N^{\nu} \qquad \forall \rho > 0.$$

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

Let also μ be a probability distribution on $\mathbf{R}^d$ such that

$$\int |x|^2 \mu(dx) \leq \sigma_x^2 < \infty. \tag{53}$$

The problem is to recover a function $f \colon \mathbf{R}^d \to \mathbf{R}$, given N observations $(x_t, y_t = f(x_t) + e_t)$ of the function [cf. (1)]; here x_t are independent random vectors with the common distribution μ, and e_t are independent of each other and of x_t real errors satisfying (2). We assume that we know in advance the parameters L, γ, ν of the class $\mathscr{F}_N^d(L, \gamma, \nu)$, as well as the quantities σ_x from (53) and σ from (2).

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

We can now use the aggregation procedure described in Section 4 to find an approximation $\hat{\lambda}$ of the minimizer λ^* of the functional

$$\psi_f(\lambda) = \int\biggl(f(x) - \sum_{k=1}^{K}\bigl[\lambda_{2k-1}L\cos(\omega_k^T x) + \lambda_{2k}L\sin(\omega_k^T x)\bigr]\biggr)^2 \mu(dx)$$

on the set $\Lambda = \{\lambda \in \mathbf{R}^{2M}\colon \|\lambda\|_1 \le 1\}$. The result of Theorem 4.1 suggests that

$$\psi(\hat{\lambda}) - \psi(\lambda^*) = O\biggl(\sqrt{\frac{\ln N}{N}}\biggr),$$

where $\psi(\lambda^*)$ is the true minimum of $\psi(\lambda)$ on Λ. Due to Jones–Barron's approximation result we know that the latter quantity is small.

The algorithm implementing the above idea is as follows:

ALGORITHM 5.1.

1. Given $N, d, L, \gamma, \nu, \sigma$ and σ_x, we set

$$\sum_N \equiv \sum(N, d, L, \gamma, \nu, \sigma, \sigma_x) = \frac{\sigma_x N^{(4\nu+\gamma+1)/4\gamma}}{d^{(\gamma+1)/4\gamma}\gamma^{1/\gamma}L^{(1-\gamma)/2\gamma}\sigma^{(\gamma+1)/2\gamma}}, \tag{54}$$

$$\eta^2 = \sqrt{\frac{d \ln \sum_N}{N}}L\sigma, \qquad \rho(\eta) = \frac{N^{\nu/\gamma}}{(\gamma\eta)^{1/\gamma}}, \tag{55}$$

$$\varepsilon = \frac{\eta}{2L\sigma_x}. \tag{56}$$

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

$$W_{\rho(\eta)} = \bigl\{w \in \mathbf{R}^d : |w| \le \rho(\eta)\bigr\}$$

with ε given by (56). The cardinality K of the net is assumed to satisfy the inequality

$$K \le (1 + 2\varepsilon^{-1}\rho(\eta))^d \tag{57}$$

(such a net for sure exists).

3. Let $M = 2K$, $\Lambda = \{\lambda \in \mathbf{R}^M\colon \|\lambda\|_1 \le 1\}$ and

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

We use the stochastic approximation algorithm described in Section 4 to find approximation $\hat{\lambda}$ of the point

$$\lambda^* = \arg\min_{\lambda\in\Lambda}\psi(\lambda), \qquad \psi(\lambda) = \int(f(x) - f_\lambda(x))^2\mu(dx).$$

When applying Algorithm 4.1, we treat the M functions $L\cos(\omega^T x)$; $L\sin(\omega^T x)$, $\omega \in \Omega$, as the functions to be aggregated and set

$$\Lambda = \{\lambda \in \mathbf{R}^M |\ \|\lambda\|_1 \le 1\}.$$

Finally, the step sizes γ_t in Algorithm 4.1 are chosen according to (34) with $R = 1$ and $\theta = 0.5$.

The convergence rate of the resulting estimator

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

of f is given by the following.

THEOREM 5.1. *Let $f \in \mathscr{F}_N^d(L, \gamma, \nu)$, and let* (2), (53) *be satisfied. Then for all large enough N [i.e., $N \geq N_0(L, \gamma, \nu, d, \sigma, \sigma_x)$] one has*

$$E\{\|\hat{f}_N(x) - f(x)\|_{2,\mu}^2\} \leq \kappa L \sigma \sqrt{\frac{d \ln \sum_n}{N}}, \tag{58}$$

with Σ_N given by (54) *and an absolute constant κ.*

PROOF.

Step 1. According to (57) and (54)–(56) we have for all large enough N,

$$M \leq (1 + 2\varepsilon^{-1}\rho(\eta))^d \leq (\kappa_1 \varepsilon^{-1}\rho(\eta))^d \leq \left(\kappa_2 \sum_N\right)^d \tag{59}$$

Step 2. Let us verify that for every $f \in \mathscr{F}_N^d(L, \gamma, \nu)$ there exists a function

$$\tilde{f}(x) = \sum_{k=1}^{M} \big[\lambda_{2k-1} L \cos(\omega_k^T x) + \lambda_{2k} L \sin(\omega_k^T x)\big]$$

with $\|\lambda\|_1 \leq 1$ and $\omega_k \in \Omega$ such that

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

Indeed, by (55) we have

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

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

$$\|f - g\|_{2,\mu} \leq \|f - g\|_\infty \leq \eta. \tag{60}$$

On the other hand, it follows from Barron's proof of (9) (see [2]) that one can find a function of the form

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

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

$$\|h - g\|_{2,\mu} \le \eta. \tag{61}$$

Next we note that for any $\omega, \omega' \in \mathbf{R}^d$,

$$\begin{aligned}\int \left|\exp(ix^T\omega) - \exp(ix^T\omega')\right|^2 \mu(dx) &\le 4|\omega - \omega'|^2 \int |x|^2 \mu(dx)\\ &= 4|\omega - \omega'|^2 \sigma_x^2.\end{aligned} \tag{62}$$

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

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

we obtain due to (62),

$$\begin{aligned}\|h - r\|_{2,\mu} &\le L \max_k \left(\int \left|\exp(ix^T\omega_k) - \exp(ix^T\zeta_k)\right|^2 \mu(dx)\right)^{1/2}\\ &\le 2L \max_k |\omega_k - \zeta_k| \sigma_x = \eta\end{aligned}$$

[see (56)]. Along with (60) and (61) this estimate yields $\|f - r\|_{2,\mu} \le 3\eta$. Now we can set $f(x) = \mathrm{Re}\{r(x)\}$.

Step 3. Applying Theorem 4.1 (where one should set $R = 1, \theta = 0.5$) to the $2M$ functions

$$\left\{L\cos(\omega^T x), L\sin(\omega^T x)\right\}_{\omega\in\Omega}$$

and taking into account that, by Step 2 of the proof, in our situation $\psi^* \equiv \min_{\lambda\in\Lambda} \psi(\lambda) \le 9\eta^2$, we get

$$E\left\{\|\hat{f}_N - f\|_{2,\mu}^2\right\} \le \kappa_3\eta^2 + \kappa_3\left(\frac{\ln M}{N}\right)^{1/2} L(\sigma + 9\sigma^{-1}\eta^2),$$

and the latter quantity, as it is immediately seen from (59) and (54)–(56), for all large enough values of N is bounded from above by $\kappa\sqrt{d\ln\Sigma_N/N}$ with properly chosen absolute constant κ. □

5.1. *Lower bound.* We are about to show that the rate of convergence given by Theorem 5.1 cannot be improved significantly in the minimax sense.

THEOREM 5.2. *Let $L > 0$. Consider the problem of estimating a univariate function $f(x): R \to R$ from observations $(x_t, y_t = f(x_t) + e_t)$, $t = 1, \dots, N$, where x_t, e_t are mutually independent, x_t are uniformly distributed on* $[0, 1]$ *and $e_t \sim \mathcal{N}(0, \sigma^2)$. Let $\mathcal{F}_N^*(L)$ be the class $\mathcal{F}_N^1(L, 1, 1)$. Then for some absolute constant and all large enough values of N for every algorithm $\mathcal{B}$ approximating*

$f \in \mathscr{F}_N^*(L)$ on the basis of the above observations one has

$$\sup_{f \in \mathscr{F}^*(R)} E\{\|\hat{f}_{\mathscr{B}} - f\|_2^2 \geq \kappa L \sigma \sqrt{\frac{\ln N}{N}}, \tag{63}$$

where $\hat{f}_{\mathscr{B}}$ is the estimator yielded by $\mathscr{B}$, the function underlying the observations being f.

SKETCH OF THE PROOF. Let

$$p = \rfloor \frac{L}{\sigma}\sqrt{\frac{N}{\ln N}} \lfloor,$$

and $\Lambda(N) = \{\lambda \in \mathbf{R}^N\}$ be a set of vectors such that $2p$ entries of λ_k of λ are equal to $2^{-1/2}(2p)^{-1}$ and other entries vanish. Note that for all large enough N the set is nonempty and $\lambda \in \Lambda(N)$ implies

$$\{\|\lambda\|_1 \leq 2^{-1/2}\}.$$

We denote $\phi(x) = (\phi_1(x), \dots, \phi_N(x))$ the vector-valued function with the components $\phi_k(x) = L\cos(2\pi k x)$. Then for any $\lambda \in \Lambda(N)$ the 1-periodic function $f_\lambda(x) = \lambda^T\phi(x)$ clearly belongs to $\mathscr{F}_N^*(L)$. On the other hand, as in the proof of Theorem 3.1, for all large enough N we can extract from $\mathscr{F}_N^*(L)$ the set $\mathscr{F}_N^+$ of cardinality greater than or equal to $N^{\kappa_1 p}$, κ_1 being the appropriately chosen absolute constant, in such a way that for every two distinct functions $f, g \in \mathscr{F}_N^+$ one has

$$\frac{L^2}{8p} \leq \|f - g\|_{2,\mu}^2 \leq \frac{L^2}{4p}.$$

Now we can use exactly the same arguments as in the proof of Theorem 3.1 to get the desired lower bound (63). □

6. Numerical examples. In this section we present simulation results for the SA aggregation algorithm as applied to the *nonparametric filtering problem*.

Nonparametric filtering problem. Consider a dynamic system

$$y_t = f(y_{t-1}, \dots, y_{t-d}) + e_t, \tag{64}$$

$e_0, e_1, \dots$ being independent noises. The problem is to predict, given observations $y_1, \dots, y_N$, the state y_{N+1}. We suppose that the order d of the system is known a priori.

Now assume that the dynamics of (64) has the following known-in-advance structure:

$$f(x) = g(p^T x), \tag{65}$$

where g is a function on $\mathbf{R}$ and p is an unknown vector of parameters.

Being a bit sloppy, we can deal with the model (64) as if the observations (y_t) were produced by the model (1) with $x_t = (y_{t-1}, \dots, y_{t-d})^T$. We can now build an estimator $\hat{f}_N$ using observations $y_1, \dots, y_N$ according to the scheme presented in Example 2 of the Introduction. Then we can form the forecast

$$\hat{y}_{N+1} = \hat{f}(x_{N+1}).$$

Though in this example x_t are dependent, under reasonable stability assumptions there exists steady state distribution μ of x_t and the situation is "not too far" from the model (1) where x_t are independent variables with common distribution μ. Note that in this example the measure μ is determined by the unknown regression function itself and can be rather sophisticated. For example, for the simple dynamics

$$(66) \qquad (\mathscr{D}_d): \begin{cases} y_t = F(p^T x_t) + \sigma n_t, & x_t = \begin{pmatrix} y_{t-1} \\ y_{t-2} \\ \cdots \\ y_{t-d} \end{pmatrix}, \\ F(z) = \cos(4\pi z) + \cos(5\pi z), & \\ n_t \sim \mathscr{N}(0, 1), & p = \frac{1}{\sqrt{d}} \begin{pmatrix} 1 \\ 1 \\ \cdots \\ 1 \end{pmatrix} \end{cases}$$

with $d = 2$, $\sigma = 0.1$ the plot of the first 2^{20} pairs $x_t = (y_t, y_{t-1})$ looks as shown in Figure 1.

The dynamics we deal with in simulations is $(\mathscr{D}_d)$ with $d = 2$ and $d = 3$. In order to build the estimator $\hat{f}_N$, we define a grid of M unit directions p_i, $i = 1, \dots, M$, in $\mathbf{R}^d$. In the case of $d = 2$ it was the uniform grid,

$$p_i = \left(\cos\left(\phi_0 + \frac{i}{M}\pi\right), \sin\left(\phi_0 + \frac{i}{M}\pi\right)\right), \qquad k = 1, \dots, M,$$

ϕ_0 being a randomly chosen "phase shift"; in the case of $d = 3$ we chose the directions p_i randomly.

We use the first

$$N = 1024$$

observations to build M nonparametric estimators $f_i(x) = \phi_i(p_i^T x)$ of the function $f^*(x) = F(p^T x)$. When building ϕ_i, we act as if the observations were

$$y_t = \phi(p_i^T x_t) + \sigma n_t$$

for some ϕ. Estimators ϕ_i are obtained by the spatial adaptive nonparametric estimator applied to this model of observations. This estimator originates from [10]; we have modified it in an evident way to handle arbitrary design of the regressors, instead of the regular design studied in [10].

After building the M estimators f_i on the basis of the first $N = 1024$ observations, we use N remaining observations to approximate the best convex

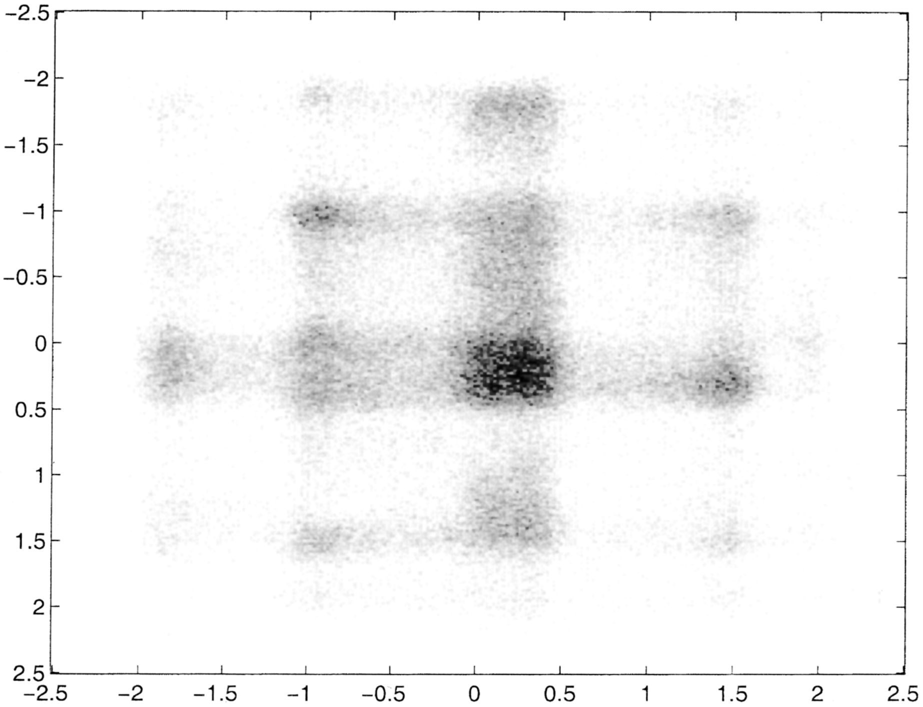

FIG. 1. *Plot of* 2^{20} *points* x_t *given by dynamics* $(\mathscr{D}_2)$, $\sigma = 0.1$.

combination of the estimators $f_1, \dots, f_M$, that is, the optimal solution to the problem (12) associated with

$$\Lambda = \left\{ \lambda \in \mathbf{R}^M | \lambda \geq 0, \sum_{i=1}^{M} \lambda_i = 1 \right\}.$$

To this end we use Algorithm 4.1 with the step sizes (34) associated with the setup

$$R = 2, \kappa = 1/3, L = \max_{1 \leq t \leq N} |y_t|.$$

The estimator

$$\bar{f}(x_t) = \sum_{i=1}^{M} (\bar{\lambda}_N)_i f_i(x_t)$$

provided by the algorithm is then used to predict the "regular" component $y_t^* \equiv f(x_t)$ of y_t, $t = 2N+1, \dots, 4N$. Below we refer to the indicated predictor as to the *structure-based* one.

TABLE 1
Empirical mean square error of prediction

Method	$\boldsymbol{\sigma = 0.1}$	$\boldsymbol{\sigma = 0.33}$
Structure-based predictor, dynamics $(\mathscr{D}_2)$	0.093	0.275
Structure-based predictor, dynamics $(\mathscr{D}_3)$	0.107	0.288

We run two series of experiments: the first for the intensity σ of the noise in $(\mathscr{D}_d)$ equal to 0.1, and the second for $\sigma = 0.33$. In both experiments, the number M of estimators to be combined was set to 400 in the case $d = 2$ and to 3144 in the case $d = 3$. The quality of a predictor $\bar{y}_t = \bar{f}(x_t)$ was measured by the empirical mean square error

$$\delta = \sqrt{\frac{1}{2N} \sum_{t=2N+1}^{4N} (y_t^* - \bar{f}(x_t))^2}.$$

The numerical results are as shown in Table 1.

In Figures 2 and 3 we present the result of the structure-based reconstruction of the true dynamics via $2N = 2048$ observations for $\sigma = 0.1$ and $\sigma = 0.33$, respectively.

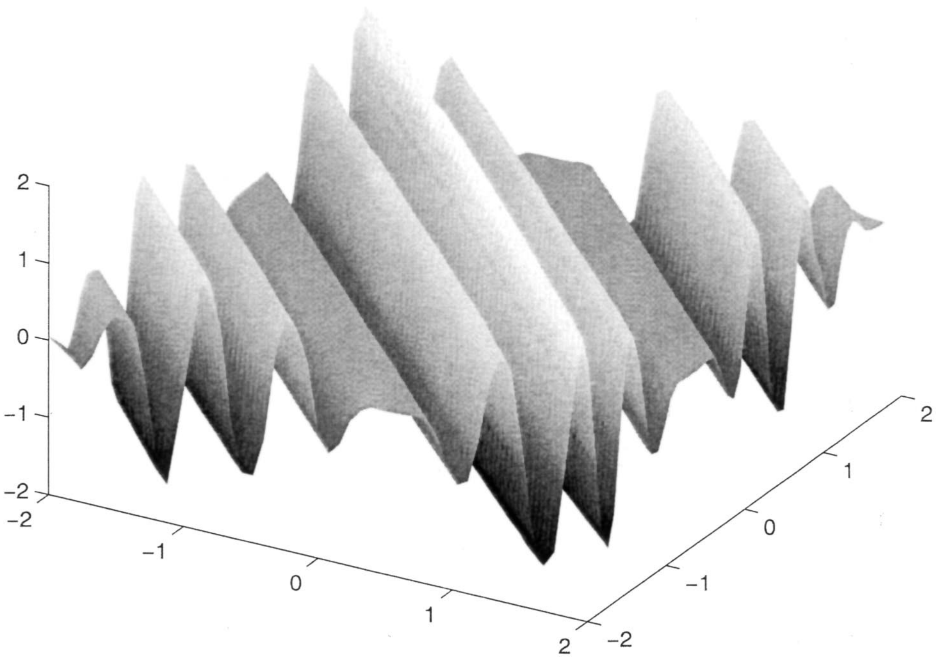

FIG. 2. *Structure-based reconstruction of dynamics* $(\mathscr{D}_2)$, $\sigma = 0.1$.

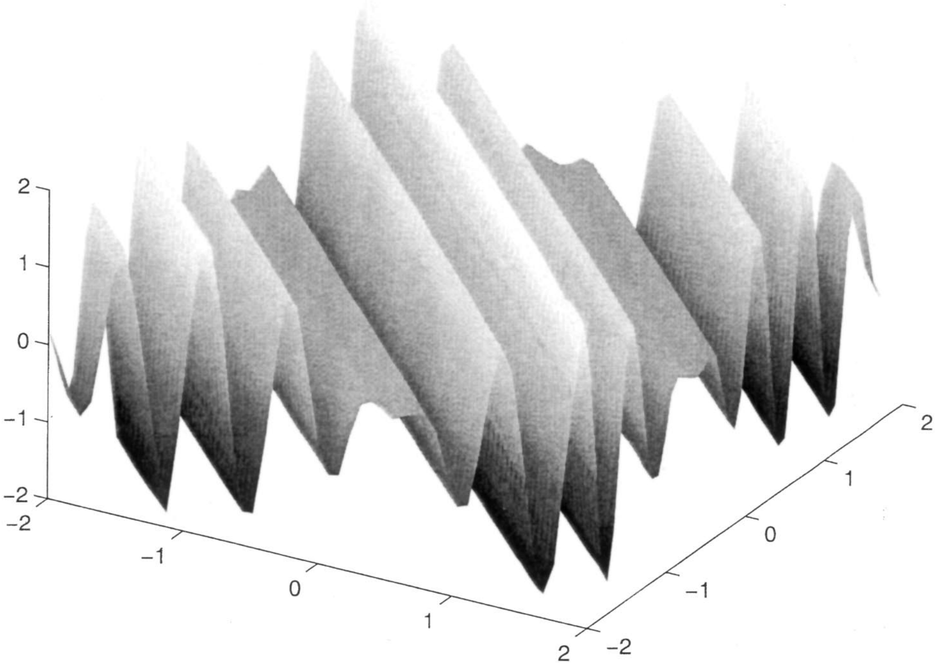

FIG. 3. *Structure-based reconstruction of dynamics* ($\mathscr{D}_2$)*,* $\sigma = 0.33$.

REFERENCES

[1] BREIMAN, L. (1993). Hinging hyperplanes for regression, classification and function approximation. *IEEE Trans. Inform. Theory* **39** 999–1013.

[2] BARRON, A. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory* **39** 930–945.

[3] BARRON, A. (1991). Complexity regularization with application to artificial neural networks. In *Nonparametric Functional Estimation and Related Topics* (G. Roussas ed.) Kluwer, Netherlands.

[4] BARRON, A. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14** 115–133.

[5] BREIMAN, L., FRIEDMAN, J. M., OLSHEN, J. H. and STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

[6] DONOHO, D., JOHNSTONE, I., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

[7] EFROIMOVICH, S. Y., and PINSKER, M. S. (1984). A learning algorithm for nonparametric filtering. *Automatika i Telemehanika* (in Russian). (English translation in *Automat. Remote Control* **11** 58–65.)

[8] FRIEDMAN, J. H. and STUETZLE, W. (1981). Projection pursuit regression. *J. Amer. Statist. Assoc.* **76** 817–823.

[9] FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.* **19** 1–141.

[10] GOLDENSHLUGER, A. and NEMIROVSKI, A. On spatial adaptive nonparametric regression. Research Report 5/94, Optimization Lab. Faculty of Industrial Engineering and Management, Technion.

[11] HÄRDLE, W. (1990). *Applied Nonparametric Regression*. Cambridge Univ. Press.

[12] HUBER, P. J. (1985). Projection pursuit (with discussion). *Ann. Statist.* **13** 435–525.

[13] JONES, L. K. (1992). A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. *Ann. Statist.* **20** 608–613.

[14] LEE, W. S., BARLETT, P. L. and WILLIAMSON, R. C. (1996). Efficient agnostic learning of neural networks with bounded fanin. *IEEE Trans. Inform. Theory* **42** 2118–2132.

[15] MORGAN, J. N. and SONQUIST, J. A. (1963). Problems in the analysis of survey data, and a proposal. *J. Amer. Statist. Assoc.* **58** 415–434.

[16] NEMIROVSKI, A. (1992). On nonparametric estimation of functions satisfying differential inequalities. In *Advances in Soviet Mathematics* (R. Khasminski, ed.) **12** 7–43. Amer. Math. Soc. Washington, DC.

[17] NEMIROVSKI, A. and YUDIN, D. (1983). *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York.

[18] PISIER, G. (1981). Remarques sur un resultat non publie de B. Maurey. In *Seminaire d'analyse fonctionelle.* **112**. Ecole Polytechnique, Palaiseau.

[19] RUBINSHTEIN, R. and SHAPIRO, A. (1995). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization via the Score Function Method*. Wiley, New York.

[20] KOROSTELEV, A. and TSYBAKOV, A. (1991). *Minimax Theory of Image Reconstruction*. Springer, New York.

LMC, 51 RUE DE MATHÉMATIQUES
DOMAINE UNIVERSITAIRE, BPS3
GRENOBLE, CEDEX 9
FRANCE
E-MAIL: juditsky@inrialpes.fr

FACULTY OF INDUSTRIAL ENGINEERING
AND MANAGEMENT AT TECHNION
TECHNION, ISRAEL INSTITUTE
OF TECHNOLOGY
TECHNION CITY, HAIFA 32000
ISRAEL
E-MAIL: nemirovs@i.e.technion.ac.il