HAL
open science

Nonparametric denoising signals of unknown local structure, II: Nonparametric function recovery

Anatoli B. Juditsky, Arkadii S. Nemirovski

To cite this version:

Anatoli B. Juditsky, Arkadii S. Nemirovski. Nonparametric denoising signals of unknown local structure, II: Nonparametric function recovery. Applied and Computational Harmonic Analysis, 2010, 29 (3), pp.354-367. 10.1016/j.acha.2010.01.003 . hal-00365531

HAL Id: hal-00365531

https://hal.science/hal-00365531v1

Submitted on 3 Mar 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Nonparametric denoising Signals of Unknown Local Structure, II: Nonparametric Regression Estimation

Anatoli Juditsky [a,*], Arkadi Nemirovski [b]

[a] *LJK, B.P. 53, 38041 Grenoble Cedex 9, France*

[b] *ISyE, Georgia Institute of Technology, 765 Ferst Drive, Atlanta GA 30332-0205 USA*

Abstract

We consider the problem of recovering of continuous multi-dimensional functions f from the noisy observations over the regular grid $m^{-1}\mathbb{Z}^d$, $m \in \mathbb{N}_*$. Our focus is at the adaptive estimation in the case when the function can be well recovered using a linear filter, which can depend on the unknown function itself. In the companion paper [26] we have shown in the case when there exists an adapted time-invariant filter, which locally recovers "well" the unknown signal, there is a numerically efficient construction of an adaptive filter which recovers the signals "almost as well". In the current paper we study the application of the proposed estimation techniques in the non-parametric regression setting. Namely, we propose an adaptive estimation procedure for "locally well-filtered" signals (some typical examples being smooth signals, modulated smooth signals and harmonic functions) and show that the rate of recovery of such signals in the ℓ_p-norm on the grid is essentially the same as that rate for regular signals with nonhomogeneous smoothness.

Key words: Nonparametric denoising, adaptive filtering, minimax estimation, nonparametric regression.

1 Introduction

Let $\mathbf{F} = (\Omega, \Sigma, P)$ be a probability space. We consider the problem of recovering unknown complex-valued random field $(s_\tau = s_\tau(\xi))_{\substack{\tau\in\mathbb{Z}^d \\ \xi\in\Omega}}$ over $\mathbb{Z}^d$ from

* Corresponding author.
Email addresses: `anatoli.juditsky@imag.fr` (Anatoli Juditsky), `nemirovs@isye.gatech.edu` (Arkadi Nemirovski).

noisy observations

$$y_\tau = s_\tau + e_\tau. \tag{1}$$

We assume that the field (e_τ) of observation noises is independent of (s_τ) and is of the form $e_\tau = \sigma\epsilon_\tau$, where (ϵ_τ) are independent of each other *standard* Gaussian complex-valued variables; the adjective "standard" means that $\Re(\epsilon_\tau)$, $\Im(\epsilon_\tau)$ are independent of each other $\mathbf{N}(0,1)$ random variables.

We suppose that the observations (1) come from a function ("signal") f of continuous argument (which we assume to vary in the d-dimensional unit cube $[0,1]^d$); this function is observed in noise along an n-point equidistant grid in $[0,1]^d$, and the problem is to recover f via these observations. This problem fits the framework of nonparametric regression estimation with a "traditional setting" as follows:

A. The objective is to recover an unknown smooth function $f : [0,1]^d \to \mathbb{R}$, which is sampled on the observation grid $\Gamma_n = \{x_\tau = m^{-1}\tau : 0 \le \tau_1, ..., \tau_d \le m\}$ with $(m+1)^d = n$, so that $s_\tau = f(x_\tau)$. The error of recovery is measured with some functional norm (or a semi-norm) $\|\cdot\|$ on $[0,1]^d$, and the risk of recovery $\widehat{f}$ of f is the expectation $E_f\|\widehat{f} - f\|^2$;

B. The estimation routines are aimed at recovering *smooth* signals, and their quality is measured by their maximal risks, the maximum being taken over f running through natural families of smooth signals, e.g., Hölder or Sobolev balls;

C. The focus is on the asymptotic, as the volume of observations n goes to infinity, behavior of the estimation routines, with emphasis on *asymptotically minimax (nearly) optimal estimates* – those which attain (nearly) best possible rates of convergence of the risks to 0 as the observation sample size $n \to \infty$.

Initially, the research was focused on recovering smooth signals with *a priori known smoothness parameters* and the estimation routines were tuned to these parameters (see, e.g., [23,34,38,24,2,31,39,22,36,21,27]). Later on, there was a significant research on *adaptive estimation*. Adaptive estimation methods are free of a priori assumptions on the smoothness parameters of the signal to be recovered, and the primary goal is to develop the routines which exhibit asymptotically (nearly) optimal behavior on a wide variety of families of smooth functions (cf. [35,28,29,30,6,8,9,25,3,7,19]). For a more compete overview of results on smooth nonparametric regression estimation see, for

instance, [33].[3]

The traditional focus on recovering smooth signals ultimately comes from the fact that such a signal *locally* can be well-approximated by a polynomial of a fixed order r, and such a polynomial is an "easy to estimate" entity. Specifically, for every integer $T \geq 0$, the value of a polynomial p at an observation point x_t can be recovered via $(2T+1)^d$ neighboring observations $\{x_\tau : |\tau_j - t_j| \leq T, 1 \leq j \leq d\}$ "at a parametric rate" – with the expected squared error $C\sigma^2(2T+1)^{-d}$ which is inverse proportional to the amount $(2T+1)^d$ of the observations used by the estimate. The coefficient C depends solely on the order r and the dimensionality d of the polynomial. The corresponding estimate $\widehat{p}(x_t)$ of $p(x_t)$ is pretty simple: it is given by a "time-invariant filter", that is, by convolution of observations with an appropriate discrete kernel $q^{(T)} = (q_\tau^{(T)})_{\tau\in\mathbb{Z}^d}$ vanishing outside the box $\mathbf{O}_T = \{\tau \in \mathbb{Z}^d : |\tau_j| \leq T, 1 \leq j \leq d\}$:

$$\widehat{p}(x_t) = \sum_{\tau\in\mathbf{O}_T} q_\tau^{(T)} y_{t-\tau},$$

then the estimation $\widehat{f}$ of $f(x_t)$ is taken as $\widehat{f} = \widehat{p}(x_t)$.

Note that the kernel $q^{(T)}$ is readily given by the degree r of the approximating polynomial, T and dimension d. The "classical" adaptation routines takes care of choosing "good" values of the approximation parameters (namely, T and r). On the other hand, the polynomial approximation "mechanism" is supposed to be fixed once for ever. Thus, in those procedures the "form" of the kernel is considered as given in advance.

In the companion paper [26] (referred hereafter as Part I) we have introduced the notion of a *well-filtered signal*. In brief, the signal $(s_\tau)_{\tau\in\mathbb{Z}^d}$ is T-well-filtered for some $T \in \mathbb{N}_+$ if there is a filter (kernel) $q = q^{(T)} \in \mathbf{O}_T$ which recovers (s_τ) in the box $\{u : |u-t| \leq 3T\}$ with the mean square error comparable with $\sigma T^{-d/2}$:

$$\max_{u:|u-t|\leq 3T} E\left\{|s_u - \sum_{\tau\in\mathbf{O}_T} q_\tau^{(T)} y_{u-\tau}|^2\right\} \leq O(\sigma^2 T^{-d}).$$

The universe of these signals is much wider than the one of smooth signals.As we have seen in Part I that it contains, in particular, "modulated smooth signals" – sums of a fixed number of products of smooth functions and multivariate harmonic oscillations of unknown (and arbitrarily high) frequencies.

[3] Our "brief outline" of adaptive approach to nonparametric regression would be severely incomplete without mentioning a novel approach aimed at recovering *non-smooth* signals possessing *sparse representations* in properly constructed functional systems [5,10,4,11,12,13,14,15,16,17,37,18]. This promising approach is completely beyond the scope of our paper.

We have shown in Part I that whenever a discrete time signal (that is, a signal defined on a regular discrete grid) is *well-filtered*, we can recover this signal at a "nearly parametric" rate *without a priori knowledge of the associated filter*. In other words, a well-filtered signal can be recovered *on the observation grid* basically as well as if it were an algebraic polynomial of a given order.

We are about to demonstrate that the results of Part I on recovering well-filtered signals of unknown structure can be applied to recovering nonparametric signals which admit well-filtered *local* approximations. Such an extension has an unavoidable price – now we cannot hope to recover the signal well outside of the observation grid (a highly oscillating signal can merely vanish on the observation grid and be arbitrarily large outside it). As a result, *in what follows we are interested in recovering the signals along the observation grid only* and, consequently, replace the error measures based on the functional norms on $[0,1]^d$ by their grid analogies.

The estimates to be developed will be "double adaptive", that is, adaptive with respect to both the unknown in advance structures of well-filtered approximations of our signals and to the unknown in advance "approximation rate" – the dependence between the size of a neighborhood of a point where the signal in question is approximated and the quality of approximation in this neighborhood. Note that in the case of smooth signals, this approximation rate is exactly what is given by the smoothness parameters. The results to follow can be seen as extensions of the results of [32,20] (see also [33]) dealing with the particular case of univariate signals satisfying differential inequalities with unknown differential operators.

2 Nonparametric regression problem

We start with the formal description of the components of the nonparametric regression problem.

Let for $\tau \in \mathbb{Z}^d$, $|\tau| = \max\{|\tau_1|, ..., |\tau_d|\}$, and let $\tau \leq m$ for some $a \in \mathbb{N}$ denote $\tau_i \leq m,\ i = 1, ..., d$. Let m be a positive integer, $n = (m+1)^d$, and let $\Gamma_n = \{x = m^{-1}\alpha : \alpha \in \mathbb{Z}^d, 0 \leq \alpha, |\alpha| \leq m\}$.

Let $C([0,1]^d)$ be the linear space of complex-valued fields over $[0,1]^d$. We associate with a signal $f \in C([0,1]^d)$ its observations along Γ_n:

$$y \equiv y_f^n(\epsilon) = \{y_\tau \equiv y_\tau^n(f,\epsilon) = f(m^{-1}\tau) + e_\tau, e_\tau = \sigma\epsilon_\tau\}_{0\leq\tau\leq m}, \tag{2}$$

where $\{\epsilon_\tau\}_{\tau\in\mathbb{Z}^d}$ are independent standard Gaussian complex-valued random noises. Our goal is to recover $f|_{\Gamma_n}$ from observations (2). In what follows, we

write

$$f_\tau = f(m^{-1}\tau), \qquad [\tau \in \mathbb{Z}^d, m^{-1}\tau \in [0,1]^d]$$

Below we use the following notations. For a set $B \subset [0,1]^d$, we denote by $\mathbb{Z}(B)$ the set of all $t \in \mathbb{Z}^d$ such that $m^{-1}t \in B$. We denote $\|\cdot\|_{q,B}$ the standard L_p-norm on B:

$$\|g\|_{p,B} = \left(\int\limits_{x \in B} |g(x)|^p dx \right)^{1/p},$$

and $|g|_{q,B}$ its discrete analogy, so that

$$|g|_{q,B} = m^{-d/q} \left(\sum_{\tau \in \mathbb{Z}(B)} |g_\tau|^q \right)^{1/q}.$$

We set

$$\Gamma_n^o = \Gamma_n \cap (0,1)^n = \{m^{-1}t : t \in \mathbb{Z}^d, t > 0, |t| < m\}.$$

Let $x = m^{-1}t \in \Gamma_n^o$. We say that a nonempty *open* cube

$$B_h(x) = \{u \mid \ |u_i - x_i| < h/2, \ i = 1, ..., d\}$$

centered at x is *admissible* for x, if $B_h(x) \subset [0,1]^n$. For such a cube, $T_h(x)$ denotes the largest nonnegative integer T such that

$$\mathbb{Z}(B) \supset \{\tau \in \mathbb{Z}^d : |\tau - t| \leq 4T\}.$$

For a cube

$$B = \{x \in \mathbb{R}^d : |x_i - c_i| \leq h/2, \, i = 1, ..., d\},$$

$D(B) = h$ stands for the edge of B. For $\gamma \in (0,1)$ we denote

$$B_\gamma = \{x \in \mathbb{R}^d : |x_i - c_i| \leq \gamma h/2, i = 1, ..., d\}$$

the γ-shrinkage of B to the center of B.

2.1 Classes of locally well-filtered signals

Recall that we say that a function on $[0,1]^d$ is smooth if it can be locally well-approximated by a polynomial. *Informally*, the the definition below sais that a continuous signal $f \in C([0,1])^d$ is *locally well-filtered* if f admits a good *local* approximation by a well-filtered discrete signal ϕ_τ on Γ_n (see Definition 1 of Section 2.1, Part I).

Definition 1 *Let $B \subset [0,1]^d$ be a cube, k be a positive integer, $\rho \geq 1$, $R \geq 0$ be reals, and let $p \in (d, \infty]$. The collection B, k, ρ, R, p specifies the family $\mathbf{F}^{k,\rho,p}(B,R)$ of locally well-filtered on B signals f defined by the following*

requirements:
(1) $f \in C([0,1]^d)$*;*
(2) There exists a nonnegative function $F \in L_p(B)$, $\|F\|_{p,B} \leq R$, *such that for every* $x = m^{-1}t \in \Gamma_n \cap \mathrm{int}B$ *and for every admissible for* x *cube* $B_h(x)$ *contained in* B *there exists a field* $\phi \in C(\mathbb{Z}^d)$ *such that* $\phi \in \mathbf{S}^t_{3T_h(x)}(0, \rho, T_h(x))$ *(where the set* $\mathbf{S}^t_L(\theta, \rho, T)$ *of* T*-well filtered signals is defined in Definition 1 of Part I) and*

$$\forall \tau \in \mathbb{Z}(B_h(x)) : |\phi_\tau - f_\tau| \leq h^{k-d/p}\|F\|_{p,B_h(x)}. \tag{3}$$

In the sequel, we use for $\mathbf{F}^{k,\rho,p}(B;R)$ also the shortened notation $\mathbf{F}[\psi]$, where ψ stands for the collection of "parameters" (k, ρ, p, B, R).

Remark The motivating example of locally well-filtered signals is that of modulated smooth signals as follows. Let a cube $B \subset [0,1]^d$, $p \in (d, \infty]$, positive integers k, ν and a real $R \geq 0$ be given. Consider a collection of ν functions $g_1, ..., g_\nu \in C([0,1]^d)$ which are k times continuously differentiable and satisfy the constraint

$$\sum_{\ell=1}^{\nu} \|D^k g_\ell\|_{p,B} \leq R.$$

Let $\omega(\ell) \in \mathbb{R}^d$, and let

$$f(x) = \sum_{\ell=1}^{\nu} g_\ell(x) \exp\{i\omega^T(\ell)x\}.$$

By the standard argument [1], whenever $x = m^{-1}t \in \Gamma_n \cap \mathrm{int}B$ and $B_h(x)$ is admissible for x, the Taylor polynomial $\Phi^x_\ell(\cdot)$, of order $k-1$, taken at x, of f_ℓ satisfies the inequality

$$u \in B_h(x) \Rightarrow |\Phi^x_\ell(u) - f_\ell(u)| \leq c_1 h^{k-d/p}\|F_\ell\|_{p,B_h(x)}, \quad \text{where} \quad F_\ell(u) = |D^k f_\ell(u)|$$

(here and in what follows, c_i are positive constants depending solely on d, k and ν). It follows that if $\Phi(u) = \sum\limits_{\ell=1}^{\nu} \Phi^x_\ell(u) \exp\{i\omega^T(\ell)u\}$ then

$$\begin{aligned} &u \in B_h(x) \Rightarrow |\Phi(u) - f(u)| \leq h^{k-d/p}\|F\|_{p,B_h(x)}, \\ &F = c_2 \sum_{\ell=1}^{\nu} F_\ell \qquad [\Rightarrow \|F\|_{p,B} \leq c_3 R]. \end{aligned} \tag{4}$$

Now observe that the exponential polynomial $\phi(\tau) = \Phi(m^{-1}\tau)$ belongs to $\mathbf{S}^t_L(0, c_4, T)$ for any $0 \leq T \leq L \leq \infty$ (Proposition 10 of Part I). Combining this fact with (4), we conclude that $f \in \mathbf{F}^{k,\rho(\nu,k,d),p}(B, c(\nu,k,d)R)$.

2.2 Accuracy measures

Let us fix $\gamma \in (0,1)$ and $q \in [1,\infty]$. Given an estimate $\widehat{f}_n$ of the restriction $f|_{\Gamma_n}$ of f on the grid Γ_n, based on observations (2) (i.e., a Borel real-valued function of $x \in \Gamma_n$ and $y \in \mathbb{C}^n$) and $\psi = (k,\rho,p,B,R)$, let us characterize the quality of the estimate on the set $\mathbf{F}[\psi]$ by the worst-case risks

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

3 Estimator construction

The recovering routine we are about to build is aimed at estimating functions from classes $\mathbf{F}^{k,\rho,p}(B,R)$ with *unknown in advance parameters* k,ρ,p,B,R. The only design parameters of the routine is an a priori upper bound μ on the parameter ρ and a $\gamma \in (0,1)$.

3.1 Preliminaries

From now on, we denote by $\Theta = \Theta_{(n)}$ the deterministic function of observation noises defined as follows. For every cube $B \subset [0,1]^d$ with vertices in Γ_n, we consider the discrete Fourier transform of the observation noises reduced to $B \cap \Gamma_n$, and take the maximum of modules of the resulting Fourier coefficients, let it be denoted $\theta_B(e)$. By definition,

$$\Theta \equiv \Theta_{(n)} = \sigma^{-1}\max_B \theta_B(e),$$

where the maximum is taken over all cubes B of the indicated type. By the origin of $\Theta_{(n)}$, due to the classical results on maxima of Gaussian processes (cf also Lemma 15 of Part I), we have

$$\forall w \geq 1: \quad \mathrm{Prob}\left\{\Theta_{(n)} > w\sqrt{\ln n}\right\} \leq \exp\left\{-\frac{cw^2\ln n}{2}\right\}, \tag{5}$$

where $c > 0$ depends solely on d.

3.2 Building blocks: window estimates

To recover a signal f via $n = m^d$ observations (2), we use point-wise window estimates of f defined as follows.

Let us fix a point $x = m^{-1}t \in \Gamma_n^o$; our goal is to build an estimate of $f(x)$. Let $B_h(x)$ be an admissible window for x. We associate with this window an estimate $\widehat{f}_n^h = \widehat{f}_n^h(x; y_f^n(\epsilon))$ of $f(x)$ defined as follows. If the window is "very small", specifically, $h \le m^{-1}$, so that x is the only point from the observation grid Γ_n in $B_h(x)$, we set $T_h(x) = 0$ and $\widehat{f}_n^h = y_t$. For a larger window, we choose the largest nonnegative integer $T = T_h(x)$ such that

$$\mathbb{Z}(B_h(x)) \supset \{\tau : |\tau - t| \le 4T\}$$

and apply Algorithm A of Part I to build the estimate of $f_t = f(x)$, the design parameters of the algorithm being $(\mu, T_h(x))$. Let the resulting estimate be denoted by $\widehat{f}_n^h = \widehat{f}_n^h(x; y_f^n(\epsilon))$.
To characterize the quality of the estimate $\widehat{f}_n^h = \widehat{f}_n^h(x; y_f^n(\epsilon))$, let us set

$$\Phi_\mu(f, B_h(x)) = \min_p \left\{ \max_{\tau \in \mathbb{Z}(B_h(x))} |p_\tau - f_\tau| : p \in \mathbf{S}^t_{3T_h(x)}(0, \mu, T_h(x)) \right\}.$$

Lemma 2 *One has*

$$(f_\tau) \in \mathbf{S}^t_{3T_h(x)}(\theta, \mu, T_h(x)), \quad \theta = \frac{\Phi_\mu(f, B_h(x))(1+\mu)}{(2T+1)^{d/2}}. \tag{6}$$

Assuming that $h > m^{-1}$ and combining (6) with the result of Theorem 4 of Part I we come to the following upper bound on the error of estimating $f(x)$ by the estimate $\widehat{f}_n^h(x; \cdot)$:

$$|f(x) - \widehat{f}_n^h(x; y_f(\epsilon))| \le C_1 \left[\Phi_\mu(f, B_h(x)) + \frac{\sigma}{\sqrt{nh^d}} \Theta_{(n)} \right] \tag{7}$$

(note that $(2T_h(x)+1)^{-d/2} \le C_0(nh^d)^{-1/2}$). For evident reasons (7) holds true for "very small windows" (those with $h \le m^{-1}$) as well.

3.3 The adaptive estimate

We are about to "aggregate" the window estimates $\widehat{f}_n^h$ into an *adaptive* estimate, applying Lepskii's adaptation scheme in the same fashion as in [30,19,20].

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,

$$\mathrm{Prob}\left\{\Theta_{(n)} > \omega\sqrt{\ln n}\right\} \leq n^{-4(\mu+1)}; \tag{8}$$

by (5), the required ω may be chosen as a function of μ, d only. We are to describe the basic blocks of the construction of the adaptive estimate.

"Good" realizations of noise. Let us define the set of "good realizations of noise" as

$$\Xi_n = \{\epsilon \mid \Theta_{(n)} \leq \omega\sqrt{\ln n}\}. \tag{9}$$

Now (7) implies the "conditional" error bound

$$\begin{aligned} &\epsilon \in \Xi_n \Rightarrow |f(x) - \widehat{f}_n^h(x; y_f(\epsilon))| \leq C_1 \left[\Phi_\mu(f, B_h(x)) + S_n(h)\right], \\ &S_n(h) = \frac{\sigma}{\sqrt{nh^d}}\omega\sqrt{\ln n}. \end{aligned} \tag{10}$$

Observe that *as h grows, the "deterministic term" $\Phi_\mu(f, B_h(x))$ does not decrease, while the "stochastic term" $S_n(h)$ decreases.*

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

$$\begin{aligned} B_*(x) &= B_{h_*(x)}(x), \\ h_*(x) &= \max\{h \mid h > 0, B_h(x) \subset [0,1]^d, \Phi_\mu(f, B_h(x)) \leq S_n(h)\}. \end{aligned} \tag{11}$$

Note that such a window does exist, since $S_n(h) \to \infty$ as $h \to +0$. Besides this, since the cubes $B_h(x)$ are open, the quantity $\Phi_\mu(f, B_h(x))$ is continuous from the left, so that

$$0 < h \leq h_*(x) \Rightarrow \Phi_\mu(f, B_h(x)) \leq S_n(h). \tag{12}$$

Thus, the ideal window $B_*(x)$ is well-defined for every x possessing admissible windows, i.e., for every $x = \Gamma_n^o = \{m^{-1}t : t \in \mathbb{Z}^d, 0 < t, |t| < m\}$.

Normal windows. Assume that $\epsilon \in \Xi_n$. 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:

$$\epsilon \in \Xi_n, 0 < h \leq h_*(x) \Rightarrow |f(x) - \widehat{f}_n^h(x; y_f(\epsilon))| \leq 2C_1 S_n(h) \tag{13}$$

(by (10) and (12)). Let us fix an $\epsilon \in \Xi_n$ (and thus – a realization y of the observations) and let us 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 window by no more than $4C_1$ times the stochastic term of the latter estimate, i.e.

$$\begin{array}{c}\text{Window } B_h(x) \text{ is normal}\\ \Updownarrow \\ \left\{\begin{array}{l} B_h(x) \text{ is admissible} \\ \forall h', 0 < h' \leq h: \quad |\widehat{f}_n^{h'}(x;y) - \widehat{f}_n^h(x;y)| \leq 4C_1 S_n(h') \quad [y = y_f(\epsilon)] \end{array}\right.\end{array} \tag{14}$$

Note that if $x \in \Gamma_n^o$, then x possesses a normal window, specifically, the window $B_{m^{-1}}(x)$. Indeed, this window contains a single observation point, namely, x itself, so that the corresponding estimate, same as every estimate corresponding to a smaller window, by construction coincides with the observation at x, so that all the estimates $\widehat{f}_n^{h'}(x;y)$, $0 < h' \leq m^{-1}$, are the same. Note also that (13) implies that

(!) *If $\epsilon \in \Xi_n$, then the ideal window $B_*(x)$ is normal.*

The adaptive estimate $\widehat{f}_n(x;y)$. 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 $B^+(x) = B_{h^+(x)}(x)$. *The adaptive estimate $\widehat{f}_n(x;y)$ is exactly the window estimate associated with the window $B^+(x)$.* Note that from (!) it follows that

(!!) *If $\epsilon \in \Xi_n$, 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

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

and we come to the conclusion as follows:
(*) *If $\epsilon \in \Xi_n$, 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* (10) *associated with the ideal window:*

$$\epsilon \in \Xi_n \Rightarrow |\widehat{f}_n(x;y) - f(x)| \leq 5C_1 \left[\Phi_\mu(f, B_{h_*(x)}(x)) + S_n(h_*(x))\right]. \tag{15}$$

Thus, the estimate $\widehat{f}_n(\cdot;\cdot)$ – which is based solely on observations and does not require any a priori knowledge of the "parameters of well-filterability 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": $\epsilon \in \Xi_n$).
Note that the adaptive estimate $\widehat{f}_n(x;y)$ we have built depends solely on "design parameters" μ, γ (recall that C_1 depends on μ, γ), the volume of observations n and the dimension d.

4 Main result

Our main result is as follows:

Theorem 3 *Let $\gamma \in (0,1)$, $\mu \geq 1$ be an integer, let $\mathbf{F} = \mathbf{F}^{k,\rho,p}(B;R)$ be a family of locally well-filtered signals associated with a cube $B \subset [0,1]^d$ with $mD(B) \geq 1$, $\rho \leq \mu$ and $p > d$. For properly chosen $P \geq 1$ depending solely on μ, d, p, γ and nonincreasing in $p > d$ the following statement holds true:*

Suppose that the volume $n = m^d$ of observations (2) is large enough, namely,

$$P^{-1} n^{\frac{2kp+d(p-2)}{2dp}} \geq \frac{R}{\sigma}\sqrt{\frac{n}{\ln n}} \geq P[D(B)]^{-\frac{2kp+d(p-2)}{2p}} \tag{16}$$

where $D(B)$ is the edge of the cube B.

Then for every $q \in [1,\infty]$ the worst case, with respect to $\mathbf{F}$, q-risk of the adaptive estimate $\widehat{f}_n(\cdot,\cdot)$ associated with the parameter μ can be bounded as follows:

$$\begin{aligned}\widehat{\mathbf{R}}_q\left(\widehat{f}_n;\mathbf{F}\right) &= \sup_{f\in\mathbf{F}}\left(E\left\{\left|\widehat{f}_n(\cdot;y_f(\epsilon)) - f(\cdot)\right|^2_{q,B_\gamma}\right\}\right)^{1/2} \\ &\leq PR\left(\frac{\sigma^2\ln n}{R^2 n}\right)^{\beta(p,k,d,q)}[D(B)]^{d\lambda(p,k,d,q)},\end{aligned} \tag{17}$$

where

$$\beta(p,k,d,q) = \begin{cases} \frac{k}{2k+d}, & \text{when } q \leq \frac{(2k+d)p}{d}, \\ \frac{k+d\left(\frac{1}{q}-\frac{1}{p}\right)}{2k+d-\frac{2d}{p}}, & \text{when } q > \frac{(2k+d)p}{d}, \end{cases}$$

$$\lambda(p,k,d,q) = \begin{cases} \frac{1}{q} - \frac{d}{(2k+d)p}, & \text{when } q \leq \frac{(2k+d)p}{d}, \\ 0, & \text{when } q > \frac{(2k+d)p}{d} \end{cases}$$

(recall that here B_γ is the concentric to B γ times smaller cube).

Note that the rates of convergence to 0, as $n \to \infty$, of the risks $\widehat{\mathbf{R}}_q\left(\widehat{f}_n; \mathbf{F}\right)$ of our adaptive estimate on the families $\mathbf{F} = \mathbf{F}^{k,\rho,p}(B; R)$ are exactly the same as those stated by Theorem 3 from [31] (see also [30,9,19,33]) in the case of recovering non-parametric *smooth* regression functions from Sobolev balls. It is well-known that in the smooth case the latter rates are optimal in order, up to logarithmic in n factors. Since the families of locally well-filtered signals are much wider than local Sobolev balls (smooth signals are trivial examples of modulated smooth signals!), it follows that the rates of convergence stated by Theorem 3 also are nearly optimal.

5 Simulation examples

In this section we present the results of a small simulation study of the adaptive filtering algorithm applied to the 2-dimensional de-noising problem. The simulation setting is as follows: we consider real-valued signals

$$y_\tau = s_\tau + e_\tau, \quad \tau = (\tau_1, \tau_2) \in \{1, ..., m\}^2,$$

$e_{(1,1)}, ..., e_{(m,m)}$ being independent standard Gaussian random variables. The problem is to estimate, given observations (y_τ), the values of the signal (f_{x_τ}) on the grid $\Gamma_m = \{m^{-1}\tau,\ 1 \le \tau_1, \tau_2 \le m\}$, and $f(x_\tau) = s_\tau$. The value $m = 128$ is common to all experiments.

We consider signals which are sums of three harmonic components:

$$s_\tau = \alpha[\sin(m^{-1}\omega_1^T\tau + \theta_1) + \sin(m^{-1}\omega_2^T\tau + \theta_2) + \sin(m^{-1}\omega_3^T\tau + \theta_3)];$$

the frequencies ω_i and the phase shifts θ_i, $i = 1, ..., 3$ are drawn randomly from the uniform distribution over, respectively, $[0, \omega_{\max}]^3$ and $[0, 1]^3$ and the coefficient α is chosen to obtain the signal-to-noise ratio equal to one.

In the simulations we present here we compared the result of adaptive recovery with $T = 10$ to that of a "standard nonparametric recovery", i.e. the recovery by the locally linear estimator with square window. We have done $k = 100$ independent runs for each of eight values of $\omega_{\max}$,

$$\omega_{\max} = \{1.0,\ 2.0,\ 4.0,\ 8.0\, 16.0,\ 32.0,\ 64.0,\ 128.0\}.$$

In Table 1 we summarize the results for the mean integrated squared error (MISE) of the estimation,

$$MISE = \sqrt{\frac{1}{100m^2}\sum_{j=1}^{100}\sum_{\tau=(1,1)}^{(m,m)}(\widehat{s}_\tau^{(j)} - s_\tau^{(j)})^2}.$$

The observed phenomenon is rather expectable: for slowly oscillating signals the quality of the adaptive recovery is slightly worse than that of "standard recovery", which are tuned for estimation of regular signals. When we rise the frequency of the signal components, the adaptive recovery stably outperforms the standard recovery. Finally, standard recovery is clearly unable to recover highly oscillating signals (cf Figures 1-4).

Appendix

We denote $C(\mathbb{Z}^d)$ the linear space of complex-valued fields over $\mathbb{Z}^d$. A field $r \in C(\mathbb{Z}^d)$ with finitely many nonzero entries r_τ is called *a filter*. We use the commun notation Δ_j, $j = 1, ..., d$, for the "basic shift operators" on $C(\mathbb{Z}^d)$:

$$(\Delta_j r)_{\tau_1,...,\tau_d} = r_{\tau_1,...,\tau_{j-1},\tau_j-1,\tau_{j+1},...,\tau_d}.$$

and denote $r(\Delta)x$ the output of a filter r, the input to the filter being a field $x \in C(\mathbb{Z}^d)$, so that $(r(\Delta)x)_t = \sum_\tau r_\tau x_{t-\tau}$.

5.1 Proof of Lemma 2.

To save notation, let $B = B_h(x)$ and $T = T_h(x)$. Let $p \in C(\mathbb{Z}^d)$ be such that $p \in \mathbf{S}^t_{3T}(0, \mu, T)$ and $|p_\tau - f_\tau| \le \Phi_\mu(f, B_h(x))$ for all $\tau \in \mathbb{Z}(B_h(x))$. Since $p \in \mathbf{S}^t_{3T}(0, \mu, T)$, there exists a filter $q \in C_T(\mathbb{Z}^d)$ such that $|q|_2 \le \mu(2T+1)^{-d/2}$ and $(q(\Delta)p)_\tau = p_\tau$ whenever $|\tau - t| \le 3T$. Setting $\delta_\tau = f_\tau - p_\tau$, we have for any τ , $|\tau - t| \le 3T$,

$$\begin{aligned}
&|f_\tau - (q(\Delta)f)_\tau| \le |\delta_\tau| + |p_\tau - (q(\Delta)p)_\tau| + |(q(\Delta)\delta)_\tau| \\
&\le \Phi_\mu(f, B_h(x)) + |q|_1 \max\{|\delta_\nu| : |\nu - \tau| \le T\} \le \Phi_\mu(f, B_h(x)) \\
&\quad + |q|_2(2T+1)^{d/2}\Phi_\mu(f, B_h(x)) \max\{|\delta_\nu| : |\nu - \tau| \le T\} \\
&\quad [\text{note that } |\tau - t| \le 3T \text{ and } |\nu - \tau| \le T \text{ implies } |\nu - t| \le 4T] \\
&\le \Phi_\mu(f, B_h(x))(1+\mu) = \frac{\Phi_\mu(f, B_h(x))(1+\mu)}{(2T+1)^{d/2}}(2T+1)^{-d/2}
\end{aligned}$$

as required in (6). $\square$

5.2 Proof of Theorem 3

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 5^0.
Let us fix a family of well-filtered signals $\mathbf{F} = \mathbf{F}_d^{k,\rho,p}(B; R)$ with the parameters satisfying the premise of Theorem 3 and a function f from this class.
Recall that by the definition of $\mathbf{F}$ there exists a function $F \geq 0$, $\|F\|_{p,B} \leq R$, such that for all $x = m^{-1}t \in (\mathrm{int}B) \cap \Gamma_n$ and all $h,\ B_h(x) \subset B$:

$$\Phi_\mu(f, B_h(x)) \leq P_1 h^{k-d/p} \Omega(f, B_h(x)),\ \Omega(f, B') = \left(\int\limits_{B'} F^p(u) du \right)^{1/p}. \quad (18)$$

From now on, P (perhaps with sub- or superscripts) are quantities ≥ 1 depending on μ, d, γ, p only and nonincreasing in $p > d$.

1^0. We need the following auxiliary result:

Lemma 4 *Assume that*

$$n^{\frac{k-d/p}{d}} \sqrt{\ln n} \geq P_1 (\mu + 3)^{k-d/p+d/2} \frac{R}{\sigma\omega}. \quad (19)$$

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

$$\begin{aligned} &(a) : h \leq (1-\gamma) D(B), \\ &(b) : P_1 h^{k-d/p} \Omega(f, B_h(x)) \leq S_n(h). \end{aligned} \quad (20)$$

Then $h(x)$ is well-defined and

$$h(x) \geq m^{-1}. \quad (21)$$

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

$$|\widehat{f}_n(x; y) - f(x)| \leq C_2 \left[S_n(h(x)) \mathbf{1}\{\epsilon \in \Xi_n\} + \sigma \Theta_{(n)} \mathbf{1}\{\epsilon \notin \Xi_n\} \right] \quad (22)$$

Proof: The quantity $h(x)$ is well-defined, since for small positive h the left hand side in (20.b) is close to 0, while the right hand side one is large. From

(19) it follows that $h = m^{-1}$ satisfies (20.a), so that $B_{m^{-1}}(x) \subset B$. Moreover, (19.b) implies that

$$P_1 m^{-k+d/p} R \le S_n(m^{-1});$$

the latter inequality, in view of $\Omega(f, B_{m^{-1}}(x)) \le R$, says that $h = m^{-1}$ satisfies (20.b) as well. Thus, $h(x) \ge m^{-1}$, as claimed in (21).

Consider the window $B_{h(x)}(x)$. By (20.a) it is admissible for x, while from (20.b) combined with (18) we get $\Phi_\mu(f, B_{h(x)}(x)) \le S_n(h)$. It follows that the ideal window $B_*(x)$ of x is not smaller than $B_{h(x)}(x)$.

Assume that $\epsilon \in \Xi_n$. Then, according to (15), we have

$$|\widehat{f}_n(x;y) - f(x)| \le 5C_1 \left[\Phi_\mu(f, B_{h_*(x)}(x)) + S_n(h_*(x))\right]. \tag{23}$$

Now, by the definition of an ideal window, $\Phi_\mu(f, B_{h_*(x)}(x)) \le S_n(h_*(x))$, and the right hand side in (23) does not exceed $10C_1 S_n(h_*(x)) \le 10C_1 S_n(h(x))$ (recall that, as we have seen, $h_*(x) \ge h(x)$), as required in (22).

Now let $\epsilon \notin \Xi_n$. 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 $B_h(x)$ which is normal and such that $h \ge m^{-1}$ (the latter – since the window $B_{m^{-1}}(x)$ always is normal, and $B_h(x)$ is the largest normal window centered at x). Applying (14) with $h' = m^{-1}$ (so that $\widehat{f}_n^{h'}(x;y) = f(x) + \sigma\epsilon_t$), we get $|(f(x) + \sigma\epsilon_t) - \widehat{f}_n(x;y)| \le 4C_1 S_n(m^{-1})$, whence

$$|f(x) - \widehat{f}_n(x;y)| \le \sigma|\epsilon_t| + 4C_1 S_n(m^{-1}) \le \sigma\Theta_{(n)} + 4C_1\sigma\omega\sqrt{\ln n} \le C_2\Theta_{(n)}$$

(recall that we are in the situation $\epsilon \notin \Xi_n$, whence $\omega\sqrt{\ln n} \le \Theta_{(n)}$). We have arrived at (22). $\square$

Now we are ready to complete the proof. Assume that (19) takes place, and let us fix q, $\frac{2k+d}{d}p \le q < \infty$.

2^0. Let us denote $\widehat{\sigma}_n = \sigma\sqrt{\frac{\ln n}{n}}$. Note that for every $x \in \Gamma_n \cap B_\gamma$ either

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

or

$$h(x) = \left(\frac{\widehat{\sigma}_n}{P_1\Omega(f, B_{h(x)}(x))}\right)^{\frac{2p}{2kp+(p-2)d}},$$

what means that

$$P_1 h^{k-d/p}(x)\Omega(f, B_{h(x)}(x)) = S_n(h(x)). \tag{24}$$

Let U, V be the sets of those $x \in B_\gamma^n \equiv \Gamma_n \cap B_\gamma$ for which the first or, respectively, the second of this possibilities takes place. If V is nonempty, let

us partition it as follows.
1) We can choose $x_1 \in V$ (V is finite!) such that $h(x) \geq h(x_1) \quad \forall x \in V$. 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 $h(x) \geq h(x_2) \quad \forall x \in V \backslash V_1$, 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 (since V is finite). 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:

(1) The cubes $B_{h(x_1)}(x_1), ..., B_{h(x_M)}(x_M)$ are mutually disjoint;
(2) For every $\ell \leq M$ and every $x \in V_\ell$ we have $h(x) \geq h(x_\ell)$ and $B_{h(x)}(x) \cap B_{h(x_\ell)}(x_\ell) \neq \emptyset$.
We claim that also
(3) For every $\ell \leq M$ and every $x \in V_\ell$ one has

$$h(x) \geq \max\left[h(x_\ell); \|x - x_\ell\|_\infty\right]. \tag{25}$$

Indeed, $h(x) \geq h(x_\ell)$ by (ii), so that it suffices to verify (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 what we need.
3^0. Let us set $B^n_\gamma = \Gamma_n \cap B_\gamma$. Assume that $\epsilon \in \Xi_n$. When substituting $h(x) = (1-\gamma)[D(B)]$ for $x \in U$, we have by (22):

$$\begin{aligned}
\left|\widehat{f}_n(\cdot; y) - f(\cdot)\right|^q_{q,B_\gamma} &\leq C_2^q m^{-\frac{d}{q}} \sum_{x \in B^n_\gamma} S^q_n(h(x)) \\
&= C_2^q m^{-\frac{d}{q}} \sum_{x \in U} S^q_n(h(x)) + C_2^q m^{-\frac{d}{q}} \sum_{\ell=1}^{M} \sum_{x \in V_\ell} S^q_n(h(x)) \\
&= C_2^q m^{-\frac{d}{q}} \sum_{x \in U} \left[\frac{\widehat{\sigma}_n}{((1-\gamma)D(B))^{d/2}}\right]^q + C_2^q m^{-\frac{d}{q}} \sum_{\ell=1}^{M} \sum_{x \in V_\ell} S^q_n(h(x)) \\
[\text{by (25)}] \quad &\leq C_3^q \widehat{\sigma}_n^q m^{-\frac{d}{q}} \sum_{\ell=1}^{M} \sum_{x \in V_\ell} \left(\max\left[h(x_\ell), \|x - x_\ell\|_\infty\right]\right)^{-\frac{dq}{2}} + C_3^q \widehat{\sigma}_n^q [D(B)]^{\frac{d(2-q)}{2}} \\
&\leq C_4^q \widehat{\sigma}_n^q \sum_{\ell=1}^{M} \int \left(\max\left[h(x_\ell), \|x - x_\ell\|_\infty\right]\right)^{-\frac{dq}{2}} dx + C_3^q \widehat{\sigma}_n^q [D(B)]^{\frac{d(2-q)}{2}}
\end{aligned}$$

$$\leq C_5^q \widehat{\sigma}_n^q \sum_{\ell=1}^{M} \int_0^\infty r^{d-1} \left(\max\left[h(x_\ell), r\right]\right)^{-\frac{dq}{2}} dr + C_3^q \widehat{\sigma}_n^q D[D(B)]^{\frac{d(2-q)}{2}},$$

due to $h(x) \geq m^{-1}$, see (21). Further, note that

$$\frac{dq}{2} - d + 1 \geq \frac{2k+d}{2} p - d + 1 \geq d^2/2 + 1$$

in view of $q \geq \frac{2k+d}{d} p$, $k \geq 1$ and $p > d$, and

$$\begin{aligned}
\left|\widehat{f}_n(\cdot; y) - f(\cdot)\right|_{q,B_\gamma}^q &\leq C_6^q \widehat{\sigma}_n^q \sum_{\ell=1}^{M} \left[h(x_\ell)\right]^{\frac{d(2-q)}{2}} + C_3^q \widehat{\sigma}_n^q [D(B)]^{\frac{d(2-q)}{2}} \\
[\text{by (24)}] \; &= C_6^q \widehat{\sigma}_n^q \sum_{\ell=1}^{M} \left[\frac{\widehat{\sigma}_n}{P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))}\right]^{\frac{d(2-q)}{2k-2d/p+d}} + C_3^q \widehat{\sigma}_n^q [D(B)]^{\frac{d(2-q)}{2}} \\
&= C_3^q \widehat{\sigma}_n^q [D(B)]^{\frac{d(2-q)}{2}} + C_6^q \widehat{\sigma}_n^{2\beta(p,k,d,q)q} \sum_{\ell=1}^{M} \left[P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))\right]^{\frac{d(q-2)}{2k-2d/p+d}}
\end{aligned}$$

by definition of $\beta(p, k, d, q)$.
Now note that $\frac{d(q-2)}{2k-2d/p+d} \geq p$ in view of $q \geq \frac{2k+d}{d} p$, so that

$$\begin{aligned}
\sum_{\ell=1}^{M} \left[P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))\right]^{\frac{d(q-2)}{2k-2d/p+d}} &\leq \left[\sum_{\ell=1}^{M} \left(P_1 \Omega(f, B_{h(x_\ell)}(x_\ell))\right)^p\right]^{\frac{dq-2d}{p(2k-2d/p+d)}} \\
&\leq \left[P_1^p R^p\right]^{\frac{d(q-2)}{p(2k-2d/p+d)}}
\end{aligned}$$

(see (18) and take into account that the cubes $B_{h(x_\ell)}(x_\ell)$, $\ell = 1, ..., M$, are mutually disjoint by (i)). We conclude that for $\epsilon \in \Xi_n$

$$\begin{aligned}
\left|\widehat{f}_n(\cdot; y_f(\epsilon)) - f(\cdot)\right|_{q,B_\gamma} &\leq C_7 \widehat{\sigma}_n [D(B)]^{\frac{d(2/q-1)}{2}} + P_2 \widehat{\sigma}_n^{2\beta(p,k,d,q)} R^{\frac{d(1-2/q)}{2k-2d/p+d}} \\
&= C_7 \widehat{\sigma}_n [D(B)]^{\frac{d(2/q-1)}{2}} + P_2 R \left(\frac{\widehat{\sigma}_n}{R}\right)^{2\beta(p,k,d,q)}. \qquad (26)
\end{aligned}$$

4^0. Now assume that $\epsilon \notin \Xi_n$. In this case, by (22),

$$|\widehat{f}_n(x; y) - f(x)| \leq C_2 \sigma \Theta_{(n)} \quad \forall x \in B_\gamma^n.$$

Hence, taking into account that $mD(B) \geq 1$,

$$\left|\widehat{f}_n(\cdot; y) - f(\cdot)\right|_{q,B_\gamma} \leq C_2 \sigma \Theta_{(n)} [D(B)]^{\frac{d}{q}}. \qquad (27)$$

5^0. When combining (26) and (27), we get

$$\left(E\left\{\|\widehat{f}_n(\cdot;y)-f(\cdot)\|^2_{q,B_\gamma}\right\}\right)^{1/2} \le C_8 \max\left[\widehat{\sigma}_n[D(B)]^{\frac{d(2/q-1)}{2}};\ P_4R\left(\frac{\widehat{\sigma}_n}{R}\right)^{2\beta(p,k,d,q)};\ J\right],$$

where

$$\begin{aligned} J^2 &= E\left\{\mathbf{1}\{\epsilon\notin\Xi_n\}C_2\sigma^2\Theta^2_{(n)}\right\} \le C_2^2\sigma^2P^{1/2}\{\epsilon\notin\Xi_n\}\left(E\left\{\Theta^4_{(n)}\right\}\right)^{1/2}\\ &\le C_9\sigma^2n^{-2(\mu+1)}\ln n \end{aligned}$$

(we have used (5) and (8)). Thus, when (19) holds, for all $d<p<\infty$ and all q, $\frac{2k+d}{d}p\le q<\infty$ we have

$$\begin{aligned} &\left(E\left\{\|\widehat{f}_n(\cdot;y)-f(\cdot)\|^2_{q,B_\gamma}\right\}\right)^{1/2}\\ &\le C_8\max\left[\widehat{\sigma}_n[D(B)]^{\frac{d(2/q-1)}{2}}, P_4R\left(\frac{\widehat{\sigma}_n}{R}\right)^{2\beta(p,k,d,q)}, \frac{C_9^{1/2}\sigma\sqrt{\ln n}}{n^{(\mu+1)}}\right]. \end{aligned} \tag{28}$$

Now it is easily seen that if $P\ge 1$ is a properly chosen function of μ,d,γ,p nonincreasing in $p>d$ and (16) takes place then

(1) assumption (19) holds,
(2) the right hand side in (28) does not exceed the quantity

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

(recall that $q\ge\frac{2k+d}{d}p$, so that $\lambda(p,k,d,q)=0$).

We conclude the bound (17) for the case of $d<p<\infty$, $\infty>q\ge\frac{2k+d}{d}p$. When passing to the limit as $q\to\infty$, we get the desired bound for $q=\infty$ as well.

Now let $d<p<\infty$ and $1\le q\le q_*\equiv\frac{2k+d}{d}p$. By the Hölder inequality and in view of $mD(B)\ge 1$ we have

$$|g|_{q,B_\gamma}\le C_{10}\,|g|_{q_*,B_\gamma}\,|B_\gamma|^{\frac{1}{q}-\frac{1}{q_*}},$$

and thus

$$\widehat{\mathbf{R}}_q\left(\widehat{f}_n;\mathbf{F}\right)\le C_{10}\widehat{\mathbf{R}}_{q_*}\left(\widehat{f}_n;\mathbf{F}\right)[D(B)]^{d\left(\frac{1}{q}-\frac{1}{q_*}\right)}.$$

Combining this observation with the (already proved) bound (17) associated with $q=q_*$, we see that (17) 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 the validity of (17) for all $p\in(d,\infty]$, $q\in[1,\infty]$. $\square$

References

[1] O.V. Besov, V.P. Il'in, and S.M. Nikol'ski. *Integral representations of functions and embedding theorems.* Moscow: Nauka Publishers, 1975 (in Russian).

[2] L. Birgé. Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie verw. Geb.*, 65:181–237, 1983.

[3] L. Birgé, P. Massart. From model selection to adaptive estimation. In: D. Pollard, E. Torgersen and G. Yang, Eds., *Festschrift for Lucien Le Cam*, Springer 1999, 55–89.

[4] E. Candés, D. Donoho. Ridgelets: a key to high-dimensional intermittency? *Philos. Trans. Roy. Soc. London Ser. A* 357:2495-2509, 1999.

[5] S. Chen, D.L. Donoho, M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* 20(1):33-61, 1998.

[6] D. Donoho, I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika* 81(3):425-455, 1994.

[7] D. Donoho, I. Johnstone. Minimax risk over ℓ_p balls for ℓ_q losses. *Probab. Theory Related Fields* 99:277-303, 1994.

[8] D. Donoho, I. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* 90(432):1200–1224, 1995.

[9] D. Donoho, I. Johnstone, G. Kerkyacharian, D. Picard. Wavelet shrinkage: Asymptopia? (with discussion and reply by the authors). *J. Royal Statist. Soc. Series B* 57(2):301–369, 1995.

[10] D. Donoho. Tight frames of k-plane ridgelets and the problem of representing objects that are smooth away from d-dimensional singularities in $\mathbb{R}^n$. *Proc. Natl. Acad. Sci. USA* 96(5):1828-1833, 1999.

[11] D. Donoho. Wedgelets: nearly minimax estimation of edges. *Ann. Statist.* 27:859-897, 1999.

[12] D. Donoho. Orthonormal ridgelets and linear singularities. *SIAM J. Math. Anal.* 31:1062-1099, 2000.

[13] D. Donoho. Ridge functions and orthonormal ridgelets. *J. Approx. Theory* 111(2):143-179, 2001.

[14] D. Donoho. Curvelets and curvilinear integrals. *J. Approx. Theory* 113(1):59-90, 2001.

[15] D. Donoho. Sparse components of images and optimal atomic decompositions. *Constr. Approx.* 17:353-382, 2001.

[16] D. Donoho, X. Huo. Uncertainty principle and ideal atomic decomposition. *IEEE Trans. on Information Theory* 47(7):2845-2862, 2001.

[17] D. Donoho, X. Huo. Beamlets and multiscale image analysis. *Lect. Comput. Sci. Eng.* 20:149-196, Springer, 2002.

[18] M. Elad, A. Bruckstein. A generalized uncertainty principle and sparse representation in pairs of bases. *IEEE Trans. on Information Theory* (to appear)

[19] A. Goldenshluger, A. Nemirovski. On spatially adaptive estimation of nonparametric regression. *Math. Methods of Statistics* 6(2):135–170, 1997.

[20] A. Goldenshluger, A. Nemirovski. Adaptive de-noising of signals satisfying differential inequalities. *IEEE Trans. on Information Theory* 43, 1997.

[21] Yu. Golubev. Asymptotic minimax estimation of regression function in additive model. *Problemy peredachi informatsii* 28(2):3–15, 1992. (English transl. in *Problems Inform. Transmission* 28, 1992.)

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

[23] I. Ibragimov and R. Khasminskii. On nonparametric estimation of regression. *Soviet Math. Dokl.* 21:810–814, 1980.

[24] I. Ibragimov and R. Khasminskii. *Statistical Estimation.* Springer-Verlag, New York, 1981.

[25] A. Juditsky. Wavelet estimators: Adapting to unknown smoothness. *Math. Methods of Statistics* 6(1):1–25, 1997.

[26] A. Juditsky and A. Nemirovski. Nonparametric Denoising of Signals with Unknown Local Structure, I: Oracle Inequalities Accepted to *Appl. Comp. Harm. Anal.*

[27] A. Korostelev, A. Tsybakov. *Minimax theory of image reconstruction.* Lecture Notes in Statistics 82, Springer, New York, 1993.

[28] O. Lepskii. On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability and Its Applications* 35(3):454–466, 1990.

[29] O. Lepskii. Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory of Probability and Its Applications*, 36(4):682–697, 1991.

[30] O. Lepskii, E. Mammen, V. Spokoiny. Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* 25(3):929–947, 1997.

[31] A. Nemirovskii. On nonparametric estimation of smooth regression functions. *Sov. J. Comput. Syst. Sci.*, 23(6):1–11, 1985.

[32] A. Nemirovski. On nonparametric estimation of functions satisfying differential inequalities. R. Khasminski, Ed. *Advances in Soviet Mathematics* 12:7–43, American Mathematical Society, 1992.

[33] A. Nemirovski. *Topics in Non-parametric Statistics.* In: M. Emery, A. Nemirovski, D. Voiculescu, *Lectures on Probability Theory and Statistics, Ecole d'Eteé de Probabilités de Saint-Flour XXVII – 1998, Ed. P. Bernard.* – Lecture Notes in Mathematics 1738:87–285.

[34] M. Pinsker. Optimal filtration of square-integrable signals in Gaussian noise. *Problemy peredachi informatsii*, 16(2):120–133. 1980. (English transl. in *Problems Inform. Transmission* 16, 1980.)

[35] M. Pinsker, S. Efroimovich. Learning algorithm for nonparametric filtering. *Automation and Remote Control* 45(11):1434–1440, 1984.

[36] M. Rosenblatt. *Stochastic curve estimation.* Institute of Mathematical Statistics, Hayward, California, 1991.

[37] J.-L. Starck, E. Candés, D. Donoho. The curvelet transform for image denoising. *IEEE Trans. Image Process.* 11(6):670-684, 2002.

[38] Ch. Stone. Optimal rates of convergence for nonparametric estimators. *Annals of Statistics*, 8(3):1348–1360, 1980.

[39] G. Wahba. *Spline models for observational data.* SIAM, Philadelphia, 1990.

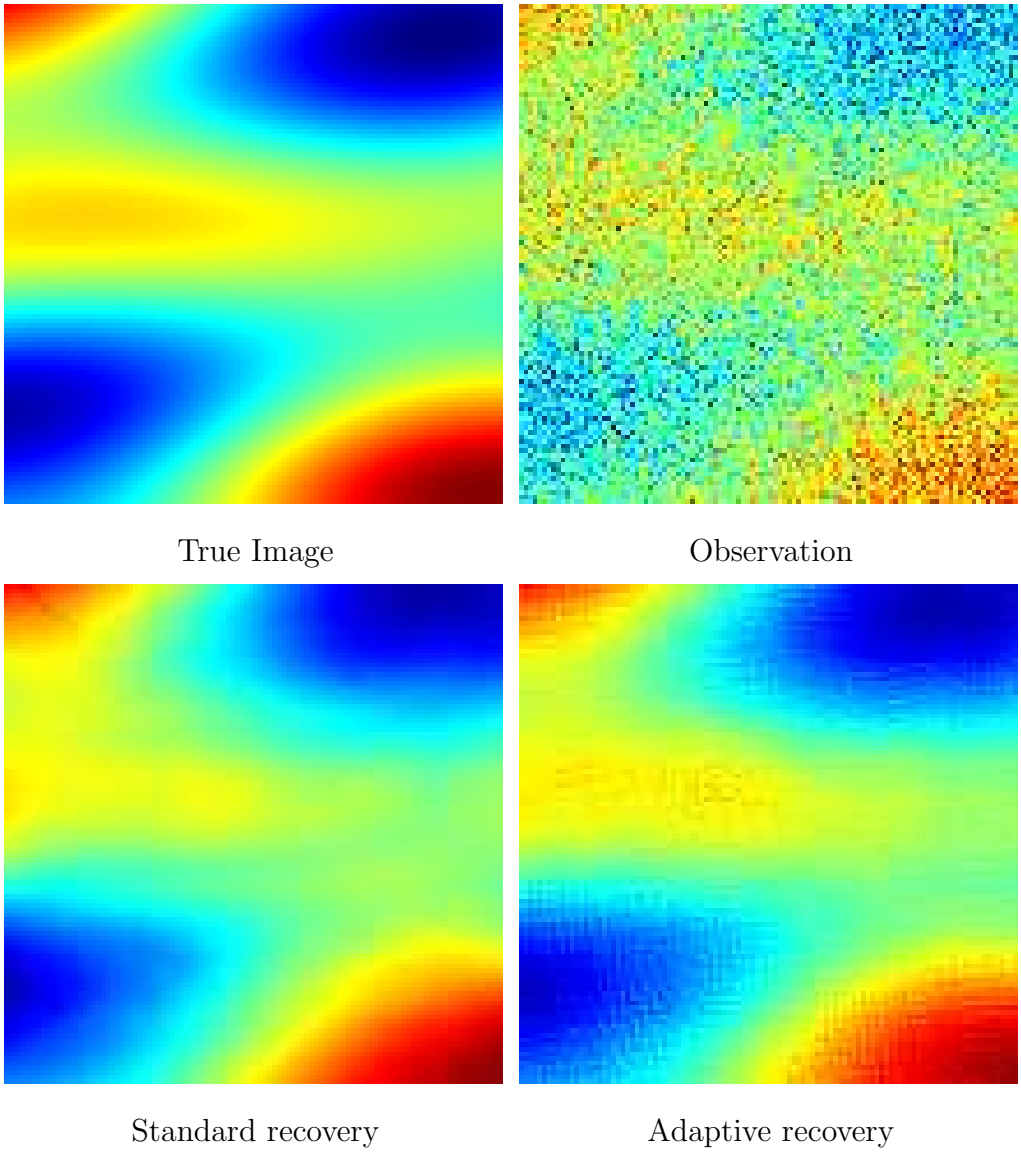

Fig. 1. Recovery for $\omega_{\max} = 2.0$

Table 1
MISE of adaptive recovery

$\omega_{\max}$	Standard recovery	Adaptive recovery
1.0	0.12	0.1
2.0	0.20	0.12
4.0	0.36	0.18
8.0	0.54	0.27
16.0	0.79	0.25
32.0	0.75	0.29
64.0	0.27	0.98
128.0	0.24	1.00

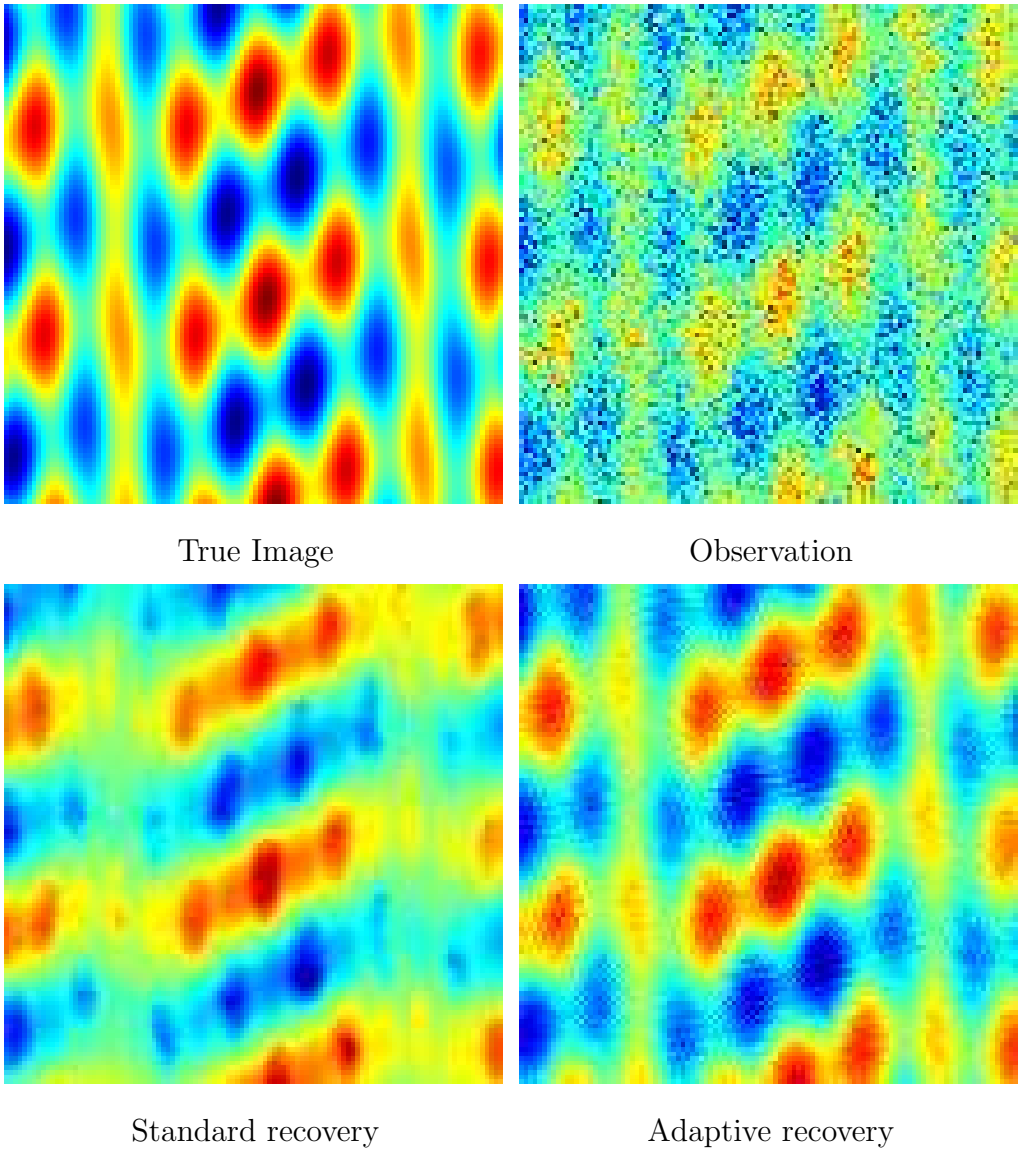

Fig. 2. Recovery for $\omega_{\max} = 8.0$

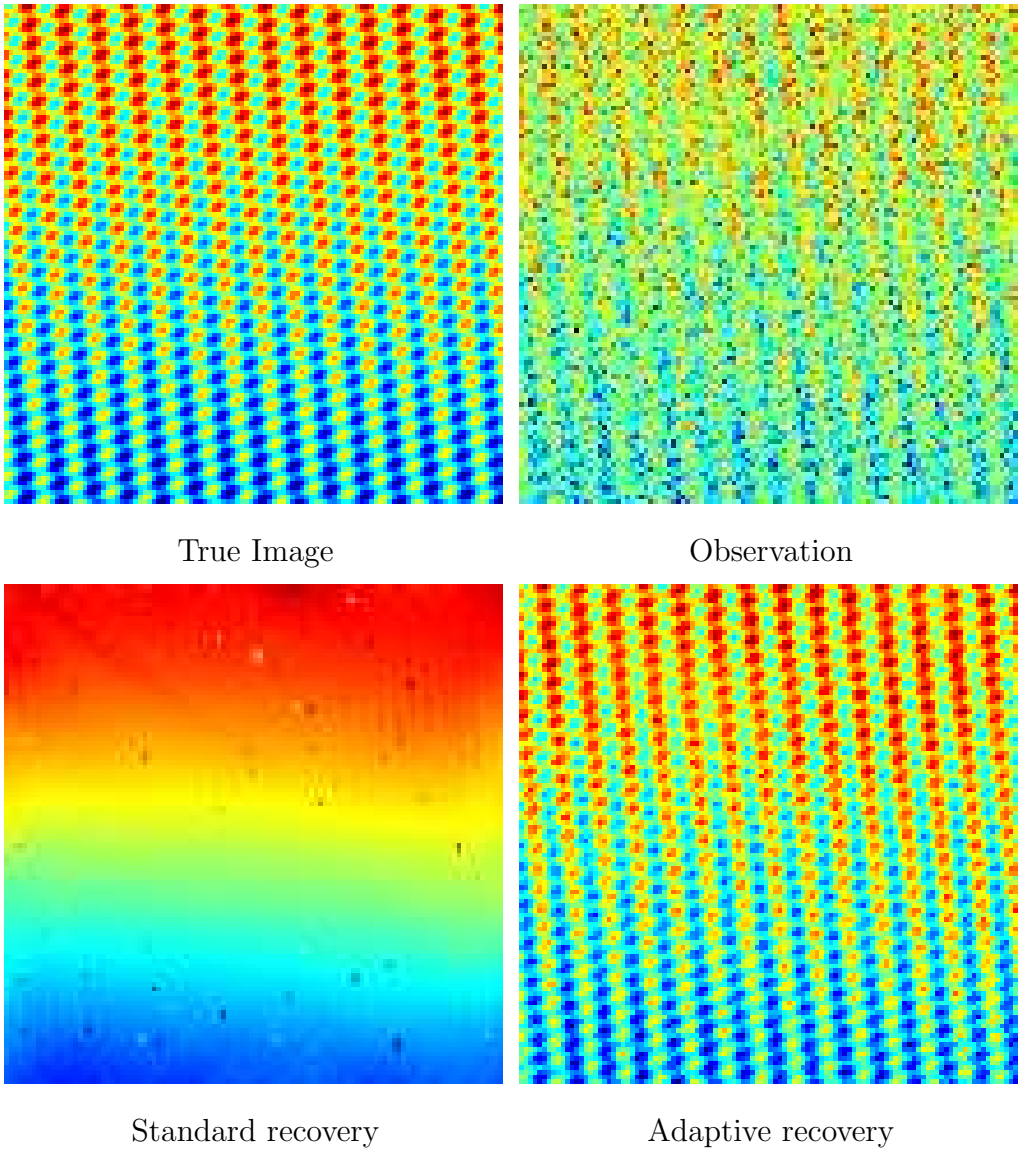

Fig. 3. Recovery for $\omega_{\max} = 32.0$

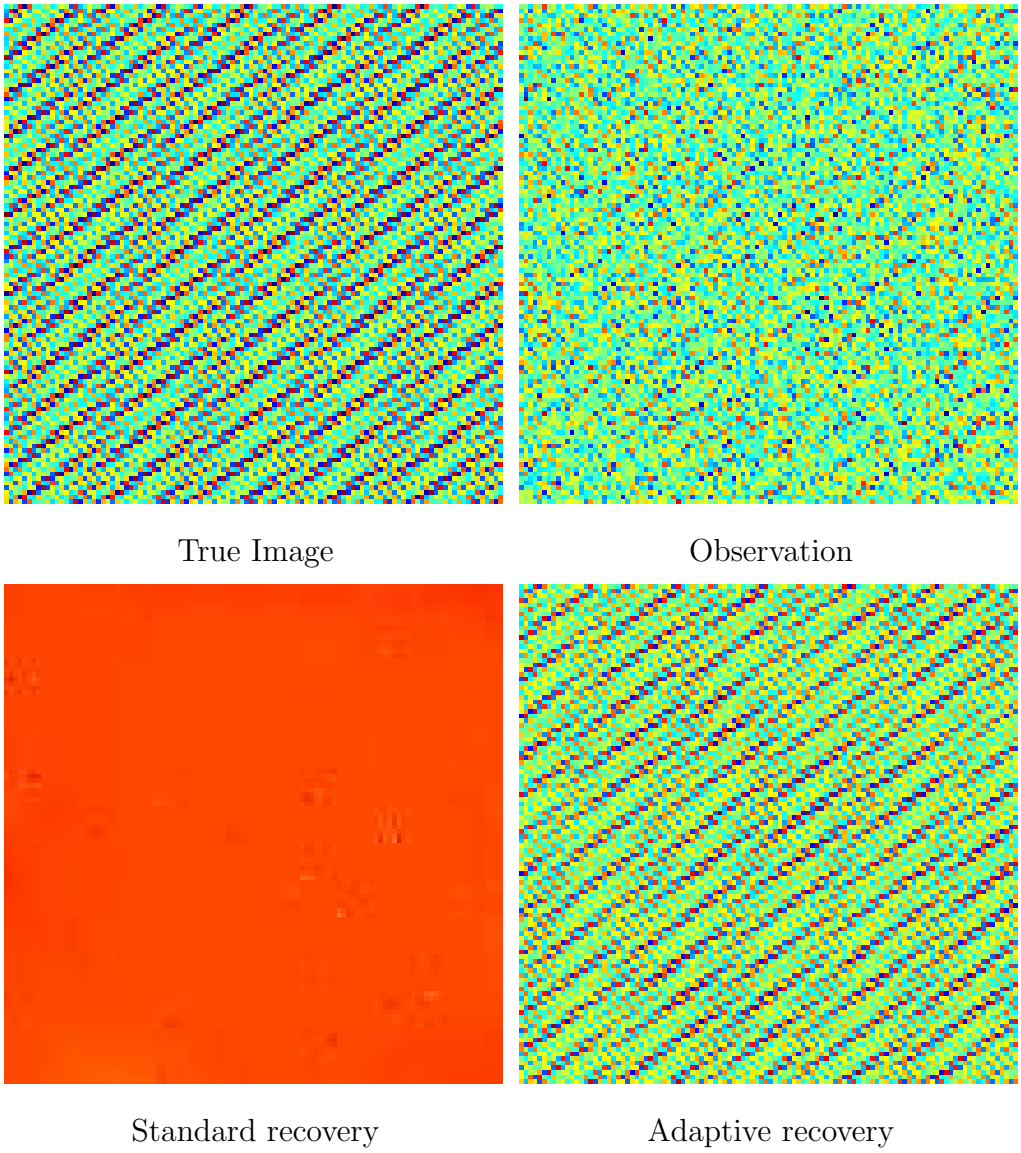

Fig. 4. Recovery for $\omega_{\max} = 128.0$