Vol. xx, pp. x

X–X

Oracle inequalities and minimax rates for non-local means and related adaptive kernel-based methods *

Ery Arias-Castro†, Joseph Salmon‡, and Rebecca Willett‡

Abstract. This paper describes a novel theoretical characterization of the performance of non-local means (NLM) for noise removal. NLM has proven effective in a variety of empirical studies, but little is understood fundamentally about how it performs relative to classical methods based on wavelets or how various parameters (*e.g.,* patch size) should be chosen. For cartoon images and images which may contain thin features and regular textures, the error decay rates of NLM are derived and compared with those of linear filtering, oracle estimators, variable-bandwidth kernel methods, Yaroslavsky's filter and wavelet thresholding estimators. The trade-off between global and local search for matching patches is examined, and the bias reduction associated with the local polynomial regression version of NLM is analyzed. The theoretical results are validated via simulations for 2D images corrupted by additive white Gaussian noise.

Key words. Non-local means (NLM), Yaroslavsky's filter, kernel smoothing, patch-based methods, local polynomial regression, oracle bounds, minimax bounds, cartoon model, textures.

1. Introduction.

The classical problem of image noise removal has drawn significant attention during the past few decades from the image processing, computational harmonic analysis, nonlinear approximation, and statistics communities. In recent years there has been a resurgence of interest in kernel-based methods, including the ubiquitous non-local means (NLM) algorithm [6], due to their practical efficacy on broad collections of "natural" images. While there is a wealth of theoretical analysis associated with nonlinear thresholding estimators based on wavelets and related sparse multiscale representations of images [13, 14, 34, 46] or on diffusion models [47, 43] and partial differential equations [37, 1], performance guarantees for NLM are lacking and this paper aims at providing some results in this direction.

In this paper, we explore the theoretical underpinnings of adaptive kernel-based image estimation and derive bounds on the mean squared error as a function of the number of pixels observed and features of the underlying image. The denoising methods we consider are all based on estimating each pixel value with a weighted sum of the surrounding pixels. Depending on how the weights in this average are selected, this framework may correspond to classical linear filters [36, 53], Yaroslavsky's filter (YF) [55], the Sigma filter [27], or the bilateral filter [49]. It also includes variable-bandwidth kernel estimators [28], often referred to as Lepski's method by statisticians and as the Intersection of Confidence Intervals (ICI) rule [22, 23] in signal processing.

As none of these methods have been explicitly designed to deal with textured regions, many authors, inspired by work on texture synthesis [16] and inpainting [8], have proposed to introduce patches (small sub-images) to take advantage of natural image redundancy, es-

*The authors started working on the paper at the Institute for Mathematics and its Applications and gratefully acknowledge support from DARPA grant no. FA8650-11-1-7150, AFOSR award no. FA9550-10-1-0390, and NSF awards no. CCF-06-43947.

†Department of Mathematics; University of California, San Diego, CA, USA.
‡Department of Electrical and Computer Engineering; Duke University, Durham, NC, USA.

pecially in textured regions. NLM [6] and UINTA [3] algorithms are typical examples of this approach, as is their extension using Lepski's method [25]. Those algorithms rely on averaging similar pixels, where the similarity is measured through patches centered on the pixel of interest. Some more elaborate methods have tried to remove artifacts appearing in regions with low redundancy [42] — a phenomenon also known as the *rare patch effect* [15] — for instance by choosing NLM parameters automatically and locally. A common tool used for this local adaptivity is the Stein Unbiased Risk Estimate (SURE) [15, 51, 52].

Most current state-of-the-art methods for denoising take advantage of the patch framework [31, 9, 10]. Despite the strong empirical performance of these methods, few performance guarantees exist: bounds with information theory flavor are derived in [54] for a simple version of NLM; a consistency result relying on beta-mixing assumptions on the image and on the noise (both modeled as random variables) is obtained in [5, 6]; [43] proposes a graph-diffusion interpretation for a simple image model; a bias/variance analysis aiming at locally choosing NLM parameters is carried out in [15]; [29, 7] obtain Cramer-Rao type efficiency results. While finishing this paper, we became aware of two related results addressing optimal performance in the context of non-parametric minimax estimation [33, 32]. [33] evaluates the performance of NLM for the piecewise constant horizon model [26], while [32] considers an anisotropic variant of NLM for the same image class. The latter shares several features with earlier work on anisotropic NLM [12, 11]. Our work is most closely related to [33], addressing the same challenge of quantifying the performance of NLM and related methods, and at the same time contains several novel contributions. We comment in more detail on [33] in Section 7.

1.1. Our contribution. We derive theoretical performance bounds for the linear filter, Yaroslavsky's filter, variable-bandwidth kernel methods and non-local means — both the original [6] and a fast patch-mean based variant of NLM [30] — in the classical "cartoon" model in which an image consists of smooth surfaces separated by a smooth discontinuity, a popular model in statistics [26]. Our results are for the local polynomial versions of these methods, which fully adapt (in some instances) to the underlying smoothness. (The systematic bias associated with classical NLM near discontinuities — and boundaries — is shown to disappear when using a local polynomial regression.) We also consider nonstandard image classes, one modeling images with thin features and another one modeling regular textures. The latter is particularly significant because it highlights some of the key advantages of patch-based methods over, say, wavelet thresholding estimators. Previous insights into the performance of NLM-like methods on textures are empirical at best; we are not aware of any theory in this vein. Our benchmarks are two oracle inequalities, though many of our theoretical results can be compared directly with similar classical results in the wavelet literature and known minimax lower bounds on mean squared error (MSE) [14, 26].

1.2. Organization of the paper. In Section 2 we describe the mathematical framework. In Section 3 we introduce the methods that we analyze in the sequel. In Section 4 we state performance guarantees in the cartoon model for each one of these methods, and in Section 5 we do the same in the context of the thin feature and regular pattern models. In Section 6 we perform some numerical experiments carefully illustrating our theoretical findings. In Section 7 we contrast our contribution to that of Maleki *et al.* [33] and discuss extensions. The proofs are gathered in Section 8, which includes general results on local polynomial regression which

may be of independent interest.

1.3. Notation. We use standard notation. For non-negative sequences (a_n) and (b_n), $a_n = O(b_n)$ (same as $a_n \preceq b_n$) if the sequence $|a_n/b_n|$ is bounded from above; $a_n \asymp b_n$ if $a_n = O(b_n)$ and $b_n = O(a_n)$; $a_n = o(b_n)$ if $a_n/b_n \to 0$ as $n \to \infty$. For real numbers a and b, $a \vee b = \max(a, b)$ while $a \wedge b = \min(a, b)$. For a Lebesgue-measurable subset $A \subset \mathbb{R}^d$, $\mathrm{Vol}(A)$ denotes its Lebesgue measure. For any vector $x \in \mathbb{R}^d$, we define its Euclidean norm and the supnorm as

$$\|x\|_2 = \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}, \qquad \|x\| := \|x\|_\infty = \max_{i=1}^{d} |x_i|.$$

For the supnorm, we use the notation $B(0,1)$ (resp. $\overline{B(0,1)}$) for the open (resp. closed) unit ball. For $\eta > 0$, we define the η-neighborhood (for the norm $\|\cdot\|$) of a set $A \subseteq \mathbb{R}^d$ as

$$B(A, \eta) = \{x \in \mathbb{R}^d : \mathrm{dist}(x, A) < \eta\}.$$

For a discrete set A, we denote its cardinality by either $|A|$ or $\#A$. For a set $A \subset \mathbb{R}^d$, $\mathbb{1}_{\{A\}}$ is the indicator function of A, while for a discrete subset $B \subset \{1, \dots, m\}$, $\mathbf{1}_B$ denotes the vector with entries indexed by B equal to one, and all others equal to zero. Additional notation is introduced in the text as needed.

2. Function estimation in additive white noise.

We cast the problem of image denoising as a non-parametric regression problem in the presence of white (centered) noise, the standard model in statistics [26]. We consider the general d-dimensional problem, and use the term "image" to denote any discretized signal on the d-dimensional square lattice, with important cases in dimensions $1 \le d \le 4$. Though patch-based methods were designed in the context of 2D images, and are still developed mostly in this context, we consider a general dimension as the same techniques may find applications in color, spectral, 3D and 4D imaging [56].

We observe noisy samples $\{y_i \in \mathbb{R} : i \in I_n^d\}$ (where $I_n := \{1, \dots, n\}$) of the target function $f : [0,1]^d \to [0,1]$ at the design points $\{x_i \in \mathbb{R}^d : i \in I_n^d\}$ corrupted by an additive noise $\{\varepsilon_i \in \mathbb{R} : i \in I_n^d\}$, as follows

$$y_i = f(x_i) + \varepsilon_i, \quad i \in I_n^d. \tag{2.1}$$

For now, we only assume that the noise $\{\varepsilon_i : i \in I_n^d\}$ are uncorrelated with mean zero and variance σ^2, though some results will require some tail bounds. Also, for concreteness, we focus on a standard model in image processing where the sample points are on the square lattice, specifically, $x_i = ((i_1 - 1/2)/n, \dots, (i_d - 1/2)/n)$ when $i = (i_1, \dots, i_d)$. Leaving n implicit, define vectors $\mathbf{y} = (y_i : i \in I_n^d)$, $\mathbf{f} = (f_i : i \in I_n^d)$ with $f_i := f(x_i)$ and $\boldsymbol{\varepsilon} = (\varepsilon_i : i \in I_n^d)$.

The vector model can thus be written

$$\mathbf{y} = \mathbf{f} + \boldsymbol{\varepsilon}\,. \tag{2.2}$$

We focus on estimating function f on the grid, namely our goal is to estimate the vector $\mathbf{f}$, we measure the performance of an estimator $\widehat{\mathbf{f}}$ in terms of (MSE):

$$\mathrm{MSE}_f(\widehat{\mathbf{f}}) = \frac{\mathbb{E}\|\widehat{\mathbf{f}} - \mathbf{f}\|_2^2}{n^d} = \frac{1}{n^d} \sum_{i \in I_n^d} \mathbb{E}(\widehat{f}_i - f_i)^2\,,$$

where the expectation $\mathbb{E}$ is with respect to the probability measure associated with the noise.

Although our analysis may be generalized to other norms, mean squared error is handy because of the point-wise (squared) bias and variance decomposition:

$$\mathbb{E}(\widehat{f}_i - f_i)^2 = \underbrace{(\mathbb{E}\widehat{f}_i - f_i)^2}_{\text{Squared Bias}} + \underbrace{\mathbb{E}\big(\mathbb{E}(\widehat{f}_i) - \widehat{f}_i\big)^2}_{\text{Variance}}, \quad \forall i \in I_n^d. \tag{2.3}$$

This leads for the vector estimate to the following decomposition:

$$\mathbb{E}\|\widehat{\mathbf{f}} - \mathbf{f}\|_2^2 = \|\mathbb{E}(\widehat{\mathbf{f}}) - \mathbf{f}\|_2^2 + \mathbb{E}\|\mathbb{E}(\widehat{\mathbf{f}}) - \widehat{\mathbf{f}}\|_2^2 .$$

To recover the function f only through a finite number of measurements, it is customary to require that the targeted function belongs to a class $\mathcal{F}$ of structured functions such as smooth, piecewise smooth, or periodic textured images. In this context, the minimax risk over the function class $\mathcal{F}$ is defined as

$$\mathcal{R}_n^*(\mathcal{F}) = \inf_{\widehat{\mathbf{f}}} \sup_{f \in \mathcal{F}} \mathrm{MSE}_f(\widehat{\mathbf{f}}),$$

where the infimum is over all the measurable function with respect to the observations. We say that an estimator is (rate-)optimal for the class $\mathcal{F}$ if its worst-case MSE over $\mathcal{F}$ is comparable to the minimax risk, *i.e.,* (assuming implicitly that n becomes large)

$$\mathcal{R}_n(\widehat{\mathbf{f}}, \mathcal{F}) := \sup_{f \in \mathcal{F}} \mathrm{MSE}_f(\widehat{\mathbf{f}}) = O(\mathcal{R}_n^*(\mathcal{F})).$$

2.1. Cartoon images. We are particularly interested in situations where the function f has discontinuities: this is typical of images, mainly because of occlusions occurring in natural scenes. We say that f is a "cartoon image" if it is a piecewise smooth image with discontinuities along smooth hypersurfaces. This model spurred the greatest part of the research in image processing and is very common when no texture is present [26]. For simplicity, we consider that f is made of two pieces with each piece being Hölder smooth. Note that all our results apply to the more general case where f is made of more than two pieces. For a function $g : \mathbb{R}^d \to \mathbb{R}$ and $s = (s_1, \dots, s_d) \in \mathbb{N}^d$, we denote the s-derivative of g at $x \in \mathbb{R}^d$ by

$$g^{(s)}(x) = \frac{\partial^{|s|}}{\partial_{x_1}^{s_1} \cdots \partial_{x_d}^{s_d}} g(x),$$

where $|s| := s_1 + \cdots + s_d$.

Definition 2.1 (Hölder function class). *For $\alpha, C_0 > 0$, we define $\mathcal{H}_d(\alpha, C_0)$ as the Hölder class of functions $g : [0,1]^d \to [0,1]$ that are $\lfloor \alpha \rfloor$ times differentiable ($\lfloor \alpha \rfloor$ is the largest integer strictly less than α) and satisfy*

$$\forall x \in [0,1]^d,\ \forall s \in \mathbb{N}^d, 1 \le |s| \le \lfloor \alpha \rfloor : \quad |g^{(s)}(x)| \le C_0; \tag{2.4}$$

$$\forall (x, x') \in [0,1]^d,\ \forall s \in \mathbb{N}^d, |s| = \lfloor \alpha \rfloor : \quad |g^{(s)}(x) - g^{(s)}(x')| \le C_0 \|x - x'\|_\infty^{\alpha - \lfloor \alpha \rfloor}. \tag{2.5}$$

The main feature of Hölder functions of order α is that they are well-approximated locally by a polynomial (in fact, their Taylor expansion) of degree $\lfloor \alpha \rfloor$, *cf.* Lemma 8.1.

Definition 2.2 (Cartoon function class). *For $\alpha, C_0 > 0$, let $\mathcal{F}^{\text{cartoon}}(\alpha, C_0)$ denote the set of functions of the form*

$$f(x) = \mathbb{1}_{\{x\in\Omega\}} f_\Omega(x) + \mathbb{1}_{\{x\in\Omega^c\}} f_{\Omega^c}(x), \tag{2.6}$$

where $f_\Omega, f_{\Omega^c} \in \mathcal{H}_d(\alpha, C_0)$, with jump (or discontinuity gap)

$$\mu(f) := \inf_{x\in\partial\Omega} |f_\Omega(x) - f_{\Omega^c}(x)| \geq 1/C_0, \tag{2.7}$$

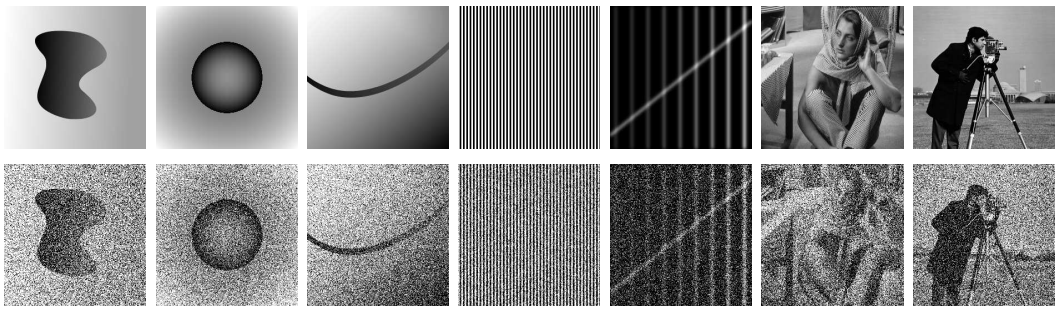
*and $\Omega \subset (0,1)^d$ is a bi-Lipschitz image of the (Euclidean) unit ball $B(0,1)$, specifically, $\Omega = \phi(B(0,1))$, where $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is injective with ϕ and ϕ^{-1} both Lipschitz with constant C_0 (*i.e., *C_0-Lipschitz) with respect to the supnorm. We refer to f_Ω as the foreground and to f_{Ω^c} as the background. Moreover $\partial\Omega$ represents the (topological) boundary of Ω.*

The condition (2.7) is a lower bound on the minimum "jump" along the discontinuity $\partial\Omega$. We require that ϕ is bi-Lipschitz to ensure that the set Ω is sufficiently smooth and does not have a serious bottleneck, which could potentially mislead the methods discussed here.

We define the signal-to-noise ratio (SNR) for a target function f with jump $\mu(f)$, and noise standard deviation σ, as being the quantity

$$\text{SNR} = \frac{\mu(f)}{\sigma}. \tag{2.8}$$

We assume throughout that $\mu \asymp 1$, so that our bounds (which scale with σ) reflect performance also as a function of SNR. In the cartoon model, we focus on the case where the noiseless image is at least piecewise Lipschitz, that is, $\alpha \geq 1$. Note that our results apply to the case where $\alpha > 1/2$, and that simple linear filtering is essentially optimal when $\alpha \leq 1/2$. The setting is illustrated in Figure 2.1(a).



(a) Blob (b) Bowl (c) Swoosh (d) Stripes (e) Ridges (f) Barbara (g) Cameraman

Figure 2.1. *Original and noisy images: cartoon (Blob, Bowl), thin features (Swoosh), texture (Stripes) and natural images (Ridges, Barbara, Cameraman). Blob and Swoosh are piecewise polynomial of order 1, while Bowl is piecewise polynomial of order 2.*

2.2. Thin features and textures. In addition to considering cartoon images as defined above, we will consider images which contain other features common in natural images, such as thin regions a few pixel wide and regular textures. We consider simple models for these and show that YF and, more generally, the NLM perform much better than linear filtering.

These models are instances of the cartoon model where the forefront Ω varies with n. Let $\mathcal{F}(\alpha, C_0)$ be defined as $\mathcal{F}^{\text{cartoon}}(\alpha, C_0)$ but without constraints on Ω.

As a simple model of thin feature, consider an image f in the cartoon family, but where Ω is a thin d_0-dimensional surface of thickness a — which will vary with n. A classical example of this kind of structure is the support bar of the Cameraman's tripod, see Figure 2.1(g). An example of function from this class is illustrated by the Swoosh image, see Figure 2.1(c).

Definition 2.3 (Thin feature function class).

$$\mathcal{F}^{\text{thing}}(\alpha, C_0, d_0, a) := \left\{ f \in \mathcal{F}(\alpha, C_0) : \Omega = \{x = (x', z) : \operatorname{dist}(z, \phi(x')) < a\} \right\},$$

where $\phi : (0,1)^{d_0} \to (0,1)^{d-d_0}$ *is* C_0*-Lipschitz.*

We may similarly define a class of regular pattern functions which themselves may not be smooth, but which occur repeatedly across the image domain. This structure would be difficult to exploit with, say, wavelet-based methods that fail to take advantage of image redundancy. However, empirical evidence suggests that non-local adaptive kernels can perform quite well on these images. A classical example of this type of image structure is the striped scarf in the Barbara image. The following is a class where Ω is made of the disjoint union of translates of a smaller region Ω_0 of diameter of order a — which will vary with n.

Definition 2.4 (Regular pattern function class).

$$\mathcal{F}^{\text{pattern}}(\alpha, C_0, a) := \left\{ f \in \mathcal{F}(\alpha, C_0) : \Omega = (0,1)^d \cap \bigcup_{v \in a\mathbb{Z}^d} (\Xi + v) \right\},$$

where $\Xi \subset (0,a)^d$ *is any set. Note that the union above is disjoint.*

An example of function from this class is illustrated by the Stripes image in Figure 2.1(a).

3. Background on kernel methods for denoising. We now describe NLM and other related methods. The story starts with kernel smoothing (*i.e.,* linear filtering). Though this age-old method (with a proper choice of kernel) is essentially optimal when the image does not have discontinuities, its performance suffers dramatically in the presence of edges, which it tends to blur. YF, and more generally NLM, attempt to choose the kernel adaptively so as to avoid averaging over the discontinuity.

These estimates are weighted averages of the pixel values of the form

$$\widehat{f}_i = \frac{\sum_{j \in I_n^d} \omega_{i,j}\, y_j}{\sum_{j \in I_n^d} \omega_{i,j}}. \tag{3.1}$$

The various methods that we study in this paper differ only in the choice of weights $\omega_{i,j}$. Adaptation to higher order of smoothness is often accomplished by a local polynomial regression (LPR) [17, 20]. The local polynomial estimator of degree r and weights $(\omega_{i,j})$ is

$$\begin{cases} \widehat{f}_i = \widehat{a}_0^{(i)} \\ \widehat{\mathbf{a}}^{(i)} = \arg\min_{\mathbf{a}} \sum_{j \in I_n^d} \omega_{i,j} \left(y_j - \sum_{0 \le |s| \le r} a_s \, (x_j - x_i)^s \right)^2, \end{cases} \tag{3.2}$$

where $x^s := x_1^{s_1} \cdots x_d^{s_d}$, for $x = (x_1, \ldots, x_d) \in \mathbb{R}^d$ and $s = (s_1, \ldots, s_d) \in \mathbb{R}^d$, and the minimization in (3.2) is over $\mathbf{a} = (a_s : 0 \leq |s| \leq r) \in \mathbb{R}^q$ where $q = \binom{r+d}{d}$. Note that, in fact, (3.2) leads to an estimator of the form (3.1) with different weights [50, p. 34]. We assume throughout that the polynomial degree r is sufficiently large to take full advantage of the smoothness of f. Specifically, if $f \in \mathcal{F}^{\text{cartoon}}(\alpha, C_0)$, we assume that $r \geq \lfloor \alpha \rfloor$. When the number of nonzero weights in (3.2) is not enough to determine $\widehat{f}_i$ uniquely, we define $\widehat{f}_i$ as y_i, namely, we do not apply any smoothing. Alternatively, one could decrease the degree of the polynomial regression until the fit is well-defined, but this is not important in our setting.

Since we know that f takes values in $[0, 1]$, we clip $\widehat{f}$ so that it also takes values in $[0, 1]$. This clipping does not increase the MSE.

3.1. Linear filtering (LF). This method can be traced back in the statistics literature to the work of Nadaraya [36] and Watson [53]. In this context the choice of the similarity between two pixels is only controlled by spatial proximity:

$$\omega_{i,j} = K_h(x_i, x_j)\,, \tag{3.3}$$

where $K_h(x, x') = K(\frac{x}{h}, \frac{x'}{h})$ for some kernel function K and some bandwidth $h > 0$, which is independent of the location in the nonadaptive (classical) version (*cf.* [20] for more details on kernel methods). Common choices include the Gaussian kernel, but we focus on the box kernel

$$K_h(x, x') = \mathbb{1}\{\|x - x'\|_\infty \leq h\}. \tag{3.4}$$

3.2. Yaroslavsky's filter (YF). YF was introduced by Yaroslavsky [55] and independently by Lee [27], with more modern variants such as SUSAN [44] and Bilateral filtering [49]. In these methods, similarity between two pixels is based on their spatial distance and on the relative proximity of image intensity at these pixels. This translates into choosing weights in (3.1) of the form

$$\omega_{i,j} = K_h(x_j, x_j)\; L_{h_y}(y_i, y_j)\,, \tag{3.5}$$

where K, L are kernels and h, h_y the associated bandwidths. (K, h) control the spatial proximity while (L, h_y) control the photometric proximity. As in classical kernel smoothing, h plays the role of spatial bandwidth, while h_y is a photometric bandwidth. In this work we only consider the simple version using the box kernel:

$$K_{h_y}(y, y') = \mathbb{1}\{|y - y'| \leq h_y\}. \tag{3.6}$$

3.3. Non-Local Means (NLM) and patch-based methods. NLM and other patch-based methods generalize the idea of including the photometric proximity in the kernel. In [6], the distance between two pixels is solely measured in terms of the discrepancy between patches surrounding the pixels considered. Though spatial proximity was already introduced in [6], it was only mentioned as a numerical parameter to solve a computational issue. However, later works (*cf.* [41]) have shown that spatial proximity can improve NLM performance. We consider NLM with spatial proximity, which includes the non-local version, the two being identical when h is sufficiently large.

A generic description is the following. Let $h_{\mathsf{P}} > 0$ and let P_i (leaving h_{P} implicit) be the hypercube of width h_{P} centered at x_i, *i.e.*,

$$\mathsf{P}_i = x_i + \left[-\frac{h_{\mathsf{P}}}{2}, \frac{h_{\mathsf{P}}}{2}\right]^d = \left\{x : \|x - x_i\|_\infty \le \frac{h_{\mathsf{P}}}{2}\right\}. \tag{3.7}$$

Such a patch corresponds to a pixel patch of width $[h_{\mathsf{P}} n] + 1$ in the digital image (where $[a]$ denotes the largest integer not exceeding $a \in \mathbb{R}$). Let $\mathbf{y}_{\mathsf{P}_i} = (y_j : x_j \in \mathsf{P}_i)$ be the vector of pixel values over the patch centered at x_i. With this notation, the weights used in NLM are:

$$\omega_{i,j} = K_h(x_i, x_j)\ L_{h_y}\left(\mathbf{y}_{\mathsf{P}_i}, \mathbf{y}_{\mathsf{P}_j}\right)\ , \tag{3.8}$$

where K, L are kernel functions and h, h_y are bandwidths, as before. One classical choice of L_{h_y} (which we consider in our theoretical results) is

$$L_{h_y}\left(\mathbf{y}_{\mathsf{P}_i}, \mathbf{y}_{\mathsf{P}_j}\right) = \mathbb{1}\{\|\mathbf{y}_{\mathsf{P}_i} - \mathbf{y}_{\mathsf{P}_j}\|_2 \le h_y\}. \tag{3.9}$$

The photometric similarity is based on the Euclidean distance between the patches (as vectors) around the pixels. We refer to this as "classical" or Euclidean NLM (or just NLM).

Computing L_{h_y} can be computationally intensive for large h_{P}. To address this, some authors have considered projecting $\mathbf{y}_{\mathsf{P}_i}$ onto a low-dimensional subspace and using this projection to compute an approximation of $L_{h_y}(\mathbf{y}_{\mathsf{P}_i}, \mathbf{y}_{\mathsf{P}_j})$. This introduces an interesting trade-off between computational complexity and accuracy which is examined in [4, 48]. In this paper, we consider a 1-dimensional projection introduced in [30] where patches are simply compared via their means alone, resulting in a photometric kernel of the form

$$L_{h_y}\left(\mathbf{y}_{\mathsf{P}_i}, \mathbf{y}_{\mathsf{P}_j}\right) = \overline{L}_{h_y}\left(\overline{y}_{\mathsf{P}_i}, \overline{y}_{\mathsf{P}_j}\right), \quad \overline{y}_{\mathsf{P}_i} := \mathrm{Ave}(\mathbf{y}_{\mathsf{P}_i}). \tag{3.10}$$

We refer to this method as NLM-average. In our theoretical analysis, we consider the photometric kernel

$$\overline{L}_{h_y}\left(\overline{y}_{\mathsf{P}_i}, \overline{y}_{\mathsf{P}_j}\right) = \mathbb{1}\{|\overline{y}_{\mathsf{P}_i} - \overline{y}_{\mathsf{P}_j}| \le h_y\}. \tag{3.11}$$

For intuition, consider a piecewise constant image with no noise. Take $x_i \in \Omega$ away from the discontinuity $\partial\Omega$. On the one hand, when $x_j \in \Omega$ is also away from the discontinuity, the patch vectors $\mathbf{y}_{\mathsf{P}_i}$ and $\mathbf{y}_{\mathsf{P}_j}$ are identical, and so are the patch averages $\overline{y}_{\mathsf{P}_i}$ and $\overline{y}_{\mathsf{P}_j}$, so the regression weight ω_{ij} is solely based on the spatial distance between x_i and x_j. On the other hand, when $x_j \in \Omega^c$ the patch vectors $\mathbf{y}_{\mathsf{P}_i}$ and $\mathbf{y}_{\mathsf{P}_j}$ are quite different (thanks to (2.7)), and similarly for the patch averages, so x_j is excluded from the regression neighborhood of x_i, meaning that $\omega_{ij} = 0$, even when the two points are spatially close to each other — this is assuming that h_y is properly chosen. When the function is piecewise smooth, the same reasoning applies as the function is morally constant within a (small) spatial window.

In our analysis, NLM Euclidean (3.9) and NLM-average (3.11) behave similarly, except for the regular pattern model, where the former is generally superior. In practice, however, we note a difference. In smooth regions, the average in (3.11) has little bias and little variance, making it significantly more robust to noise than the Euclidean distance (3.9). Near edges or

patterns, however, the bias of the average in (3.11) can outweigh the variance, making NLM Euclidean (3.9) superior. This insight is supported by our experimental results in Section 6.

The spatial bandwidth h is typically larger than the patch width h_{P}. Common sizes used in practice are 21×21 kernel windows (also referred to as the searching zone) and 7×7 patches (in pixel units) and common kernels are the box-kernel for K and the Gaussian kernel for L. Though we assume box kernels for both, our results extend readily to other kernel functions.

4. Oracle inequalities and minimax results for cartoon images. We analyze the performance of the kernel-based methods described in Section 3 within the mathematical framework detailed in Section 2. Qualitatively speaking, our theoretical results are congruent with what is observed in practice; see our experiments in Section 6.

Indeed, we show that standard kernel smoothing blurs edges, which is in fact well-known both in theory and practice. YF performs well when the SNR is large, and poorly otherwise. This filter relies on a clear gap between the pixel values on either side of the discontinuity: when the SNR is large, there is indeed a gap, which ceases to exist when the SNR is of order 1 (see Figure 4.1 for an illustration). The latter situation is where NLM shines. Indeed, patches of size larger than just one pixel gather more information about the area surrounding the pixel, which NLM (implicitly) uses to assess whether two pixels are on the same size of the discontinuity. For example, comparing patches in Figure 4.2, we see that the means of sufficiently large patches allow us to estimate reliably whether each center pixel is in Ω or not, even with an SNR of order 1.

In what follows, we focus on the local polynomial variants described in (3.2) to avoid a systematic bias that conventional weighted average variants suffer from. It is well-known that this bias appears near the boundary of the image, though this can be corrected with a proper extension of the image. More importantly, this bias arises also near the discontinuity. Note that enforcing the spatial windows to have the same (symmetric) shape puts a real constraint on the resulting performance of the algorithm, as discussed in Section 4.3. The choice of kernels K and L for LPR variants (3.2) is unimportant as long as they satisfy some basic properties. For concreteness we use the box kernels for both the spatial and photometric components, namely (3.4) and (3.9) or (3.11)

When minimizing MSE, we tune each method in an optimal way to strike the best (for that method) balance between bias and variance (2.3). Indeed, the larger the bandwidth, the larger the bias and the smaller the variance. The issue with kernel smoothing — whether in the form of weighted average (3.1) or LPR (3.2) — is that it suffers from a substantial bias when the smoothing window (those points where the weights are equal to one) includes points on the 'other side' of the discontinuity. At the same time, the window cannot be too small, for otherwise the variance will be overwhelming. Different kernel methods handle this differently.

4.1. Linear kernel smoothing blurs edges. It is well-known that linear kernel smoothing blurs discontinuities. This comes from the fact that the window size is fixed — the same at all pixels — so points near and points far from the discontinuity are treated in the same way. This lack of adaptivity leads to a substantial MSE. How does that statement translate into a mathematical result within our framework? The following is proved in [2] for $d \leq 2$, though the result (at least the upper bound) is probably older. We do provide a proof in Section 8 for LPR.

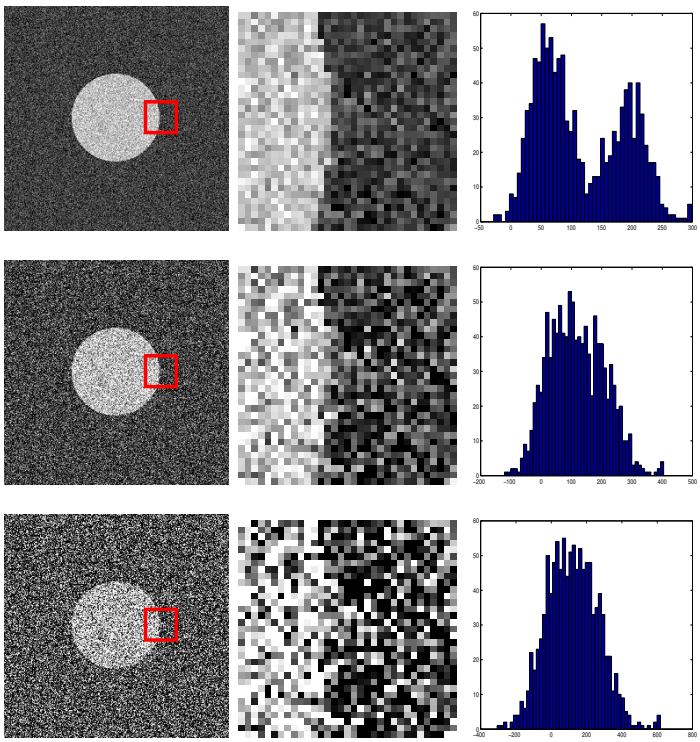


Figure 4.1. *On the left column are disk shape cartoon images with increasing levels of noise (rows are respectively with $SNR = 4, 2, 1$ from top to bottom). A searching zone is displayed in red, for a pixel near the discontinuity. The middle column is a close-up inside the searching zone, while the right one provides histograms of pixel values within it.*

Theorem 4.1. *Let $\widehat{\mathbf{f}}_h^{\mathrm{LF}}$ denote the linear estimator, in the form of either local average (3.1) or LPR (3.2), with weights as in (3.3). We have*

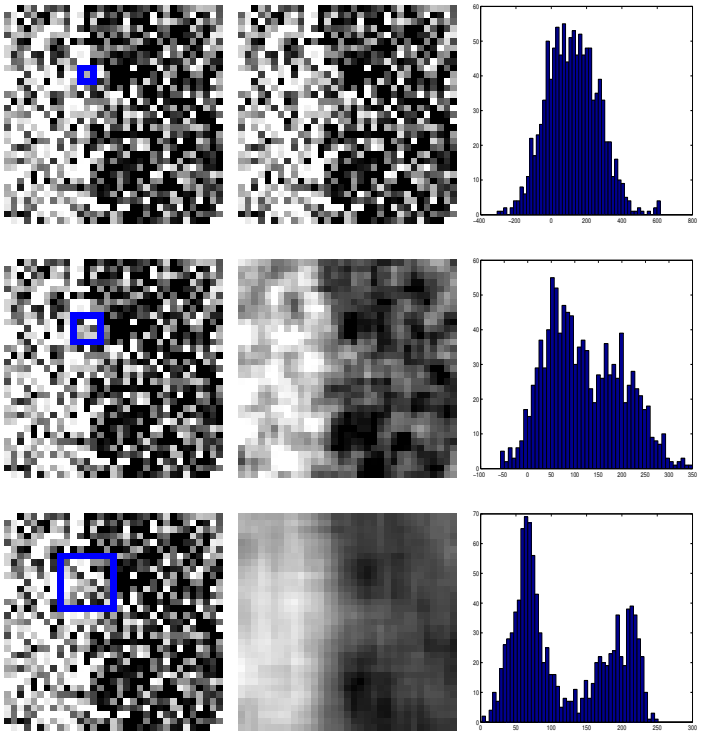
$$\inf_h \mathcal{R}_n(\widehat{\mathbf{f}}_h^{\mathrm{LF}}, \mathcal{F}^{\mathrm{cartoon}}(\alpha, C_0)) \asymp \mathcal{R}^{\mathrm{LF}} := (\sigma^2/n^d)^{1/(d+1)},$$

and the optimal choice of bandwidth is $h \asymp h^{\mathrm{LF}} := (\sigma^2/n^d)^{1/(d+1)}$.

Note that the bound does not depend on the regularity $\alpha \geq 1$ of the target function. As apparent in the proof, and intuitively clear, this is due to the fact that kernel smoothing blurs edges: to strike a good bias-variance trade-off, the smoothing window cannot be too small, transforming sharp edges into ramps. The resulting bias is then larger than the bias over the smooth regions, which is where α appears.

4.2. Oracle kernel. What can we hope to achieve with adaptive kernel methods? Statisticians have used the notion of an oracle to answer this question [13, 21, 50]. We saw that what limits linear filtering is a large bias near the discontinuity, due to the mixing of pixels from both sides. What if we had access to an oracle that would identify for us the foreground and the background?

The *membership oracle* tells us which sample points belong to Ω or to Ω^c. With access to this oracle, we simply process the smooth pieces, f_Ω and f_{Ω^c}, separately. By doing so



(a) Searching zone and averaging window (b) Local averaging (c) Histogram of the averaged pixels values in (b)

Figure 4.2. *We use the same setting as in Figure 4.1, last row (SNR=1). The noisy image is displayed in the first column, with examples of kernel supports. The second column is the result of the local (box) kernel averaging using support of width 1, 3 and 7 from top to bottom. The last column provides histograms of the filtered pixels shown in (b).*

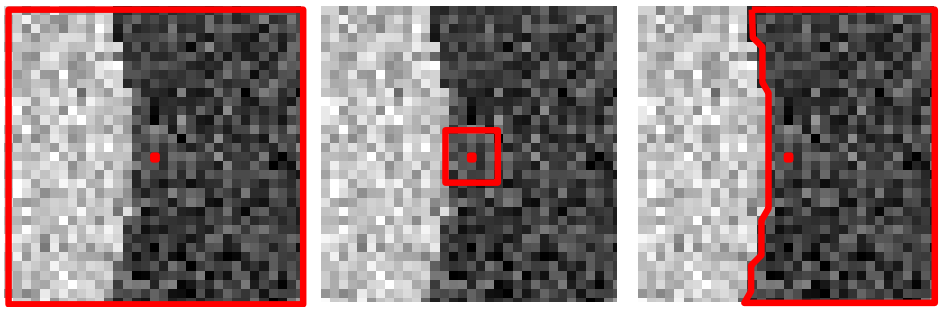
we achieve the minimax rate for the class $\mathcal{H}_d(\alpha, C_0)$: the information this oracle provides is sufficient to do as well as if there were no discontinuity. This is illustrated in Figure 4.3 (best viewed in color).

Theorem 4.2. *Let $\widehat{\mathbf{f}}_h^{\mathrm{MO}}$ denote LPR estimator (3.2) with weights as in (3.3) when x_i and x_j belong to the same side of the discontinuity, and set to zero otherwise. We have*

$$\inf_h \mathcal{R}_n(\widehat{\mathbf{f}}_h^{\mathrm{MO}}, \mathcal{F}^{\mathrm{cartoon}}(\alpha, C_0)) \asymp \mathcal{R}^{\mathrm{MO}} := (\sigma^2/n^d)^{2\alpha/(d+2\alpha)},$$

and the optimal choice of bandwidth is $h \asymp h^{\mathrm{MO}} := (\sigma^2/n^d)^{1/(d+2\alpha)}$.

The lower bound is a well-known minimax bound [26, Theorem 5.1.2, p. 133]. If we consider a class of piecewise polynomial functions, then this oracle estimator, without spatial proximity (*i.e.*, $h = \infty$), achieves the parametric rate of σ^2/n^d. It is worth noting that LPR plays a crucial role here. Indeed, the window around a point near the discontinuity — comprised of all points belonging to the same side of the discontinuity — will be irregularly



(a) Kernel smoothing (b) *Bandwidth oracle* (c) *Membership oracle*

Figure 4.3. *The kernel supports for the linear filter, the bandwidth oracle and the membership oracle*

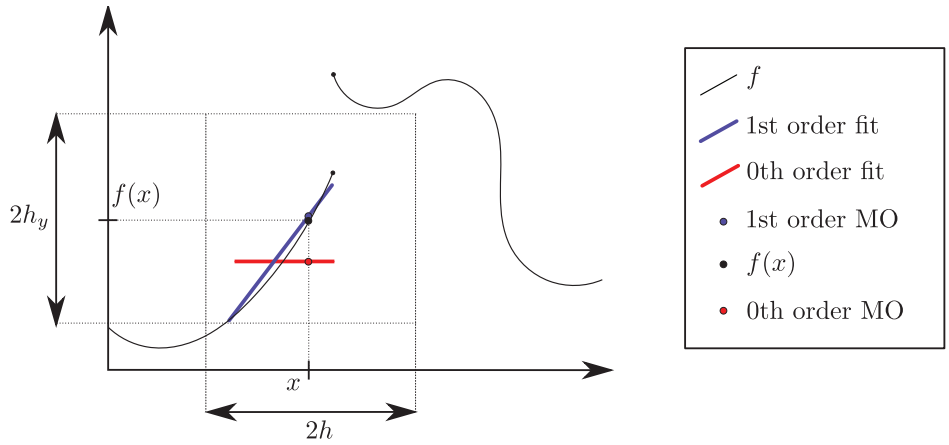


Figure 4.4. *Membership oracles of order 0 and 1 on a non-noisy 1D signal. Note how the bias is reduced by going to the order 1.*

shaped. For instance, imagine a point on a linear surface adjacent to the discontinuity. For a symmetric window (sufficiently small not to include the discontinuity), the linear variations around the pixel of interest will average out and we can accurately estimate the pixel value. For an *asymmetric* window caused by the discontinuity, the linear variations will not average out, inducing a small bias and leading to a higher risk of order $(\sigma^2/n^d)^{3/(d+3)}$ when $\alpha \geq 3/2$. This phenomenon can be observed in practice and is illustrated in Figure 4.4.

Note that the oracle only has to provide the membership information locally, within the smoothing window. The insight we get from this is that we only need to know over which pixel values to average to attain the same error rate as we would without discontinuities. This is exactly what adaptive kernel methods [28], including patch-based methods, PDE methods [37, 1, 19] and graph diffusion methods [47, 43] aim at doing.

4.3. Variable bandwidth kernel methods. These methods [35, 45, 22, 23, 24], including Lepski's method [28] and variants [38, 39], choose the bandwidth adaptively at every location, the goal being to avoid smoothing over discontinuities and to adapt to the regularity of the signal when unknown. Wavelet shrinkage methods are often thought to perform some sort of

variable-bandwidth kernel smoothing [13]. What can we hope to achieve with these methods? Clearly, we cannot do better than if we knew the discontinuity, meaning if we had access to the *membership oracle*. In that case, at each point we would choose the bandwidth equal to its distance to the discontinuity (BO below stands for *bandwidth oracle*). See Figure 4.3 for a comparison of the MO and BO spatial supports.

Theorem 4.3. *Let* $\widehat{\mathbf{f}}_h^{\mathrm{BO}}$ *denote LPR estimator* (3.2) *with weights chosen as in* (3.3) *when* $\|x_i - x_j\|_\infty < \operatorname{dist}(x, \partial\Omega) := \inf_{y\in\partial\Omega} \|x - y\|_\infty$*, and set to zero otherwise. We have*

$$\inf_h \mathcal{R}_n(\widehat{\mathbf{f}}_h^{\mathrm{BO}}, \mathcal{F}^{\mathrm{cartoon}}(\alpha, C_0)) \asymp \mathcal{R}^{\mathrm{BO}} := (A_n\sigma^2/n) \vee (\sigma^2/n^d)^{2\alpha/(d+2\alpha)},$$

where $A_n = \log n$ *when* $d = 1$ *and* $A_n = 1$ *when* $d \geq 2$*, for an optimal choice of maximal bandwidth* $h \asymp h^{\mathrm{MO}}$*.*

Note that BO achieves the error rate of MO only when $d = 1$, when $d = 2$ and $\alpha = 1$, or when $d \geq 2$ and $\sigma^2 = O(n^{-2\alpha(1-1/d)+1})$, which is polynomially small when $d = 2$ and $\alpha > 1$ or when $d \geq 3$. Thus in general, the *bandwidth oracle* is substantially weaker than the *membership oracle*. That said, BO is still always substantially better than LF — unless σ^2 is a positive power of n. Of course, we can always ignore the information that the oracle provides, in which case we would not do worse than linear filtering!

4.4. Yaroslavsky's filter is oracle-optimal under low noise. As the practitioner knows, YF can be quite good on natural images. In fact, it can dramatically outperforms the linear filter and compares favorably with methods such as wavelet thresholding, particularly when the noise level is small. We substantiate this empirical evidence with a sharp theoretical study of its performance, showing it achieves the *membership oracle* bound in such situations (*i.e.,* when σ is small).

Assume that for a fixed cumulative distribution function F, the noise satisfies the following

$$\mathbb{P}(|\varepsilon_i| \leq t) \geq F(t/\sigma), \ \forall t, \ \forall i \in I_n^d, \tag{4.1}$$

The following result states that YF achieves a performance comparable to that of MO as long as the noise variance is small. We only require that the noise distribution in (4.1) has quickly decaying tails.

Theorem 4.4. *Let* $\widehat{\mathbf{f}}_{h,h_y}^{\mathrm{YF}}$ *denote the LPR estimator* (3.2) *with Yaroslavsky's weights* (3.5)*. Suppose that, for some constants* $C, b > 0$*,* (4.1) *holds with* $1 - F(t) \leq C\exp(-(t/C)^b)$ *for* t *large enough. Then there is another constant* $C' > 0$ *such that, if* $\sigma \leq (C'\log n)^{-1/b}$*,*

$$\inf_{h,h_y} \mathcal{R}_n(\widehat{\mathbf{f}}_{h,h_y}^{\mathrm{YF}}, \mathcal{F}^{\mathrm{cartoon}}(\alpha, C_0)) \leq (1 + o(1))\mathcal{R}^{\mathrm{MO}},$$

where an optimal choice of bandwidths is $h \asymp h^{\mathrm{MO}}$ *and* $h_y \asymp 1$*.*

Gaussian noise satisfies the requirements of Theorem 4.4 with $C = \sqrt{2}$ and $b = 2$, resulting in the constraint $\sigma = O(1/\sqrt{\log n})$, which is quite mild. This explains why YF tends to perform well in practice.

This excellent performance hinges on the fact that the photometric kernel is able to mimic the *membership oracle* when the noise level is small. When the noise level is of order 1 or larger, this is no longer true, as illustrated in Figure 4.1. There, we clearly see that in a window containing points from both Ω and its complement, the pixel values are mixed in the histogram if the noise level is too large, making a clear separation impossible. We formally argue this point after the proof of Theorem 4.4 in Section 8.2.4.

It is worth noting that the proof helps clarify exactly the artifacts encountered in practice by the YF for strong noise (for instance, see Figure 6.5). Indeed, the output often looks like the original scene contaminated by something like "salt and pepper" noise. As mentioned in the proof, this is because the YF does not alter pixels with extreme values.

4.5. Performance analysis for Non-Local Means. In the previous section we established that YF performs as well as MO when the noise level is small, while it is useless otherwise. A natural strategy consists of, first, reducing the noise level by averaging and, then, applying YF. This is almost exactly what NLM-average does. We precisely quantify the MSE performance of both NLM-average and Euclidean NLM in this section. Note that we state our results for i.i.d. Gaussian noise for simplicity, though they are valid for many other distribution families such as uniform and double-exponential.

Theorem 4.5. *Let $\widehat{\mathbf{f}}^{\mathrm{NLM}}_{h,h_y}$ denote LPR estimator (3.2) with NLM weights (3.8) and photometric kernel either Euclidean (3.9) or Average (3.11). Assuming σ is bounded away from 0, we have*

$$\inf_{h,h_y} \mathcal{R}_n(\widehat{\mathbf{f}}^{\mathrm{NLM}}_{h,h_y}, \mathcal{F}^{\mathrm{cartoon}}(\alpha, C_0)) \asymp \mathcal{R}^{\mathrm{NLM}} := (B_n/n) \vee (\sigma^2/n^d)^{2\alpha/(d+2\alpha)},$$

where $B_n := (\sigma^4 \log n)^{1/d}$ (Euclidean) or $:= (\sigma^2 \log n)^{1/d}$ (Average), and an optimal choice of bandwidths is $h \asymp h^{\mathrm{MO}}$, $h_y \asymp h_y^{\mathrm{NLM}} := \sigma^3 \sqrt{\log n}$ (Euclidean) or $:= (2C_0)^{-d}\mu/2$ (Average), and $h_P \asymp h_P^{\mathrm{NLM}} := B_n/n$.

Note that there is a strong correspondence between this bound and the *bandwidth oracle* bound in Theorem 4.3. In particular, assuming σ is fixed, $\mathcal{R}^{\mathrm{NLM}} \asymp \mathcal{R}^{\mathrm{BO}}$ when $d = 1$ and $\mathcal{R}^{\mathrm{NLM}} \asymp (\log n)^{1/d} \mathcal{R}^{\mathrm{BO}}$ when $d \geq 2$.

Note that our bandwidth h is not infinite as in [33]. There, the authors use an infinite window for searching for matching patches: this is optimal in their setting because they consider piecewise *constant* images. In our setting, images are piecewise smooth, and a smaller bandwidth can not only help us reduce the risk of our estimate, but also lead to more computationally efficient estimation algorithms.

Is it possible to refine the analysis so that we get a better bound? We do not have a full answer to this question. It happens that the upper bound in Theorem 4.5 is sharp (except for the log factor) when $d = 1, 2$. Perhaps surprisingly, the situation is different in dimension $d \geq 3$ (see Section 8.2.5 after the proof of Theorem 4.5). Thus, within the cartoon model in dimension $d \leq 2$, NLM is comparable to (successful) variable bandwidth methods such as Lepski's [28] when the SNR is of order 1.

5. Performance analysis for thin features and textures. In the cartoon model of Section 2 with SNR of order 1, the performance of NLM is comparable to that of variable bandwidth

kernel smoothing, and actually that of wavelets as well [25, 40]. In natural images, however, NLM can perform substantially better. We explain this by the fact that the cartoon model we considered so far, though useful as a benchmark, does not account for some features that are common in natural images, and particularly, very thin regions a few pixels wide and regular textures.

Below, we do as if the image contained regions of cartoon type and regions with thin features and/or texture, and keep the same bandwidths that we found to be optimal in the cartoon model in the previous results.

5.1. Thin features. Both YF and NLM achieve a good performance on thin features. We focus on sample points within the feature and focus on the interesting case where the thickness is of smaller order of magnitude than the bandwidth h.

Theorem 5.1. *Consider $f \in \mathcal{F}^{\text{thin}}(\alpha, C_0, d_0, a)$ with band Ω; assume all parameters are fixed except $a \geq 4/n$ and $a \to 0$ as $n \to 0$. In terms of* point-wise risk (2.3) *at $x_i \in \Omega$, we have the following:*

1. *The linear filter with bandwidth h^{LF} has risk of order 1 if $a = o(h^{\text{LF}})$.*
2. *BO with maximal bandwidth h^{MO} has a point-wise risk of order $a^{2\alpha} \vee \sigma^2(na)^{-d}$, if $\operatorname{dist}(x_i, \Omega^c) \geq a/C$ for some $C > 3$, if $na \to \infty$.*
3. *MO with bandwidth h^{MO} has risk of order $(h^{\text{MO}}/a)^{d-d_0}\mathcal{R}^{\text{MO}}$ if $a = o(h^{\text{MO}})$.*
4. *The latter is true of YF with bandwidths $h^{\text{MO}}, h_y \asymp 1$, if the noise satisfies the conditions of Theorem 4.4.*
5. *This is also the case of NLM (Euclidean or Average) with bandwidths $h = h^{\text{MO}}, h_y = h_y^{\text{NLM}}$, and patch size $h_P = h_P^{\text{NLM}}$, if $\operatorname{dist}(x_i, \Omega^c) \geq h_P^{\text{NLM}}$.*

In view of this result, we can say that linear filtering essentially erases the feature. In contrast, YF still performs very well under low noise, and NLM performs well in this case and for higher noise settings. Note that when $h_P^{\text{NLM}} = o(a)$, the bound above holds for most points within the thin features. Though not stated here, we found that NLM is able to handle such bands under special circumstances — when $d \geq 3$ and the band is straight.

5.2. Regular patterns and textures. We consider very general patterns where YF will do as well as in the cartoon model, situations where most other methods are essentially useless. Euclidean NLM performs well too, under additional assumptions on the pattern.

Proposition 5.1. *Consider $f \in \mathcal{F}^{\text{pattern}}(\alpha, C_0, a)$ with all parameters are fixed except for a, which satisfies $a = o(h^{\text{MO}})$ and $na \to \infty$. Let $N_\Omega := \#\{i : x_i \in \Omega\}$ and N_{Ω^c} is defined similarly. If $N_\Omega \vee N_{\Omega^c} \leq C(N_\Omega \wedge N_{\Omega^c})$, with $C > 1$ fixed, we have the following:*

1. *MO with $h = h^{\text{MO}}$ achieves an MSE of order $\mathcal{R}^{\text{MO}}$.*
2. *The latter is true of YF with bandwidths $h^{\text{MO}}, h_y \asymp 1$, if the noise satisfies the conditions of Theorem 4.4.*
3. *Suppose in addition that for every $x_i \in \Omega$ and $x_j \in \Omega^c$,*

$$\|\mathbb{1}(P_i \cap \Omega) - \mathbb{1}(P_j \cap \Omega)\|_2^2 \geq (\sigma^2 \log n)/C', \tag{5.1}$$

for some $C' > 1$ fixed. Then NLM (Euclidean) with bandwidths $h = h^{\text{MO}}, h_y = h_y^{\text{NLM}}$ and patch size h_P^{NLM}, achieves an MSE of order $(na)^d\,\mathcal{R}^{\text{MO}}$.

To understand the condition (5.1), consider a simple binary image $g := \mathbb{1}_{\{\Omega\}}$. In order for NLM to perform effectively on f, then any two patches from g with one centered around a pixel in Ω and the other centered around a pixel in Ω^c must be sufficiently distinct – and the necessary degree of distinction increases with the noise level. See Figure 5.1 for an illustration of this concept. We mention that (5.1) is satisfied by such patterns as a chessboard or stripes. A regular pattern in a real image (*e.g.,* on Barbara's blouse) are often referred to as textures, and we just saw that NLM is able to effectively denoise such patterns under some regularity conditions. For random models of textures, such as Markov random fields, we do not expect NLM to do well unless the pattern is not very random. The reason is that few patches are close in Euclidean distance to a given patch.

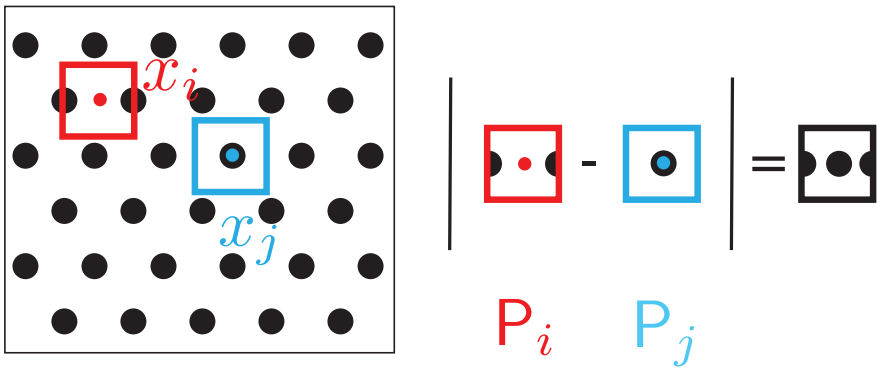
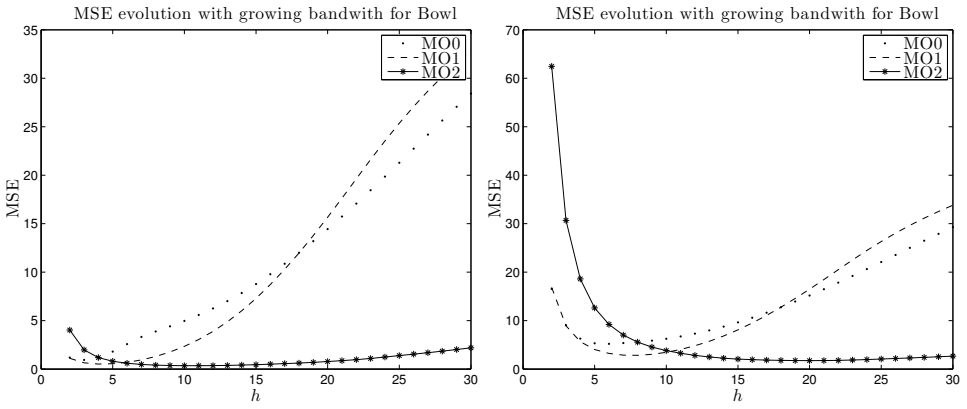


Figure 5.1. *Illustration of degree of discrepancy required between two patches, one centered at a point in Ω and one centered at a point in Ω^c.*

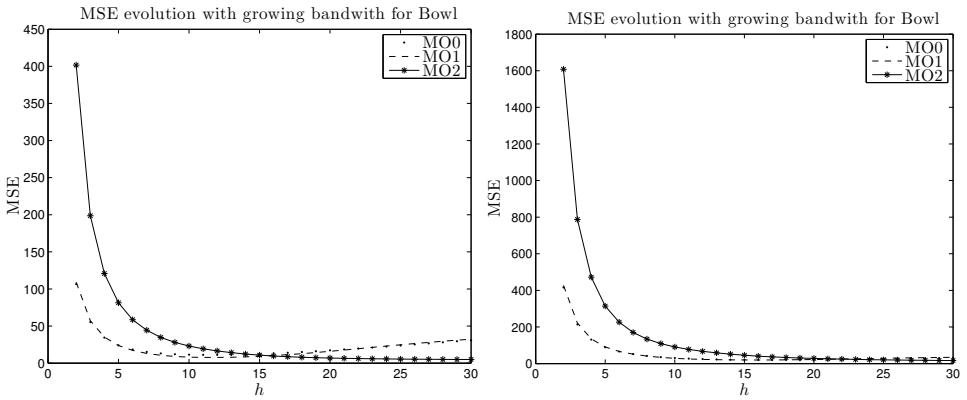
6. Experiments. In this section we provide numerical results for images with dimension $d = 2$, whose pixel intensities are between 0 and 255. In all our experiments, the noise is Gaussian with standard deviation $\sigma \in \{5, 20, 50, 100\}$ (Note that this corresponds, for normalized images in $[0, 1]$ to noise with $\sigma \in \{5/255, 10/255, 50/255, 100/255\}$).

On both toy and classical images, we have compared the behavior of the following methods: linear filtering (LF), Yaroslavsky's filter (YF), Euclidean non-local means (NLM), average non-local means (NLM-average) and the membership oracle (MO). In all cases we have implemented LPR version of the methods for the orders $r \in \{0, 1, 2\}$. Note that, as expected, for linear filtering LPR of order 0 and 1 are exactly identical because the support of the kernel is symmetric. However, for other methods the symmetry of the support is no longer guaranteed and the estimators differ. The higher order LPR versions are computed by solving the linear system in (8.5). A small numerical constant (10^{-8}) is added to the diagonal elements of $\mathbf{X}^T\mathbf{X}$ so that inverting this matrix is always a well conditioned problem.

For fair comparisons we have used the same (box) kernel for all the methods. The patch size is 7×7 (*i.e.,* $h_{\mathsf{P}} = 7$) and is kept fixed for all the methods, when patches are needed. For the spatial bandwidth h we have chosen to use the values obtained by considering the best h associated with the MO (on the Bowl image) for each noise level $\sigma \in \{5, 20, 50, 100\}$ and polynomial order $r \in \{0, 1, 2\}$ considered. Those values are provided by the MSE optimization in Figure 6.1 and given in Tab. 6.1. The photometric bandwidth h_y is chosen by hand, and



(a) $\sigma = 5$ (b) $\sigma = 20$



(c) $\sigma = 20$ (d) $\sigma = 100$

Figure 6.1. *Evolution of the MSE with respect to the searching zone for the image Bowl for four noise levels ($\sigma \in \{5, 20, 50, 100\}$).*

	$\sigma = 5$	$\sigma = 20$	$\sigma = 50$	$\sigma = 100$
$r = 0$	7	13	23	35
$r = 1$	9	17	25	33
$r = 2$	23	41	59	61

Table 6.1

*This table provides the spatial bandwidth h used in experiments. Those are the one minimizing MSE for the MO on the Bowl image (*cf. *Figure 6.1).*

differs from method to method: $\sqrt{10}\sigma$ (YF), $2h_{\mathsf{P}}\sigma$ (NLM-average), $\sqrt{3.5}h_{\mathsf{P}}\sigma$ (NLM), 30 (MO).

In practice, h needs to be larger for larger values of r in order to ensure that the local polynomial regression is stable. Note that there are $q = \binom{r+d}{d}$ polynomial coefficients in each search window. If we apply the rule of thumb of 10 observations per unknown parameter, the search window needs to include about $10q$ pixels. This is illustrated in Figure 6.1, where we see the optimal h increasing with r.

Since for natural (not cartoon-like) images the *membership oracle* is irrelevant, we have used a modified YF oracle instead. This oracle has access to the original image to compute the

weights as in (3.3), and then performs LPR on the noisy pixel values with these weights. Note that in the cases of piecewise constant images, this coincides exactly with the *membership oracle* as soon as the bandwidth is large enough.

The experiments conducted show that the linear filter is always outperformed in practice by the YF, the NLM and the NLM-average. For low noise level ($\sigma = 5$), the YF with $r = 2$ outperforms all the other methods (*cf.* for instance Figure 6.6 and Table 6.2). However, in the presence of strong noise the NLM and the NLM-average are the clear winners both on most images. Interestingly the NLM-average can even improve on the NLM for very strong noise. On the other hand, one can see that the NLM-averagemimics the behavior of the linear filter for textured images with strong noise (*cf.* Figure 6.9), due to the fact that different side of periodic features are averaged together.

Figure 6.2, Figure 6.3, Figure 6.4, and Figure 6.5 demonstrate the importance of the jump parameter μ in practice. On the left end of the Swoosh, the jump is larger and we reconstruct it accurately across all noise levels. On the right end, the jump is smaller and the performance degrades with σ, exactly as predicted by our theory.

In the spirit of reproducible research the MATLAB code used for those results are available on the authors' webpages.

7. Discussion. As described in the Introduction, the bounds described in this paper and in the independent work [33, 32] address *fundamental performance limits of NLM and related photometric image filtering methods*. These methods have an established history of strong empirical performance on natural images, but until now little was known about how these methods performed asymptotically, especially with respect to related methods based on computational harmonic analysis (*e.g.,* wavelet or curvelet denoising).

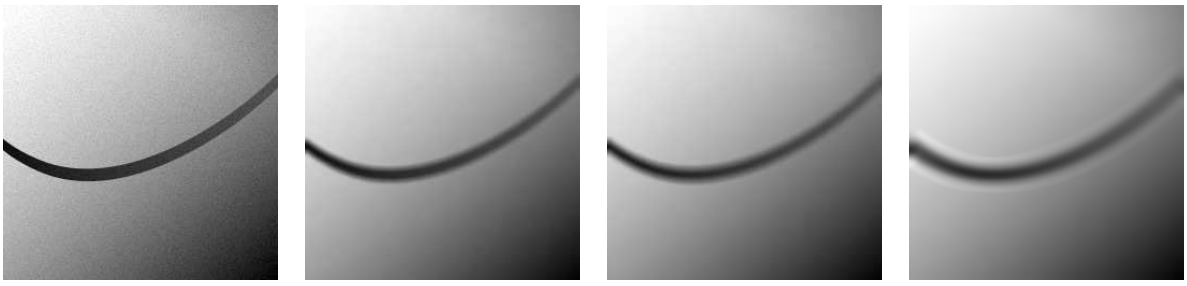
Both our bounds and the bounds in [33] suggest that NLM has some limitations for piecewise smooth images. In fact, our analysis shows (perhaps surprisingly) that YF can significantly outperform NLM at low noise levels. This fact can be explained by noting that YF is a special case of NLM with a patch width $h_{\mathsf{P}} = 1$, and the performance of NLM hinges upon our ability to measure the similarity of two patches based on noisy observations. In low noise, this similarity can already be estimated quite accurately with $h_{\mathsf{P}} = 1$. In stronger noise, the similarity measured through larger patches is more robust to noise — but larger patches also introduce some bias.

There are several distinctions between the work in this paper and the closely related work in [33] that bear mentioning. First, we consider the *cartoon model*, where the functions are piecewise Hölder smooth images with a discontinuity set corresponding to a Lipschitz mapping of the unit ball, while Maleki *et al.* [33] consider the *horizon edge model*, where the functions are piecewise constant with a discontinuity set corresponding to the graph of a Lipschitz function. Though they actually consider smoother edges, their analysis reduces to the case of Lipschitz smoothness. We consider images in arbitrary dimension — showing that NLM behaves differently when $d \geq 3$ — while Maleki *et al.* consider the case of 2D images ($d = 2$). Because they consider functions that are piecewise constant, they use the weighted average version (3.1) without spatial localization (*i.e.,* $h = \infty$). Because we focus on smooth — not necessarily constant — regions, we need to localize both YF and NLM. Applying the more complex LPR (3.2) enables YF to successfully adapt to the degree of regularity in each smooth

	Blob	Sinusoid	Bowl	Ridges	Stripes	Barbara	Cam.
				$\sigma = 5$			
LF0	35.27	40.29	57.71	48.80	21077.49	408.13	437.96
LF1	48.50	55.50	74.18	110.08	22787.90	473.15	529.64
LF2	72.11	82.89	105.09	246.70	15424.99	586.47	663.59
YF0	1.45	1.67	1.74	13.70	1.68	20.48	13.59
YF1	1.11	1.30	1.01	8.57	1.21	20.03	13.39
YF2	0.94	1.17	0.87	7.94	0.69	20.79	13.48
NLM-Av.0	1.61	2.55	3.30	4.15	327.70	255.23	188.74
NLM-Av.1	1.49	1.79	2.33	4.08	202.88	202.96	99.13
NLM-Av.2	1.35	1.67	1.99	3.45	329.83	242.91	151.60
NLM0	1.55	1.74	1.86	3.88	3.89	19.54	13.52
NLM1	1.47	1.55	1.73	3.65	3.11	19.79	13.72
NLM2	1.51	1.51	1.59	3.59	1.68	20.27	13.74
MO0	1.58	1.77	0.97	16.16	1.23	35.10	28.60
MO1	1.11	1.26	0.52	19.13	0.81	37.84	28.99
MO2	0.97	1.31	0.35	12.41	0.36	39.65	29.75
				$\sigma = 20$			
LF0	77.77	91.66	109.58	305.04	13956.38	607.11	684.54
LF1	104.25	134.27	141.41	533.25	14455.88	725.30	818.62
LF2	139.80	208.44	188.67	707.47	16996.42	901.23	994.78
YF0	15.43	17.05	11.61	118.99	9.77	189.69	104.70
YF1	17.93	20.80	11.46	158.55	7.81	219.63	113.33
YF2	18.97	24.66	14.34	174.83	6.59	242.48	122.42
NLM-Av.0	6.99	9.27	17.76	18.73	332.71	345.66	307.67
NLM-Av.1	6.02	7.72	14.12	19.61	406.11	334.80	275.87
NLM-Av.2	4.58	7.53	13.05	15.15	399.46	352.14	306.27
NLM0	5.76	6.37	12.66	20.44	11.54	121.17	92.36
NLM1	5.53	5.70	13.28	21.68	9.31	129.10	96.88
NLM2	5.02	4.53	13.16	19.44	5.92	137.95	101.09
MO0	4.00	4.41	5.03	31.78	4.65	41.67	34.24
MO1	2.96	3.25	2.82	35.75	2.89	44.92	34.06
MO2	2.26	2.74	1.88	33.60	1.83	45.68	34.88
				$\sigma = 50$			
LF0	149.70	211.28	195.46	847.84	17633.15	900.24	997.56
LF1	162.93	232.97	211.39	939.06	15081.34	955.73	1048.22
LF2	209.56	290.03	273.74	1501.85	15705.82	1157.37	1221.38
YF0	112.17	138.30	146.13	591.42	857.69	652.97	523.88
YF1	129.07	155.85	164.73	655.67	722.93	699.37	574.87
YF2	146.05	178.84	199.35	998.40	741.85	811.85	629.87
NLM-Av.0	23.66	29.89	52.96	64.32	807.78	419.12	389.60
NLM-Av.1	21.51	27.56	36.86	69.56	770.17	414.81	372.68
NLM-Av.2	18.17	26.07	39.07	67.12	820.21	425.60	385.13
NLM0	21.64	27.35	36.32	162.17	40.92	367.48	230.35
NLM1	29.09	31.35	30.78	179.78	25.50	381.14	234.01
NLM2	25.33	30.60	30.15	245.86	20.72	398.52	243.81
MO0	7.68	8.32	11.23	48.99	10.90	50.50	42.64
MO1	7.72	7.98	9.20	57.78	8.70	55.67	44.46
MO2	5.56	6.11	6.43	67.90	6.01	49.81	41.51
				$\sigma = 100$			
LF0	239.01	319.36	300.81	1340.28	17131.55	1198.68	1249.50
LF1	225.89	307.90	285.37	1277.60	17776.32	1159.50	1218.33
LF2	225.36	305.06	291.83	1550.89	17079.65	1188.50	1248.76
YF0	308.15	367.90	365.84	1206.35	8848.92	1108.23	1080.66
YF1	299.59	359.21	352.73	1156.57	9197.93	1077.16	1064.45
YF2	296.39	355.73	356.40	1375.29	8813.04	1099.59	1077.78
NLM-Av.0	64.41	76.19	118.55	202.15	8223.43	556.50	495.62
NLM-Av.1	66.78	74.22	94.12	204.95	8385.63	554.88	492.05
NLM-Av.2	66.27	73.42	98.31	224.59	8118.55	555.58	495.58
NLM0	91.67	131.36	167.44	819.97	91.49	911.60	628.08
NLM1	118.08	135.37	183.32	786.01	90.29	926.67	662.68
NLM2	101.83	127.34	171.13	956.96	88.01	918.08	646.76
MO0	14.19	15.12	22.74	80.33	19.90	61.09	54.41
MO1	18.31	17.96	23.06	91.72	22.38	76.21	65.36
MO2	17.76	17.65	18.95	88.60	22.09	72.00	62.53

Table 6.2

MSE comparisons of the denoising methods considered for LPR of order 0, 1 and 2. The compared methods are the Linear Filter (LF), the Yaroslavsky Filter (YF), the NLM-average (NLM-Av.), the classical NLM and the Membership Oracle (MO). Results are averaged over 5 Gaussian noise replicas.



(a) Noisy, MSE = 2.50e+01

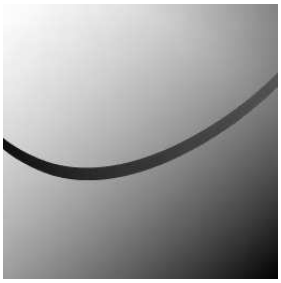
(b) LF0, MSE = 4.03e+01

(c) LF1, MSE = 5.55e+01

(d) LF2, MSE = 8.29e+01

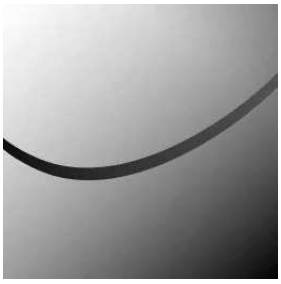
(e) YF0, MSE = 1.69e+00

(f) YF1, MSE = 1.31e+00

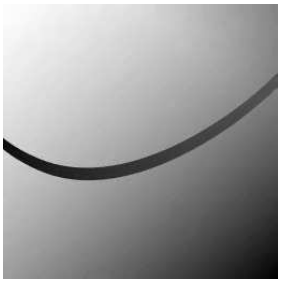


(g) YF2, MSE = 1.19e+00

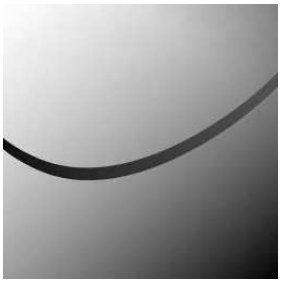
(h) NLM-Av.0, MSE = 2.56e+00



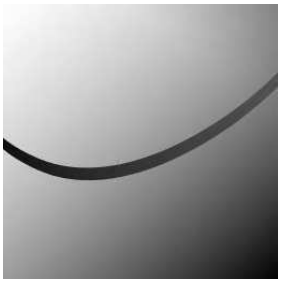
(i) NLM-Av.1, MSE = 1.90e+00



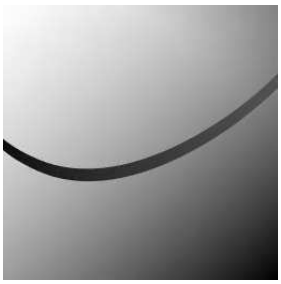
(j) NLM-Av.2, MSE = 1.75e+00



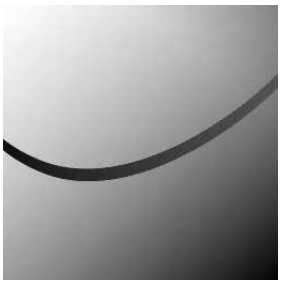
(k) NLM0, MSE = 1.54e+00



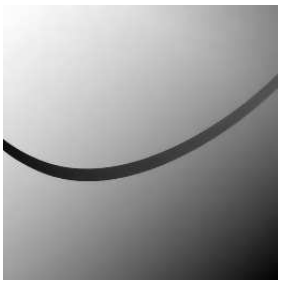
(l) NLM1, MSE = 1.36e+00



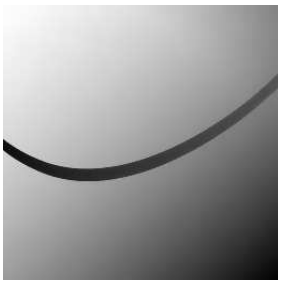
(m) NLM2, MSE = 1.37e+00



(n) MO0, MSE = 1.79e+00

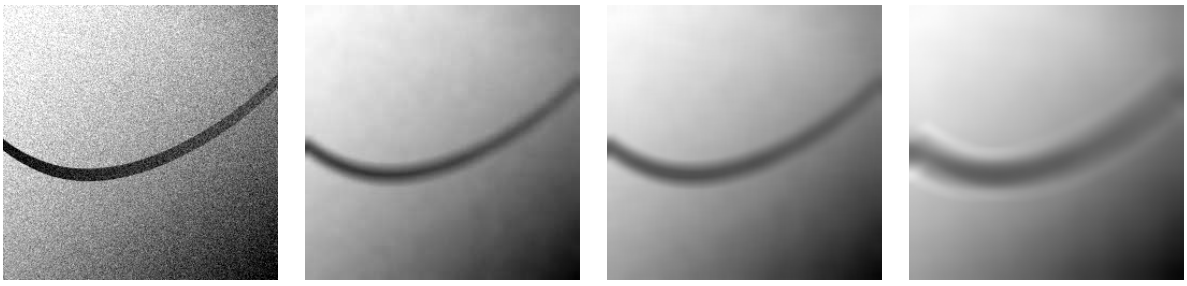


(o) MO1, MSE = 1.26e+00

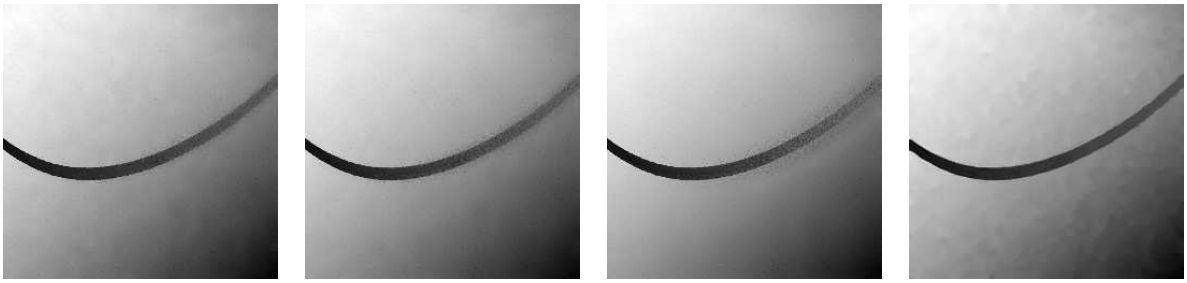


(p) MO2, MSE = 1.31e+00

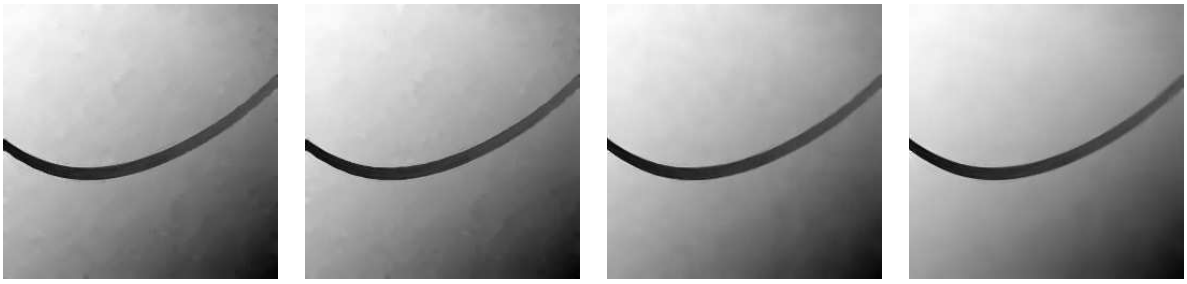
Figure 6.2. *Toy thin feature image (Swoosh) corrupted Gaussian noise with* $\sigma = 5$.



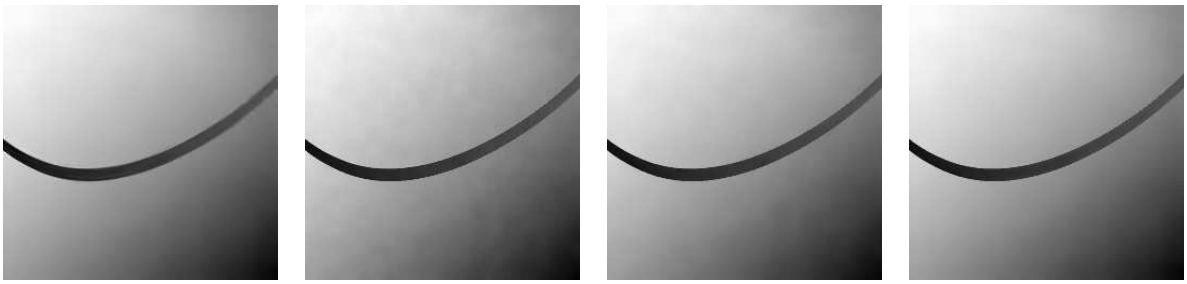
(a) Noisy, MSE = 3.99e+02 (b) LF0, MSE = 9.18e+01 (c) LF1, MSE = 1.35e+02 (d) LF2, MSE = 2.09e+02



(e) YF0, MSE = 1.71e+01 (f) YF1, MSE = 2.08e+01 (g) YF2, MSE = 2.49e+01 (h) NLM-Av.0, MSE = 9.86e+00

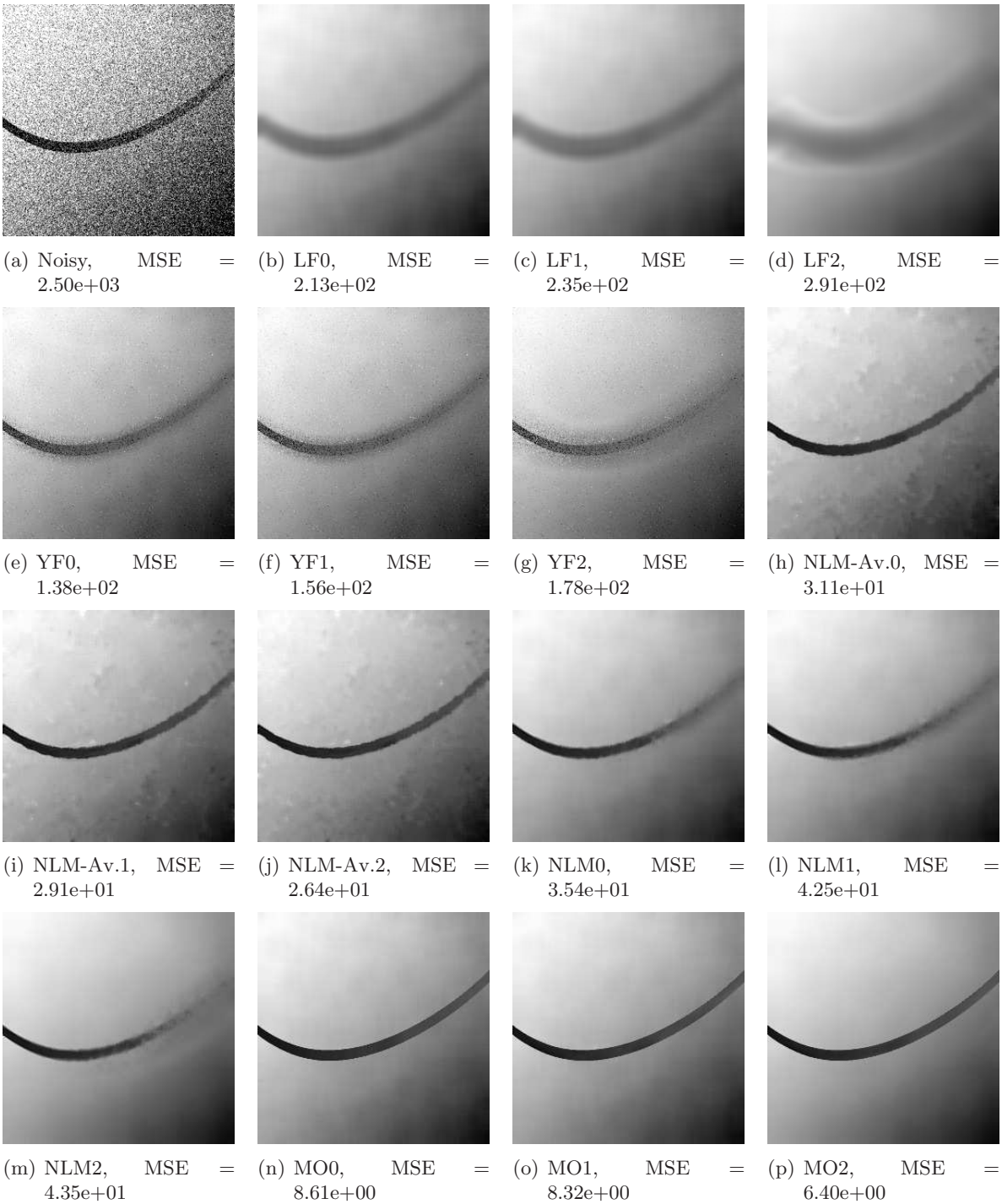


(i) NLM-Av.1, MSE = 8.37e+00 (j) NLM-Av.2, MSE = 7.87e+00 (k) NLM0, MSE = 6.19e+00 (l) NLM1, MSE = 6.00e+00



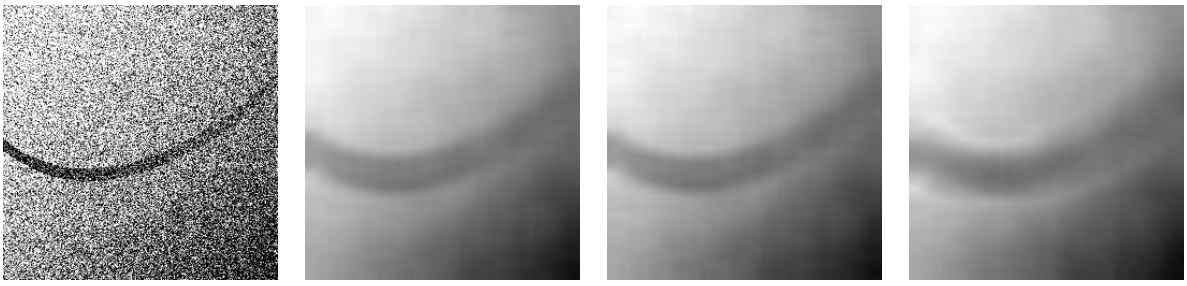
(m) NLM2, MSE = 4.53e+00 (n) MO0, MSE = 4.54e+00 (o) MO1, MSE = 3.39e+00 (p) MO2, MSE = 2.85e+00

Figure 6.3. *Toy thin feature image (Swoosh) corrupted Gaussian noise with* $\sigma = 20$.



(a) Noisy, MSE = 2.50e+03
(b) LF0, MSE = 2.13e+02
(c) LF1, MSE = 2.35e+02
(d) LF2, MSE = 2.91e+02
(e) YF0, MSE = 1.38e+02
(f) YF1, MSE = 1.56e+02
(g) YF2, MSE = 1.78e+02
(h) NLM-Av.0, MSE = 3.11e+01
(i) NLM-Av.1, MSE = 2.91e+01
(j) NLM-Av.2, MSE = 2.64e+01
(k) NLM0, MSE = 3.54e+01
(l) NLM1, MSE = 4.25e+01
(m) NLM2, MSE = 4.35e+01
(n) MO0, MSE = 8.61e+00
(o) MO1, MSE = 8.32e+00
(p) MO2, MSE = 6.40e+00

Figure 6.4. *Toy thin feature image (Swoosh) corrupted Gaussian noise with* $\sigma = 50$.



(a) Noisy, MSE = 9.98e+03

(b) LF0, MSE = 3.22e+02

(c) LF1, MSE = 3.11e+02

(d) LF2, MSE = 3.07e+02

(e) YF0, MSE = 3.67e+02

(f) YF1, MSE = 3.58e+02

(g) YF2, MSE = 3.54e+02

(h) NLM-Av.0, MSE = 8.55e+01

(i) NLM-Av.1, MSE = 8.27e+01

(j) NLM-Av.2, MSE = 8.11e+01

(k) NLM0, MSE = 1.78e+02

(l) NLM1, MSE = 1.78e+02

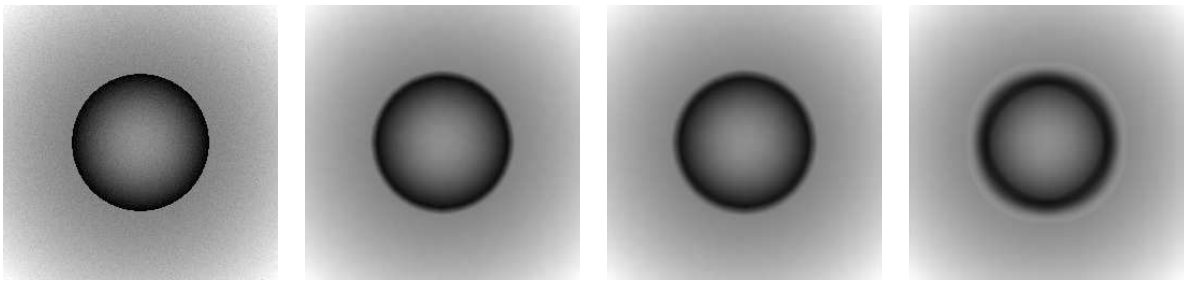
(m) NLM2, MSE = 1.68e+02

(n) MO0, MSE = 1.51e+01

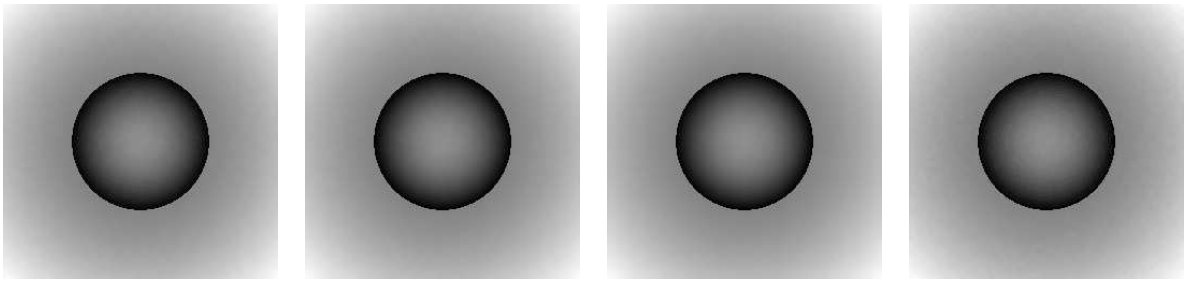
(o) MO1, MSE = 1.82e+01

(p) MO2, MSE = 1.86e+01

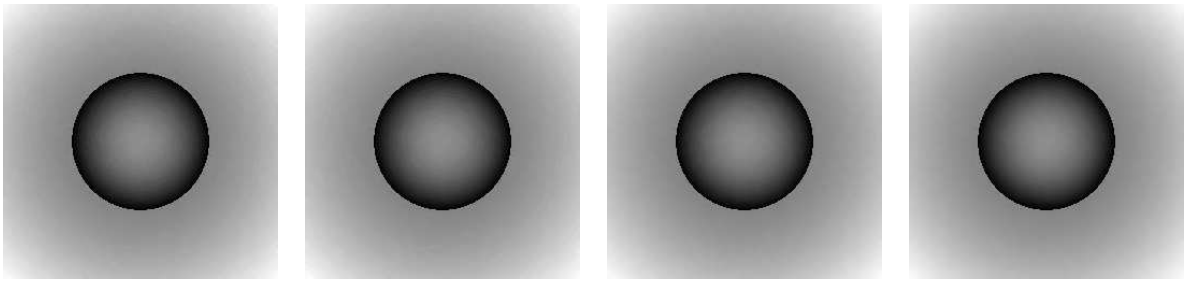
Figure 6.5. *Toy thin feature image (Swoosh) corrupted Gaussian noise with* $\sigma = 100$.



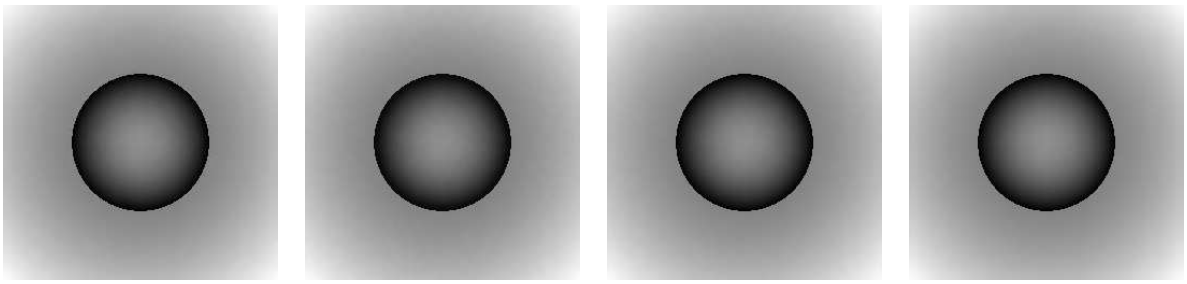
(a) Noisy, MSE = 2.50e+01 (b) LF0, MSE = 5.76e+01 (c) LF1, MSE = 7.40e+01 (d) LF2, MSE = 1.05e+02



(e) YF0, MSE = 1.71e+00 (f) YF1, MSE = 9.88e-01 (g) YF2, MSE = 8.74e-01 (h) NLM-Av.0, MSE = 3.41e+00

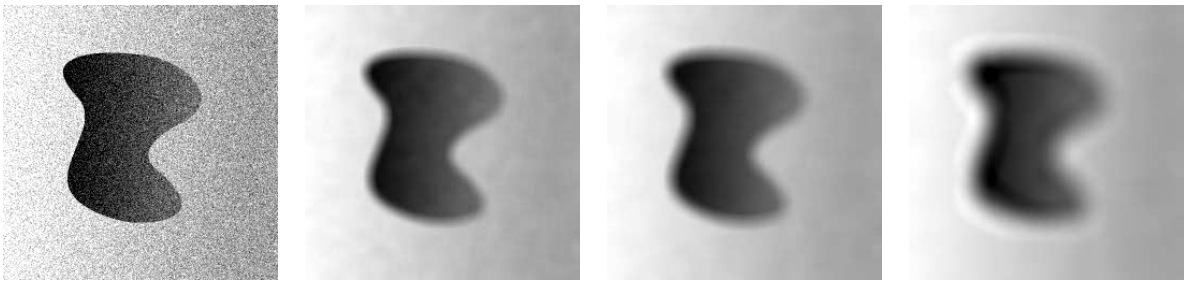


(i) NLM-Av.1, MSE = 2.50e+00 (j) NLM-Av.2, MSE = 2.15e+00 (k) NLM0, MSE = 1.67e+00 (l) NLM1, MSE = 1.60e+00

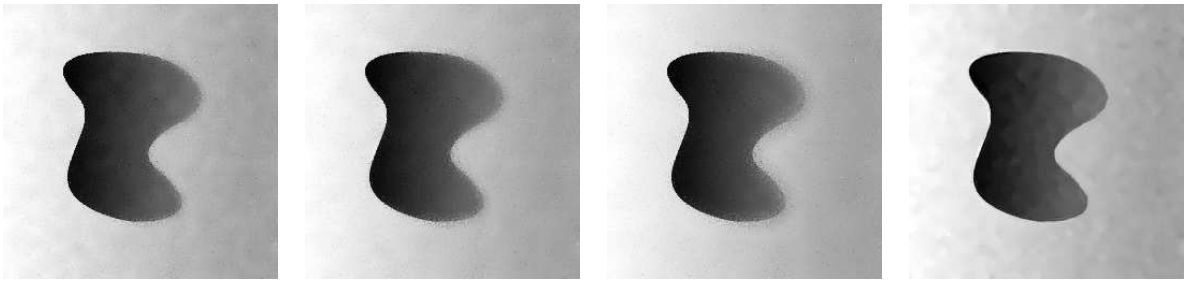


(m) NLM2, MSE = 1.50e+00 (n) MO0, MSE = 9.56e-01 (o) MO1, MSE = 5.01e-01 (p) MO2, MSE = 3.48e-01

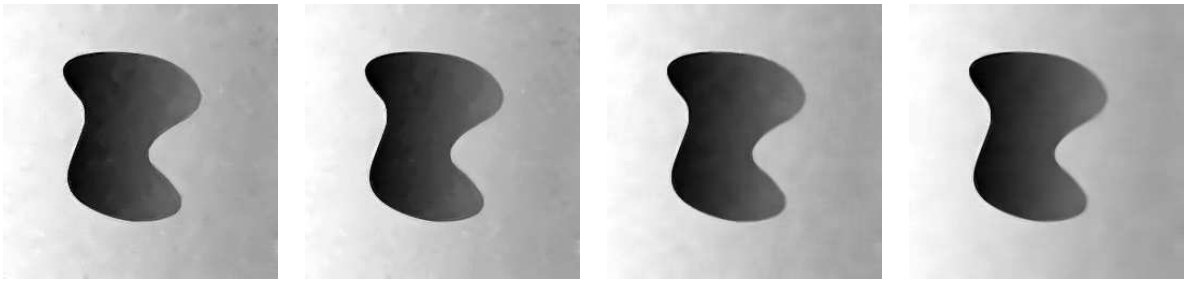
Figure 6.6. *Toy cartoon image (Bowl) corrupted Gaussian noise with* $\sigma = 5$.



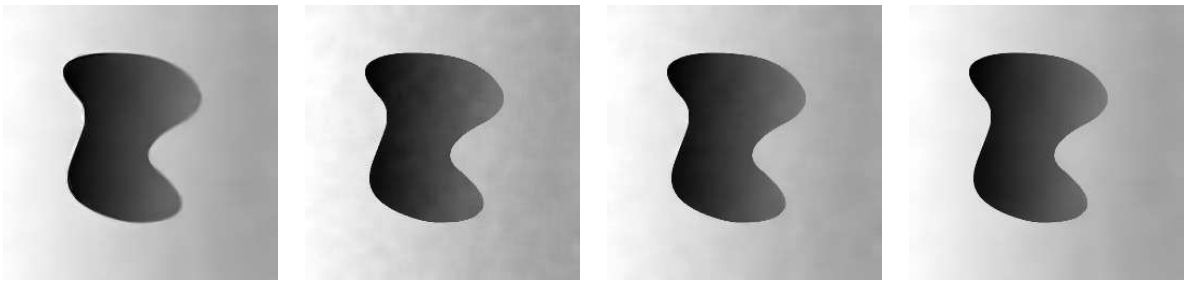
(a) Noisy, MSE = 3.99e+02 (b) LF0, MSE = 7.78e+01 (c) LF1, MSE = 1.04e+02 (d) LF2, MSE = 1.40e+02



(e) YF0, MSE = 1.53e+01 (f) YF1, MSE = 1.80e+01 (g) YF2, MSE = 1.90e+01 (h) NLM-Av.0, MSE = 7.55e+00



(i) NLM-Av.1, MSE = 6.62e+00 (j) NLM-Av.2, MSE = 5.10e+00 (k) NLM0, MSE = 5.86e+00 (l) NLM1, MSE = 5.83e+00

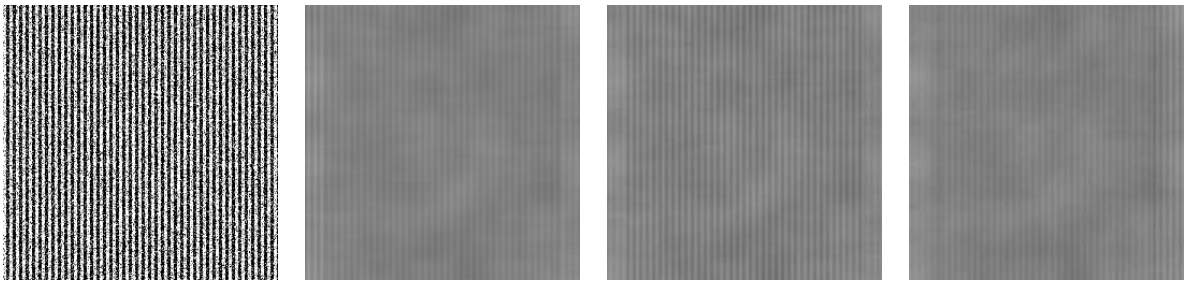


(m) NLM2, MSE = 4.92e+00 (n) MO0, MSE = 4.12e+00 (o) MO1, MSE = 3.03e+00 (p) MO2, MSE = 2.37e+00

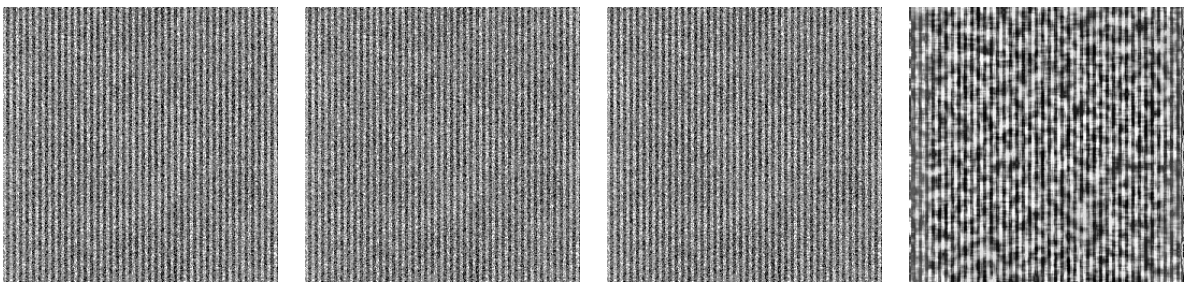
Figure 6.7. *Toy cartoon image (Blob) corrupted Gaussian noise with* $\sigma = 20$.

(a) Noisy, MSE = 2.50e+03
(b) LF0, MSE = 9.00e+02
(c) LF1, MSE = 9.55e+02
(d) LF2, MSE = 1.15e+03
(e) YF0, MSE = 6.53e+02
(f) YF1, MSE = 6.99e+02
(g) YF2, MSE = 8.10e+02
(h) NLM-Av.0, MSE = 4.20e+02
(i) NLM-Av.1, MSE = 4.17e+02
(j) NLM-Av.2, MSE = 4.26e+02
(k) NLM0, MSE = 4.50e+02
(l) NLM1, MSE = 4.76e+02
(m) NLM2, MSE = 4.96e+02
(n) MO0, MSE = 5.28e+01
(o) MO1, MSE = 5.86e+01
(p) MO2, MSE = 5.18e+01

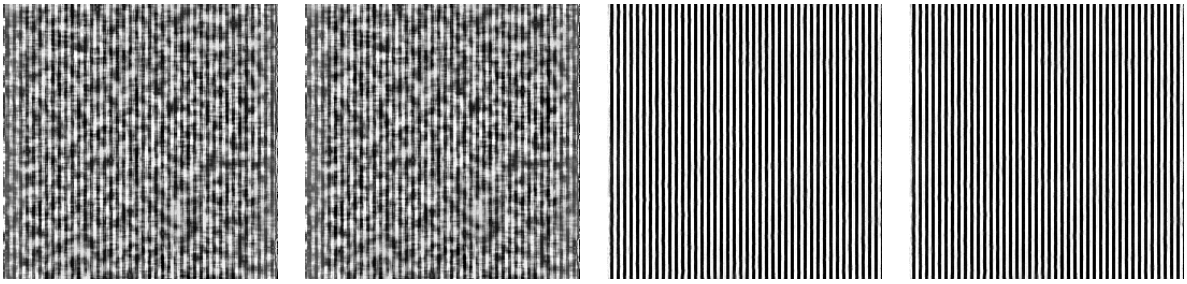
Figure 6.8. *Barbara image corrupted Gaussian noise with* $\sigma = 50$.



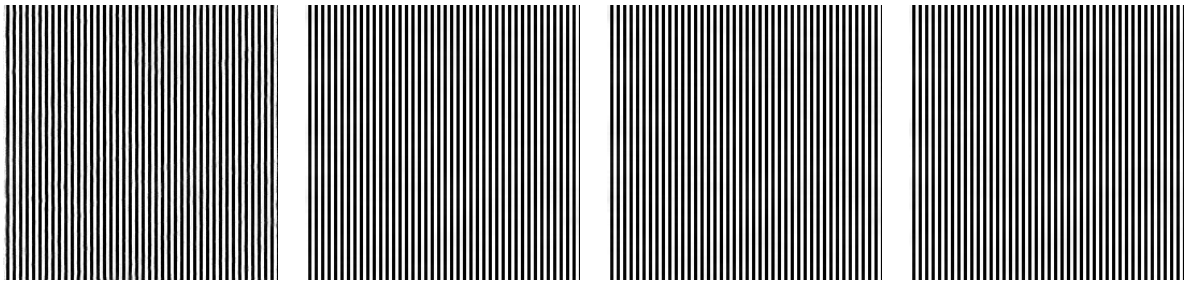
(a) Noisy, MSE = 9.98e+03 (b) LF0, MSE = 1.71e+04 (c) LF1, MSE = 1.78e+04 (d) LF2, MSE = 1.71e+04



(e) YF0, MSE = 8.91e+03 (f) YF1, MSE = 9.26e+03 (g) YF2, MSE = 8.87e+03 (h) NLM-Av.0, MSE = 7.39e+03



(i) NLM-Av.1, MSE = 7.51e+03 (j) NLM-Av.2, MSE = 7.25e+03 (k) NLM0, MSE = 2.42e+02 (l) NLM1, MSE = 2.10e+02



(m) NLM2, MSE = 2.33e+02 (n) MO0, MSE = 2.11e+01 (o) MO1, MSE = 2.37e+01 (p) MO2, MSE = 2.28e+01

Figure 6.9. *Toy texture image (Stripes) corrupted Gaussian noise with* $\sigma = 100$.

region. We note that Maleki et al. [33] do not consider the case of low noise and simply show that YF achieves the same performance as LF — which is also our conclusion in strong noise.

Also, to simplify the analysis, Maleki *et al.* consider a slightly modified version of (Euclidean) NLM and derive lower bounds for oracles versions of YF and NLM. The lower bounds for NLM were also challenging for us and we provide heuristics. We mention that our results imply that the simpler NLM-average achieves the same performance as NLM in the horizon model.

Our analysis of NLM for our image classes with thin features or regular patterns is also a significant novel aspect of our work. For horizon and cartoon image classes, it is already known that curvelets and related transforms yield near-minimax optimal performance guarantees — thus using NLM in this setting might seem superfluous. However, NLM has a significant empirical advantage over curvelets for certain kinds of repeating textures. We develop a model for images with these features, and note that it does not approach cartoon or horizon model asymptotically. For this image class, we demonstrate that NLM performs asymptotically as well as it does for the cartoon class.

The current bounds are based on ideal bandwidths which depend on the unknown smoothness parameter α. Thus we have demonstrated that the adaptive filtering techniques considered adapt to the discontinuity Ω, but not to α. We anticipate that adaptivity to α is indeed possible and leave that analysis for future work.

Last but not least, we provide extensive numerical experiments illustrating the various points we make, thus bridging theory with practice.

8. Proofs.

In this section, $C, C_1, C_2, \dots$ denote finite positive constants that do not change with n and whose actual value may change with each appearance.

8.1. Preliminary results.

We first gather some basic results.

8.1.1. Some analysis.

Functions in $\mathcal{H}_d(\alpha, C_0)$ are uniformly well-approximated locally by polynomials of degree $\lfloor \alpha \rfloor$, specifically their Taylor expansions. For $g \in \mathcal{H}_d(\alpha, C_0)$ and $x \in [0,1]^d$, the Taylor expansion of g at x of degree $t \in \mathbb{N}$ is defined as follows:

$$T_x^r g(x') = \sum_{|s| \le t} g^{(s)}(x) \prod_{i=1}^d \frac{(x_i' - x_i)^{s_i}}{s_i!}.$$

Lemma 8.1. *For any $g \in \mathcal{H}_d(\alpha, C_0)$,*

$$|g(x') - T_x^{\lfloor \alpha \rfloor} g(x')| \le c_\alpha C_0 \|x' - x\|_\infty^\alpha, \quad \forall x, x' \in [0,1]^d,$$

where

$$c_\alpha := \sum_{s \in \mathbb{N}^d : |s| = \lfloor \alpha \rfloor} \frac{1}{s_1! \cdots s_d!}.$$

Proof. Though this sort of result is well-known, we provide a proof for completeness. A Taylor approximation of degree $\lfloor \alpha \rfloor$ gives:

$$g(x') = T_x^{\lfloor \alpha \rfloor} g(x') + \sum_{|s| = \lfloor \alpha \rfloor} (g^{(s)}(z) - g^{(s)}(x)) \prod_{i=1}^d \frac{(x_i' - x_i)^{s_i}}{s_i!},$$

for some z on the segment joining x and x'. Hence,

$$|g(x') - T_x^{\lfloor\alpha\rfloor} g(x')| \le c_\alpha \|x' - x\|_\infty^{\lfloor\alpha\rfloor} \max_{|s|=\lfloor\alpha\rfloor} |g^{(s)}(z) - g^{(s)}(x)|.$$

Now apply (2.5) and the fact that $\|z - x\|_\infty \le \|x' - x\|_\infty$ to get

$$|g^{(s)}(z) - g^{(s)}(x)| \le C_0 \|x' - x\|_\infty^{\alpha - \lfloor\alpha\rfloor}, \quad \forall s \in \mathbb{N}^d, |s| = \lfloor\alpha\rfloor.$$

■

8.1.2. Some geometry. For a measurable set $A \subset \mathbb{R}^d$,

$$\rho(A) := \inf_{h\in(0,1)} \inf_{x\in A} \sup \left\{ \frac{\mathrm{Vol}(B(y,s))}{\mathrm{Vol}(B(x,h))} : B(y,s) \subset B(x,h) \cap A \right\}. \tag{8.1}$$

The following lemma bounds ρ from below for sets whose boundary is sufficiently regular.

Lemma 8.2. *Let $\phi : \mathbb{R}^d \to \mathbb{R}^d$ be injective, with ϕ and ϕ^{-1} both C-Lipschitz. Then for $\Omega = \phi(B(0,1))$ and ρ defined in (8.1), we have $\min(\rho(\Omega), \rho(\Omega^c)) \ge (2C)^{-d}$.*

Proof. Fix $x \in \Omega$ and $h > 0$. Since ϕ is Lipschitz with constant C, we have $\phi(B(\phi^{-1}(x), h/C)) \subset B(x,h)$. Note that $z := \phi^{-1}(x) \in B(0,1)$ and, by the triangle inequality, $B(z, h/C) \cap B(0,1) \supset B(z', t)$, where $z' := (1 - h/(2C))z$ and $t := h/(2C)$. Because ϕ^{-1} is C-Lipschitz, we have $\phi^{-1}(B(\phi(z'), t/C)) \subset B(z', t)$, so that

$$B(y,s) \subset \phi(B(z',t)) \subset \phi(B(z, h/C) \cap B(0,1)) \subset B(x,h) \cap \Omega,$$

where $y := \phi(z')$ and $s := t/C$. We obtain a lower bound for Ω^c in a similar way. ■

Next is a result on the number of sample points within a certain distance of a subset. Let $\mathcal{X}_{n^d}$ be the set of sample points, that is, $\mathcal{X}_{n^d} = \{x_i : i \in I_n^d\}$.

Lemma 8.3. *For any subset $A \subset (0,1)^d$ of the form $A = B(A', \eta)$ for some $A' \subset (0,1)^d$ and $4/n \le \eta \le 1$,*

$$8^{-d} n^d \mathrm{Vol}(A) \le |A \cap \mathcal{X}_{n^d}| \le 4^d n^d \mathrm{Vol}(A).$$

Proof. Let $z_1, \dots, z_k \in (0,1)^d$ be a maximal η-packing for A' (*i.e.,* the balls $B(z_j, \eta/2)$ for $j = 1, \dots, k$ are disjoint and included in $A' \subset A$, and for any $z \in A'$, there is j such that $z \in B(z_j, \eta)$). By the triangle inequality, we have

$$\bigcup_{j=1,\dots,k} B(z_j, \eta/2) \subset A \subset \bigcup_{j=1,\dots,k} B(z_j, 2\eta).$$

On the one hand, taking volumes on all sides, we get $k\eta^d \le \mathrm{Vol}(A) \le k 2^d (2\eta)^d$, since the unit $(\|\cdot\|_\infty)$ ball has volume 2^d. This turns into $4^{-d}\mathrm{Vol}(A) \le k\eta^d \le \mathrm{Vol}(A)$. On the other hand, counting sample points on all sides, using the fact that

$$\eta^d n^d \le |B(z,\eta) \cap \mathcal{X}_{n^d}| \le (2\eta)^d n^d, \quad \forall z \in (0,1)^d, \ \forall \eta \in (2/n, 1),$$

we get

$$k(\eta/2)^d n^d \leq \sum_{j=1}^{k} |B(z_j, \eta/2) \cap \mathcal{X}_{n^d}| \leq |A \cap \mathcal{X}_{n^d}| \leq \sum_{j=1}^{k} |B(z_j, 2\eta) \cap \mathcal{X}_{n^d}| \leq k(4\eta)^d n^d.$$

Combining these, we get the desired result. ■

Lemma 8.4. *Suppose $1 \leq d_0 \leq d$ are integers and let $\phi : \mathbb{R}^{d_0} \to \mathbb{R}^d$ be injective, with ϕ and ϕ^{-1} (on the range of ϕ) both C-Lipschitz with $C \geq 1$. Then there is another constant $C' > 1$ such that, for $A := \phi((0,a)^{d_0})$ and $h \in (0,1)$,*

$$\frac{1}{C'} a^{d_0} h^{d-d_0} \leq \mathrm{Vol}(B(A,h)) \leq C' a^{d_0} h^{d-d_0}.$$

Consequently, if $\phi : \mathbb{R}^d \to \mathbb{R}^d$ is as above and $A := \phi(\partial B(0,1))$, the same the result holds with $d - d_0 = 1$.

Proof. We first observe that, for any $z \in \mathbb{R}^{d_0}$ and $h > 0$, since ϕ is C-Lipschitz,

$$\phi(B(z,h)) \subset B(\phi(z), Ch). \tag{8.2}$$

Now, let $z_1, \ldots, z_m$ denote a maximal h-packing of $(0,a)^{d_0}$. Note that $m \asymp (a/h)^{d_0}$ when $h \leq 1$. By definition $\|z_i - z_j\| \geq h$, so that $\|\phi(z_i) - \phi(z_j)\| \geq h/C$ since ϕ^{-1} is C-Lipschitz with $C \geq 1$. Hence,

$$\bigsqcup_{i=1,\ldots,k} B(\phi(z_i), h/C) \subset B(A, h/C) \subset B(A,h),$$

implying

$$\sum_{i=1}^{m} \mathrm{Vol}(B(\phi(z_i), h/C)) \leq \mathrm{Vol}(B(A,h)).$$

We then conclude by the fact that $\sum_{i=1}^m \mathrm{Vol}(B(\phi(z_i), h/C)) \asymp mh^d \asymp a^{d_0} h^{d-d_0}$. For the upper bound, we use the fact that $(0,a)^{d_0} \subset \cup_{i=1,\ldots,k} B(z_i, h)$, so that

$$A \subset \bigcup_{i=1,\ldots,k} \phi(B(z_i,h)) \subset \bigcup_{i=1,\ldots,k} B(\phi(z_i), Ch),$$

by (8.2). Hence, using the triangle inequality,

$$\mathrm{Vol}(B(A,h)) \leq \sum_{i=1}^{m} \mathrm{Vol}(B(\phi(z_i), Ch + h)) \asymp mh^d \asymp a^{d_0} h^{d-d_0}.$$

For the second part, we use the fact that $\partial B(0,1) = \cup_\ell \phi_\ell((0,1)^{d-1})$ for a finite set of functions ϕ_ℓ satisfying the requirements and the fact that the composition $\phi \circ \phi_\ell$ is also Lipschitz. ■

8.1.3. Some statistics. We establish here some bounds on the point-wise MSE (2.3) of LPR (3.2). We mention that much finer results exist in dimension $d = 1$ for the case where the underlying function f is smooth; see [17] and references therein.

Lemma 8.5 (Variance).

For any sufficiently large constant $C > 0$, depending only on d, r, the following is true. Consider the LPR estimator of the form (3.2), with weights $\omega_{i,j} \in \{0,1\}$. Assume that $B_i^{\text{in}} \subset A_i := \{j : \omega_{i,j} = 1\} \subset B_{i,h} := \{j : x_j \in B(x_i, h)\}$, for some discrete ball B_i^{in} satisfying $|B_i^{\text{in}}| \geq |B_{i,h}|/C$ for some constant C.

$$\frac{1}{C}\sigma^2(nh)^{-d} \leq \operatorname{Var}(\widehat{f}_i) \leq C\sigma^2(nh)^{-d}. \tag{8.3}$$

Proof. We assume without loss of generality that $x_i = 0$ and drop the subscript i for simplicity. Below C denotes a generic constant that may change with each appearance. Let

$$q = \sum_{s=0}^{r} \binom{s+d-1}{d-1} = \binom{r+d}{d}, \tag{8.4}$$

which is the number of monomials in d variables of degree r or less. Let $\mathbf{X}$ denote the $|A| \times q$ matrix with coefficients $(x_j^s : j \in A, |s| \leq r)$. By definition of the local polynomial estimator (3.2) and the usual least squares formula, we have

$$\widehat{f} = \mathbf{e}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}, \tag{8.5}$$

where $\mathbf{e} = (1, 0, \dots, 0) \in \mathbb{R}^q$ and $\mathbf{y} := (y_j : j \in A)$ (assuming that $\mathbf{X}$ is full-rank, which we prove further down). In particular,

$$\operatorname{Var}(\widehat{f}) = \sigma^2\mathbf{e}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{e},$$

since $y_j = f(x_j) + \varepsilon_j$, with the noise (ε_j) being uncorrelated and having identical variance σ^2.

Let $z_j = x_j/h$ and $\mathbf{Z} = (z_j^s : j \in A, |s| \leq r)$, and also let $\mathbf{H} = \operatorname{diag}(h^{|s|}, |s| \leq r)$, so that $\mathbf{X} = \mathbf{Z}\mathbf{H}$, leading to

$$\operatorname{Var}(\widehat{f}) = \sigma^2\mathbf{e}^T\mathbf{H}^{-1}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{H}^{-1}\mathbf{e} = \sigma^2\mathbf{e}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{e}, \tag{8.6}$$

since $\mathbf{H}^{-1}\mathbf{e} = \mathbf{e}$. This is because $\mathbf{H}$ is an invertible diagonal matrix with first element equal to 1. The reason we work with $\mathbf{Z}$ instead of $\mathbf{X}$ is that, under the conditions assumed here, $z_j \in [-1, 1]^d$ (because $j \in A$) and $(nh)^{-d}\mathbf{Z}^T\mathbf{Z}$ is bounded from above and below in terms of its spectrum. Indeed, define matrices $\mathbf{Z}_1 = (z_j^s : j \in B^{\min}, |s| \leq r)$, $\mathbf{Z}_2 = (z_j^s : j \in A \setminus B^{\min}, |s| \leq r)$, $\mathbf{Z}_3 = (z_j^s : j \in B_h, |s| \leq r)$ and $\mathbf{Z}_4 = (z_j^s : j \in B_h \setminus A, |s| \leq r)$. Let $\prec$ denote the ordering for positive semi-definite matrices. Since

$$\mathbf{Z}_1^T\mathbf{Z}_1 \prec \mathbf{Z}_1^T\mathbf{Z}_1 + \mathbf{Z}_2^T\mathbf{Z}_2 = \mathbf{Z}^T\mathbf{Z} = \mathbf{Z}_3^T\mathbf{Z}_3 - \mathbf{Z}_4^T\mathbf{Z}_4 \prec \mathbf{Z}_3^T\mathbf{Z}_3,$$

it suffices that we focus on proving a lower bound on the spectrum of $\mathbf{Z}_1^T\mathbf{Z}_1$ and upper bound on the spectrum of $\mathbf{Z}_3^T\mathbf{Z}_3$. Consider therefore the case where A itself is a discrete ball, say

$A = \{j : x_j \in B(x, ah)\}$, where $a \in (C^{-1/d}, 1)$ by assumption. Let $z = x/h$. First, assume that a and h remain fixed. Then for $s, t \in \mathbb{N}^d$ such that $|s| \vee |t| \leq r$, we have

$$\frac{1}{(nh)^d}(\mathbf{Z}^T\mathbf{Z})_{st} = (nh)^{-d}\sum_{j\in A} z_j^{s+t} \to M_{st} := \int_{B(z,a)} u^{s+t}du, \quad \text{when } nh \to \infty, \tag{8.7}$$

recognizing a Riemann sum on the LHS. So, if $\mathbf{M} = (M_{st} : |s| \vee |t| \leq r)$, we have the convergence $(nh)^{-d}\ \mathbf{Z}^T\mathbf{Z} \to \mathbf{M}$, when $nh \to \infty$. $\mathbf{M}$ is a well-defined positive semi-definite matrix since its elements are bounded by 1 — because $B(z, a) \subset B(0, 1)$ — so we only need to show that it is positive uniformly over $a \in (C^{-1/d}, 1)$. Let $\lambda_{z,a}$ denote the smallest eigenvalue of $\mathbf{M}$ with integral over $B(z, a)$, with $z \in B(0, 1)$ and $a \in (C^{-1/d}, 1)$. We want to show that $\lambda_{z,a}$ is bounded away from 0. Suppose this is not the case, that there are sequences (z_m, a_m) such that $\lambda_{z_m,a_m} \to 0$ as $m \to \infty$. By compacity, we may assume that $(z_m, a_m) \to (z_\infty, a_\infty) \in \overline{B(0, 1)} \times [C^{-1/d}, 1]$. Then $\lambda_{z_\infty,a_\infty} = 0$, by continuity. Let $\mathbf{M}_\infty$ be the associated matrix. Then there is $b_\infty \in \mathbb{R}^q$ nonzero such that

$$0 = b_\infty^T\mathbf{M}_\infty b_\infty = \int_{B(z_\infty,a_\infty)} \sum_{s,t} b_{\infty s} b_{\infty,t} u^{s+t}du = \int_{B(z_\infty,a_\infty)} \Big(\sum_s b_{\infty,s}u^s\Big)^2 du,$$

where the sums are over $s \in \mathbb{N}^d$ such that $|s| \leq r$. This leads to a contradiction since the polynomial in the second integral cannot be zero on a nonempty ball.

So far, we assumed that a and z were fixed. Assume this is not the case. The upper bound on the largest eigenvalue of $\mathbf{Z}^T\mathbf{Z}$ is bounded in the exact same way, using the fact that $\|z_j\| \leq 1$. For the lower bound we still have that

$$\liminf\ (nh)^{-d}\sum_{j\in A} z_j^{s+t} \geq \inf_{z',a'} \int_{B(z',a')} u^{s+t}, du$$

where the infimum is over z' and a' such that $a' \in (C^{-1/d}, 1)$ and $B(z', a') \subset B(0, 1)$. And our arguments apply to the RHS. We conclude that there is $C_1 \in (0, \infty)$ such that, for nh large enough,

$$\frac{1}{C_1}(nh)^d \leq \lambda_{\min}(\mathbf{Z}^T\mathbf{Z}) \leq \lambda_{\max}(\mathbf{Z}^T\mathbf{Z}) \leq C_1(nh)^d. \tag{8.8}$$

We then redefine C as $\max(C, C_1)$ and conclude with (8.6). ■

Lemma 8.6 (Bias: Upper Bound). *Assume that $f \in \mathcal{F}^{\text{cartoon}}(\alpha, C_0)$, with foreground Ω, and that the conditions of Lemma 8.5 also hold. If moreover either $A_i \subset \Omega$ or $A_i \subset \Omega^c$, then, for some constant $C > 0$, the following inequality holds*

$$(\mathbb{E}\widehat{f}_i - f_i)^2 \leq \min(1, Ch^{2\alpha}). \tag{8.9}$$

Proof. We continue with the notation introduced in the proof of Lemma 8.6. WLOG, assume $A \subset \Omega$. In that case, f is smooth in the window, since $f = f_\Omega$. By Lemma 8.1, $f(x_j) = T_0^{\lfloor\alpha\rfloor} f(x_j) + g(x_j)$, where $T_0^{\lfloor\alpha\rfloor} f$ is a polynomial of degree at most $\lfloor\alpha\rfloor \leq r$, and

$|g(x_j)| \le c_\alpha C_0 h^\alpha$ for all $j \in A$. Now, for a polynomial p of degree at most r, let $\mathbf{p} = (p(x_j) : j \in A)$ so that $\mathbf{p} = \mathbf{X}\mathbf{a}$ for some $\mathbf{a} \in \mathbb{R}^q$, and we have

$$\mathbf{e}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{p} = \mathbf{e}^T\mathbf{a} = a_0 = p(0).$$

With this reproducing formula and the fact that $T_0^{\lfloor\alpha\rfloor} f(0) = f(0)$,

$$\mathbb{E}\widehat{f} - f(0) = \mathbf{e}^T(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{g} = \mathbf{e}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{g}.$$

Because of (8.8), we have

$$|\mathbf{e}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{g}| \le \frac{\|\mathbf{e}\|_2 \cdot \|\mathbf{Z}^T\mathbf{g}\|_2}{\lambda_{\min}(\mathbf{Z}^T\mathbf{Z})} \le C_1(nh)^{-d}\|\mathbf{Z}^T\mathbf{g}\|_2, \tag{8.10}$$

where C_1 is the constant of Lemma 8.5. But the entries $(\mathbf{Z}^T\mathbf{g})_s = \sum_{j\in A} g(x_j)z_j^s$ are uniformly bounded by $|A| \cdot c_\alpha C_0 h^\alpha = O((nh)^d h^\alpha)$, so that the RHS in (8.10) is of order $O(h^\alpha)$. This imply that $(\mathbb{E}\widehat{f} - f(0))^2 \le C_2 h^{2\alpha}$ for some constant C_2, and we conclude by redefining C as $\max(C, C_1, C_2)$. ■

Lemma 8.7 (Bias: Lower Bound). *Let $f = \mathbf{1}_\Omega$, where $\Omega = (0, 1/2) \times (0, 1)^{d-1}$ and linear filtering (meaning $\omega_{i,j} = 1$ if, and only if, $\|x_i - x_j\| \le h/2$). Then there is a constant $C > 0$ such that, when $\operatorname{dist}(x_i, \partial\Omega) \le h/C$, we have*

$$(\mathbb{E}\widehat{f}_i - f_i)^2 \ge 1/C. \tag{8.11}$$

Proof. We continue with the notation introduced in the proof of Lemma 8.6. In particular, we translate everything so that $x_i = 0$. WLOG, assume that $x_i \in \Omega$ and let $\delta = \operatorname{dist}(x_i, \partial\Omega)$. Let $A_\Omega = \{j : x_j \in B_h \cap \Omega\}$ and define A_{Ω^c} similarly. Using the reproducing formula, we get

$$\mathbb{E}\widehat{f} - f(x_i) = \mathbf{e}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{1}_{A_\Omega} - \mathbf{e}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{1} = -\mathbf{e}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{1}_{A_{\Omega^c}}. \tag{8.12}$$

Assume that $\delta < h$, in which case $A_{\Omega^c} = \{j : x_j \in (\delta, h) \times (-h, h)^{d-1}\}$, in which case the RHS in (8.12) is equal to $-G(\delta/h)$, where

$$G(a) := \mathbf{e}^T(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T\mathbf{1}_{\{j:\, z_j \in (a,1)\times(-1,1)^{d-1}\}}.$$

It suffices to show that there is $C > 0$ such that $G(a) \ge 1/C$ when $a \le 1/C$. Assume this is not the case, in which case there is $a_m \to 0$ such that $G(a_m) \to 0$. As in the proof of Lemma 8.6, recognizing Riemann sums we see that, as $m \to \infty$,

$$\frac{1}{(nh)^d}\mathbf{Z}^T\mathbf{Z} \to \mathbf{M} := \int_{(-1,1)^d} \mathbf{p}(z)\mathbf{p}(z)^T dz, \quad \mathbf{p}(z) := (z^s : |s| \le r),$$

and

$$\frac{1}{(nh)^d}\mathbf{Z}^T\mathbf{1}_{\{j:\, z_j \in (a,1)\times(-1,1)^{d-1}\}} \to \mathbf{v} := \int_{(0,1)\times(-1,1)^{d-1}} \mathbf{p}(z)dz.$$

Let

$$\mathbf{u} = \int_{(-1,0)\times(-1,1)^{d-1}} \mathbf{p}(z)dz.$$

We have $\mathbf{u} + \mathbf{v} = \mathbf{M} \cdot \mathbf{1}$ (where $\mathbf{1} = (1, \dots, 1)$), so that $\mathbf{e}^T\mathbf{M}^{-1}(\mathbf{u} + \mathbf{v}) = 1$, and therefore $\mathbf{e}^T\mathbf{M}^{-1}\mathbf{v} = 1/2$ by symmetry. Hence, $G(a_m) \to 1/2$, which is a contradiction. ■

This lemma states a lower bound on the squared bias of linear filtering when f is an indicator function of a half hypercube. The result is actually much more general. Indeed, any function f in the cartoon class has a foreground Ω whose boundary is well-approximated by a hyperplane — since $\partial\Omega$ is Lipschitz — and because f is smooth on Ω and Ω^c, it is approximately piecewise constant locally. Hence, near the discontinuity, f resembles the function in Lemma 8.7.

8.1.4. Some probability. The following result asserts that the maximum of m identically distributed random variables with exponentially decaying tails is at most a power of $\log m$.

Lemma 8.8. *Suppose $X_1, \dots, X_m$ are such that for some $a, b, c > 0$,*

$$\mathbb{P}(|X_r| > t) \le c \exp(-(t/a)^b), \quad \forall t > c, \quad \forall r = 1, \dots, m.$$

Then for m sufficiently large,

$$\mathbb{P}\left(\max(|X_1|, \dots, |X_m|) > a(2\log m)^{1/b}\right) \le c/m.$$

Proof. Define $x_m = a(2\log m)^{1/b}$. By the union bound,

$$\begin{aligned}\mathbb{P}(\max(|X_1|, \dots, |X_m|) > x_m) &\le \mathbb{P}(|X_1| > x_m) + \cdots \mathbb{P}(|X_m| > x_m) \\ &\le mc \exp(-(x_m/a)^b) \\ &= cm^{-1} \to 0.\end{aligned}$$

■

Lemma 8.9. *For $X_i \sim \mathcal{N}(0, \sigma_i^2)$ for $i = 1, \dots, m$, and any $C > 0$, we have*

$$\mathbb{P}\left(\max_{1\le i\le m} |X_i| > \max_{1\le i\le m} \sigma_i \sqrt{2C\log m}\right) \le m^{1-C}.$$

Proof. Fix $t \ge 1$ and let $\sigma = \max_{1\le i\le m} \sigma_i$. By the union bound and the fact that $\mathbb{P}(\mathcal{N}(0,1) > t) \le \exp(-t^2/2)$, we have

$$\mathbb{P}\left(\max_{1\le i\le m} |X_i| > t\right) \le \sum_{i=1}^m \mathbb{P}(|X_i| > t) \le \sum_{i=1}^m \exp(-t^2/(2\sigma_i^2)) \le m\exp(-t^2/(2\sigma^2)).$$

We then plug in $t = \sigma\sqrt{2C\log m}$. ■

Lemma 8.10. *Suppose $X_i \sim \chi^2_k$ for $i = 1, \dots, m$. There is a constant C_1 such that, for any $C > C_1$, if $k \geq (64/9)C \log(m)$, we have*

$$\mathbb{P}\left(\max_{1\leq i\leq m} X_i > k + 2\sqrt{Ck\log m}\right) \leq m^{1-C/2} \tag{8.13}$$

$$\mathbb{P}\left(\min_{1\leq i\leq m} X_i < k - 2\sqrt{Ck\log m}\right) \leq m^{1-C/2} \tag{8.14}$$

Proof. Let us prove the first inequality. Since the moment generating function of a χ^2_k is $t \to (1-2t)^{-k/2}\mathbb{1}(t < 1/2)$, Chernoff's bound gives

$$\mathbb{P}(X_i > t) \leq \exp\left(-(t-k)/2 + (k/2)\log(t/k)\right), \quad \forall t > k.$$

We then use the inequality $\log(1+x) \leq x - x^2/2 + x^3/3$, valid for $x \in (0,1)$ and input $t = k + 2\sqrt{Ck\log m}$, to get

$$\begin{aligned} -(t-k)/2 + (k/2)\log(t/k) &\leq -(t-k)^2/(4k) + (t-k)^3/(6k^2) \\ &= -C\log m + (4/3)C^{3/2}\sqrt{\log(m)/k}\log m, \end{aligned}$$

and bound the second term by $(C/2)\log m$. We then obtain

$$\mathbb{P}(X_i > t) \leq m^{-C/2},$$

and apply the union bound as before. The second inequality is proved in the same way considering that $\log(1-x) \geq -x - x^2/2 - x^3/3$ holds for $x \in (0,1)$. ■

Lemma 8.11. *Suppose $X_i \sim \chi^2_k(\delta_i^2)$ (non-central chi-square) for $i = 1, \dots, m$. There is a constant C_1 such that, for any $C > C_1$, if $k \geq 16C\log(m)$ and $\delta_{\min} := \min_i \delta_i \geq 2\sqrt{C\log m}$, we have*

$$\mathbb{P}\left(\min_{1\leq i\leq m} X_i < \delta^2_{\min}/4 + k - 3\sqrt{Ck\log m}\right) \leq 2m^{1-C/2}.$$

Similarly, if $\delta_{\max} = \max_i \delta_i \leq \sqrt{C\log m}$, we have

$$\mathbb{P}\left(\max_{1\leq i\leq m} X_i > k + 3\sqrt{Ck\log m}\right) \leq m^{1-C/2}.$$

Proof. We first notice that $X_i \equiv (Z_i + \delta_i)^2 + Y_i$, where $Z_i \sim \mathcal{N}(0,1)$ and $Y_i \sim \chi^2_{k-1}$ are independent. Hence,

$$\min_{1\leq i\leq m} X_i \geq \min_{1\leq i\leq m} (Z_i + \delta_i)^2 + \min_{1\leq i\leq m} Y_i.$$

Let $E_i = \{\max_{1\leq i\leq m} |Z_i| \geq \sqrt{C\log m}\}$. By Lemma 8.9, we have

$$\mathbb{P}(E_i) \leq m^{1-C/2}.$$

Let

$$F_i = \{\min_{1\leq i\leq m} Y_i \leq k - 1 - 2\sqrt{C(k-1)\log m}\}.$$

To control the Y_i's, we apply inequality (8.14) to get

$$\mathbb{P}(F_i) \le m^{1-C/2}.$$

Under $E_i^c \cap F_i^c$, we have $\min_i X_i \ge \delta^2/4 + k - 3\sqrt{Ck \log m}$ and

$$\mathbb{P}(E_i^c \cap F_i^c) = 1 - \mathbb{P}(E_i \cup F_i) \ge 1 - \mathbb{P}(E_i) - \mathbb{P}(F_i) \ge 1 - 2m^{1-C/2}.$$

This proves the bound on $\min_i X_i$; arguments for $\max_i X_i$ are similar and simpler. ■

8.2. Proofs of the main results.

8.2.1. Proof of Theorem 4.1. We start with the upper bound. Fix $f \in \mathcal{F}^{\text{cartoon}}(\alpha, C_0)$ with foreground Ω. Let $Q = \{i : \text{dist}(x_i, \partial\Omega) \le h\}$. For $i \in Q$, we use the fact that $|\widehat{f}_i - f_i| \le 1$, which implies $\mathbb{E}(\widehat{f}_i - f_i)^2 \le 1$. For $i \notin Q$, from Lemma 8.5 and Lemma 8.6, coupled with the bias-variance decomposition (2.3), we get

$$\mathbb{E}(\widehat{f}_i - f_i)^2 \le C(h^{2\alpha} + \sigma^2(nh)^{-d}).$$

Using Lemma 8.3, we have $|Q| \le 4^d n^d |B(\partial\Omega, h)|$, while and $|B(\partial\Omega, h)| = O(h)$ by Lemma 8.4 and the fact that $|\partial\Omega|$ is of order 1. Summing over all $i \in I_n^d$, we get

$$\text{MSE}_f(\widehat{f}) \le \frac{n^d - |Q|}{n^d} \cdot C(h^{2\alpha} + \sigma^2(nh)^{-d}) + \frac{|Q|}{n^d} \cdot C(1 + \sigma^2(nh)^{-d}) \le C_1(h + \sigma^2(nh)^{-d}).$$

Minimizing the RHS with respect to h yields the upper bound in Theorem 4.1.

For the lower bound, redefine $Q = \{i : \text{dist}(x_i, \partial\Omega) \le h/C_1\}$, where C_1 is the constant of Lemma 8.7. For $i \notin Q$, we use Lemma 8.5 and the bias-variance decomposition (2.3), to get

$$\mathbb{E}(\widehat{f}_i - f_i)^2 \ge \frac{1}{C_2}\sigma^2(nh)^{-d}.$$

For $i \in Q$ we use Lemma 8.7 and the bias-variance decomposition (2.3), to get

$$\mathbb{E}(\widehat{f}_i - f_i)^2 \ge \frac{1}{C_1},$$

Using Lemma 8.3 again, we have the following lower bound on the MSE (for n large enough),

$$\text{MSE}_f(\widehat{f}) \ge \frac{n^d - |Q|}{n^d} \cdot \frac{1}{C_2}\sigma^2(nh)^{-d} + \frac{|Q|}{n^d} \cdot \frac{1}{C_1} \ge C_3(h + \sigma^2(nh)^{-d}).$$

Minimizing the RHS with respect to h leads to the lower bound in Theorem 4.1.

8.2.2. Proof of Theorem 4.2. The proof for the upper bound is the same as that of Theorem 4.1 in smooth regions, leading to an upper bound on the MSE of the form

$$\text{MSE}_f(\widehat{f}) \le C(h^{2\alpha} + \sigma^2(nh)^{-d}).$$

Then minimizing this quantity over h gives the stated result. The lower bound is a well-known minimax bound [26, Theorem 5.1.2, p. 133].

8.2.3. Proof of Theorem 4.3. Let $\delta(x) = \operatorname{dist}(x, \Omega)$. The proof is similar to that of Theorem 4.2, except that the variance varies by location. The point bias is of order $O(h^\alpha)$ everywhere, because the smoothing window is of radius at most h, with all points in the window being on the same side of the discontinuity. However, the point variance is of order $O(\sigma^2 \lceil n\delta(x_i)\rceil^{-d})$, since the window is of radius $\delta(x_i)$ (immediate consequences of Lemma 8.5).

Let us sum the point variances over all the pixels in the image. The situation is different according to the dimension. We start with $d = 1$, so that $\Omega = (a, b) \subset (0, 1)$. For δ small enough, there are exactly four points at distance less than δ from $\partial\Omega$ (two on each side of the two jump locations). Let's consider the sample points $x_i \in [b, 1)$, and let j be such that $x_{j-1} < b \le x_j$. Note that $j = bn(1 + o(1))$, and we assume that b is fixed. For $i \in [j, j + nh]$, the variance is bounded by $C\sigma^2/(i - j + 1)$, while for $i \ge j + nh$ (in the smooth region), the variance is of order $O(\sigma^2/(nh))$ as before. Hence, summing over $i \ge j$, the averaged variance in that region is bounded by

$$\frac{C\sigma^2}{n - nh - j}\left(\sum_{i=j}^{j+[nh]} \frac{1}{i - j + 1} + \frac{n - nh - j}{nh}\right) = O(\frac{\sigma^2}{n})\left(\sum_{k=1}^{nh} \frac{1}{k} + 1\right) = O(\frac{\sigma^2 \log(n)}{n}).$$

The same is true for all the other three regions.

When $d \ge 2$, the story is just slightly different. Define $Q_\ell = \{i : \delta(x_i) \le h2^{-\ell}\}$ and let ℓ_0 be such that $h2^{-\ell_0} < 2/n \le h2^{-\ell_0+1}$. Stratifying, we have the following bound on the averaged variance

$$\frac{C\sigma^2}{n^d}\left(\sum_{\ell=0}^{\ell_0} \sum_{i \in Q_\ell \setminus Q_{\ell+1}} (nh2^{-\ell-1})^{-d} + \sum_{i \notin Q_0} (nh)^{-d}\right) = \frac{C\sigma^2}{n^d} \sum_{\ell=0}^{\ell_0} |Q_\ell \setminus Q_{\ell+1}|(nh)^{-d} 2^{d(\ell+1)} + \frac{C\sigma^2}{(nh)^d}.$$

By Lemma 8.3 and Lemma 8.4, we have $|Q_\ell \setminus Q_{\ell+1}| \le |Q_\ell| \le C_1 n^d \cdot h2^{-\ell}$, for some constant C_1. Hence, the first sum on the RHS of the last equation is bounded by

$$C_2\sigma^2 h(nh)^{-d} \sum_{\ell=0}^{\ell_0} 2^{(d-1)\ell} \le C_3\sigma^2 h(nh)^{-d} \cdot 2^{(d-1)\ell_0} = O(\sigma^2/n).$$

This leads to an upper bound on the MSE of the form

$$\mathrm{MSE}_f(\widehat{f}) \le C(h^{2\alpha} + \sigma^2 A_n/n + \sigma^2(nh)^{-d}). \tag{8.15}$$

Minimizing this quantity over h gives the upper bound stated in Theorem 4.3.

For the lower bound, we know from minimax results underlying the lower bound in Theorem 4.2 that there are functions f in the cartoon class where the bias in the smooth regions is of order at least h^α. As for the variance, our upper bound for the averaged variance is easily seen to be lower bounded (up to a multiplicative constant). This leads to a lower bound identical to (8.15) modulo a multiplicative constant, and optimizing it leads to the lower bound in Theorem 4.3. We omit details.

8.2.4. Proof Theorem 4.4. Fix $i \in I_n^d$ and let $\eta_i = \max_{j \in B(i,nh)} |\varepsilon_j|$. Then by the union bound and (4.1),

$$\mathbb{P}(\eta_i \geq t) \leq |B(i, nh)| \max_{j \in B(i,nh)} \mathbb{P}(|\varepsilon_j| \geq t) \leq (2nh+1)^d \, (1 - F(t/\sigma)) =: p. \tag{8.16}$$

Hence, with probability at least $1 - p$, the event $E_i := \{\eta_i \leq t\}$ holds true. WLOG, assume that $x_i \in \Omega$. Since f_Ω is C_0-Lipschitz, we have $|f(x_i) - f(x_j)| = |f_\Omega(x_i) - f_\Omega(x_j)| \leq C_0 h$ when $x_j \in \Omega \cap B(x_i, h)$, and by the triangle inequality, $|y_i - y_j| \leq C_0 h + |\varepsilon_i - \varepsilon_j| \leq C_0 h + 2t$ under E_i. Suppose there is $x_j \in \Omega^c \cap B(x_i, h)$. In that case, there is $x \in \partial\Omega \cap B(x_i, h)$ and we have

$$\begin{aligned} |f(x_i) - f(x_j)| &\geq |f_\Omega(x) - f_{\Omega^c}(x)| - |f_\Omega(x_i) - f_\Omega(x)| - |f_{\Omega^c}(x_j) - f_{\Omega^c}(x)| \\ &\geq \mu(f) - 2C_0 h \geq 1/C_0 - 2C_0 h, \end{aligned}$$

again by the triangle inequality and the fact that f_{Ω^c} is also C_0-Lipschitz, this implies that

$$|y_i - y_j| \geq 1/C_0 - 2C_0 h - |\varepsilon_i - \varepsilon_j| \geq 1/C_0 - 2C_0 h - 2t,$$

under E_i. We see that we need to take $h_y \geq C_0 h + 2t$ to ensure that sample points $x_j \in \Omega \cap B(x_i, h)$ are selected, while we require that $h_y < 1/C_0 - 2C_0 h - 2t$ so that points $x_j \in \Omega^c \cap B(x_i, h)$ are disregarded. These two inequalities are, for example, satisfied when $h_y = 1/(3C_0)$ and $t = 1/(6C_0)$, and h sufficiently small — by our assumptions, $h = o(1)$. Assume h_y and t are chosen that way. Then, when E_i holds, the photometric kernel in Yaroslavsky's filter is able to exactly mimic the *membership oracle*.

We now turn to bounding the MSE. First, we have

$$\mathbb{E}(\widehat{f}_i^{\mathrm{YF}} - f_i)^2 = \mathbb{E}[(\widehat{f}_i^{\mathrm{YF}} - f_i)^2 \mathbb{1}_{\{E_i\}}] + \mathbb{E}[(\widehat{f}_i^{\mathrm{YF}} - f_i)^2 \mathbb{1}_{\{E_i^c\}}].$$

Since $\widehat{f}_i^{\mathrm{YF}} = \widehat{f}_i^{\mathrm{MO}}$ on E_i,

$$\mathbb{E}[(\widehat{f}_i^{\mathrm{YF}} - f_i)^2 \mathbb{1}_{\{E_i\}}] = \mathbb{E}[(\widehat{f}_i^{\mathrm{MO}} - f_i)^2 \mathbb{1}_{\{E_i\}}] \leq \mathbb{E}(\widehat{f}_i^{\mathrm{MO}} - f_i)^2.$$

And since $|\widehat{f}_i^{\mathrm{YF}} - f_i| \leq 1$ because of our clipping, we have

$$\mathbb{E}[(\widehat{f}_i^{\mathrm{YF}} - f_i)^2 \mathbb{1}_{\{E_i^c\}}] \leq \mathbb{P}(E_i^c) \leq p.$$

It remains to check that p is negligible compared to MO risk given in Theorem 4.2. Indeed, using the fact that $t \asymp 1$, that $h \leq 1$ and that $\sigma \leq (C' \log n)^{-1/b}$, we have

$$p = O(nh)^d \exp[-(t/(C\sigma))^b] = \exp[(d - t^b(C'/C^b)) \log n] = o(\sigma^2/n^d)^{2\alpha/(d+2\alpha)},$$

when C' is sufficiently large, implicitly assuming that σ is at least a polynomial in n, for otherwise the trivial estimator $\widehat{y} = y$ is optimal.

This concludes the proof.

When the noise level is not small. Assume σ is fixed, for simplicity. Note YF is identical to LF when $h_y \to \infty$ sufficiently fast. Assume therefore that $h_y \leq h_0$ for some fixed $h_0 < \infty$.

We now argue that YF is essentially useless when this is the case. Concretely, assume the reverse of (4.1), meaning

$$\mathbb{P}(|\varepsilon_i| \le t) \le F(t/\sigma),\ \forall t,\ \forall i \in I_n^d. \tag{8.17}$$

We show that, when $F(2h_0/\sigma) < 1$, YF has an overall squared bias (and therefore MSE) of order 1, which is comparable to the trivial estimator $\widehat{f} = y$. In other words, for large noise and relatively small h_y, YF can perform *worse* than LF. For example, the bias is at least h_0 at locations i satisfying $|\varepsilon_i| \ge h_0 + h_y$. Indeed, we are averaging over values y_j such that $|y_j - y_i| \le h_y$, so that $|\widehat{f}_i - y_i| \le h_y$ and therefore $|\widehat{f}_i - f_i| \ge |\varepsilon_i| - |\widehat{f}_i - y_i| \ge (h_0 + h_y) - h_y = h_0$. Moreover, by (8.17)

$$\mathbb{P}\left(|\varepsilon_i| \ge h_0 + h_y\right) \ge 1 - F(2h_0/\sigma) > 0.$$

Hence, integrating the squared bias over these sample points alone leads to a lower bound of order 1.

8.2.5. Proof of Theorem 4.5. For simplicity, we ignore boundary issues and in particular assume that all patches are of same size, with $m_{\mathsf{P}} \asymp (nh_{\mathsf{P}})^d$ sample points each, and similarly for spatial windows, with $m_h \asymp (nh)^d$ sample points each.

Upper bound for NLM-average. For $i \in I_n^d$ such that $\mathsf{P}_i \cap \partial\Omega \ne \emptyset$, we use the fact that $|\widehat{f}_i - f_i| \le 1$ to get

$$\mathbb{E}(\widehat{f}_i - f_i)^2 \le 1.$$

Consider i with $\mathsf{P}_i \cap \partial\Omega = \emptyset$. WLOG, assume $\mathsf{P}_i \subset \Omega$. Take any $j \in B(i, nh)$. By definition,

$$\overline{y}_{\mathsf{P}_j} - \overline{y}_{\mathsf{P}_i} = \overline{f}_{\mathsf{P}_j} - \overline{f}_{\mathsf{P}_i} + \overline{\varepsilon}_{\mathsf{P}_j} - \overline{\varepsilon}_{\mathsf{P}_i}. \tag{8.18}$$

For the noise part we have

$$\overline{\varepsilon}_{\mathsf{P}_j} - \overline{\varepsilon}_{\mathsf{P}_i} \sim \mathcal{N}(0, \sigma^2 |\mathsf{P}_i \Delta \mathsf{P}_j| / m_{\mathsf{P}}^2), \tag{8.19}$$

where Δ denotes the symmetric difference. By Lemma 8.9, we have that

$$\max_{j \in B(x_i, h)} |\overline{\varepsilon}_{\mathsf{P}_j} - \overline{\varepsilon}_{\mathsf{P}_i}| \le \zeta := 2\sigma\sqrt{C \log(m_h)/m_{\mathsf{P}}}, \tag{8.20}$$

with probability at least $1 - m_h^{1-C}$. In the sequel, we fix C large and denote by E_i the event (8.20). For the signal part, we have the following

$$\begin{aligned}
\overline{f}_{\mathsf{P}_i} - \overline{f}_{\mathsf{P}_j} &= \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_i} f(x_k) - \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_j} f(x_k) \\
&= \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_0} (f(x_k + x_i) - f(x_k + x_j)),
\end{aligned}$$

where P_0 is a generic patch centered at 0. If $x_j \in \Omega$ with $\mathsf{P}_j \subset \Omega$, then, since f_Ω is C_0-Lipschitz,

$$\begin{aligned}
|\overline{f}_{\mathsf{P}_i} - \overline{f}_{\mathsf{P}_j}| &\le \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_0} |f_\Omega(x_k + x_i) - f_\Omega(x_k + x_j)| \\
&\le C_0 |x_i - x_j| \le C_0 h.
\end{aligned} \tag{8.21}$$

If $x_j \in \Omega^c$, then there is a point $x \in B(x_i, h) \cap \partial\Omega$, and we have $f(x_k) = f_\Omega(x) + [f_\Omega(x_k) - f_\Omega(x)]$ for $x_k \in \Omega$ and $f(x_k) = f_{\Omega^c}(x) + [f_{\Omega^c}(x_k) - f_{\Omega^c}(x)]$ for $x_k \in \Omega^c$, with $|f_\Omega(x_k) - f_\Omega(x)| \le C_0 h$, $|f_{\Omega^c}(x_k) - f_{\Omega^c}(x)| \le C_0 h$ and $|f_\Omega(x) - f_{\Omega^c}(x)| \ge 1/C_0$. Hence,

$$\begin{aligned}
\overline{f}_{\mathsf{P}_i} - \overline{f}_{\mathsf{P}_j} &= \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_i} f(x_k) - \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_j} f(x_k) \\
&= f_\Omega(x) + \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_i} [f_\Omega(x_k) - f_\Omega(x)] \\
&\quad - f_\Omega(x) \frac{|\mathsf{P}_j \cap \Omega|}{|\mathsf{P}_j|} - \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_j \cap \Omega} [f_\Omega(x_k) - f_\Omega(x)] \\
&\quad - f_{\Omega^c}(x) \frac{|\mathsf{P}_j \cap \Omega^c|}{|\mathsf{P}_j|} - \frac{1}{m_{\mathsf{P}}} \sum_{x_k \in \mathsf{P}_j \cap \Omega^c} [f_{\Omega^c}(x_k) - f_{\Omega^c}(x)] \\
&= (f_\Omega(x) - f_{\Omega^c}(x)) \frac{|\mathsf{P}_j \cap \Omega^c|}{|\mathsf{P}_j|} + R,
\end{aligned}$$

where $|R| \le 2C_0 h$. We now use Lemma 8.2 to bound the fraction above from below by $(2C_0)^{-d}$, to get

$$|\overline{f}_{\mathsf{P}_i} - \overline{f}_{\mathsf{P}_j}| \ge (2C_0)^{-d} \mu - 2C_0 h. \tag{8.22}$$

Using the decomposition (8.18), coupled with the triangle inequality and (8.20), (8.21) and (8.22), we see that we need to choose h_y such that

$$C_0 h + \zeta \le h_y < (2C_0)^{-d} \mu - 2C_0 h - \zeta. \tag{8.23}$$

The lower bound is to ensure that all the points $x_j \in B(x_i, h)$ such that $\mathsf{P}_j \subset \Omega$ are included in the neighborhood of x_i (*i.e.*, $\omega_{ij} = 1$), while the upper bound is to ensure that no points in Ω^c are included (under E_i). For points $x_j \in B(x_i, h)$ such that $\mathsf{P}_j \cap \Omega^c \ne \emptyset$, they may or may not be included, depending on how large that intersection is. Note that (8.23) is satisfied when h_y is a sufficiently small constant since we have $h \to 0$, $\zeta \to 0$ and $\mu \asymp 1$ under our assumptions. In any case, we assume that (8.23) holds.

In terms of MSE, we proceed as follows. Let $B_i = \{j : x_j \in B(x_i, h)\}$, $B_i^0 = \{j : x_j \in B(x_i, h), \mathsf{P}_j \subset \Omega\}$ and $A_i = \{j : \omega_{i,j} = 1\}$ — the latter is a random subset of B_i. We saw that $A_i \supset B_i^0$ under E_i, which implies

$$E_i \subset \{A_i \supset B_i^0\} \subset \bigcup_{B_i^0 \subset A \subset B_i} \{A_i = A\},$$

leading to

$$\mathbb{1}_{\{E_i\}} \le \sum_{B_i^0 \subset A \subset B_i} \mathbb{1}_{\{\{A_i = A\}\}}. \tag{8.24}$$

Using (8.24) and the fact that $|\widehat{f}_i - f_i| \le 1$, we have

$$\begin{aligned}
\mathbb{E}(\widehat{f}_i - f_i)^2 &= \mathbb{E}[(\widehat{f}_i - f_i)^2 \mathbb{1}_{\{E_i\}}] + \mathbb{E}[(\widehat{f}_i - f_i)^2 \mathbb{1}_{\{E_i^c\}}] \\
&\le \sum_{B_i^0 \subset A \subset B_i} \mathbb{P}(A_i = A)\, \mathbb{E}[(\widehat{f}_A - f_i)^2] + \mathbb{P}(E_i^c),
\end{aligned}$$

where $\widehat{f}_A$ is the local polynomial estimator based on $A \subset I_n^d$. For the second term, $\mathbb{P}(E_i^c) \leq m_h^{1-C}$ by (8.20). For the first term, by Lemma 8.2, we know that $B(x_i, h) \cap \Omega$ contains a ball of radius $C_1 h$ with $C_1 > 0$ depending only on C_0 and d. Hence, by the triangle inequality, $B(x_i, h) \setminus B(\Omega^c, h_{\mathsf{P}})$ contains a ball of radius $C_1 h - h_{\mathsf{P}} \geq C_1 h/2$ (eventually), implying that B_i^0 contains a discrete ball of radius at least $(C_1 h/3)n \asymp nh$. Therefore $|B_i^0|/|B_i| \asymp 1$ and we may apply Lemma 8.5 and Lemma 8.6 to each A in the sum above, to get

$$\mathbb{E}(\widehat{f}_A - f_i)^2 \leq C_2(h^{2\alpha} + \sigma^2(nh)^{-d}),$$

for a constant C_2. Hence, using the fact that $\sum_{B_i^0 \subset A \subset B_i} \mathbb{P}(A_i = A) \leq 1$, we have

$$\mathbb{E}(\widehat{f}_i - f_i)^2 \leq C_2(h^{2\alpha} + \sigma^2(nh)^{-d}) + m_h^{1-C}.$$

By our choice for h, $h^{2\alpha} + \sigma^2(nh)^{-d} \asymp (\sigma^2/n^d)^{2\alpha/(d+2\alpha)}$, and we may choose C large enough in the exponent that the last term on the RHS is negligible, leading to an MSE at i of order $O(\sigma^2/n^d)^{2\alpha/(d+2\alpha)}$.

The MSE is of the same order of magnitude when $x_i \in \Omega^c$, and summing over all $i \in I_n^d$, we get

$$\mathrm{MSE}_f(\widehat{f}) \leq \frac{|Q|}{n^d} + O(\sigma^2/n^d)^{2\alpha/(d+2\alpha)},$$

where $Q := \{i : \mathsf{P}_i \cap \partial\Omega \neq \emptyset\}$. Since $Q \subset \{i : \mathrm{dist}(x_i, \partial\Omega) < h_{\mathsf{P}}\}$, by Lemma 8.3 and Lemma 8.4, we have $|Q| \leq C_2 n^d \cdot h_{\mathsf{P}}$, so that

$$\mathrm{MSE}_f(\widehat{f}) \leq O(h_{\mathsf{P}} + (\sigma^2/n^d)^{2\alpha/(d+2\alpha)}),$$

which is what we needed to prove. Optimizing over h_{P} subject to (8.23) being satisfied, we achieve the desired result.

Upper bound for NLM. We follow the same arguments. Here we focus on $i \in I_n^d$ such that $\mathrm{dist}(x_i, \partial\Omega) > 2h_{\mathsf{P}}$ (instead of h_{P}), and assume WLOG that $x_i \in \Omega$. Take $j \in B(i, nh)$ such that $\mathsf{P}_j \cap \mathsf{P}_i = \emptyset$. Note that this is true when $x_j \in \Omega^c$. By definition,

$$\mathbf{y}_{\mathsf{P}_j} - \mathbf{y}_{\mathsf{P}_i} = \mathbf{f}_{\mathsf{P}_j} - \mathbf{f}_{\mathsf{P}_i} + \boldsymbol{\varepsilon}_{\mathsf{P}_j} - \boldsymbol{\varepsilon}_{\mathsf{P}_i}.$$

Since $\boldsymbol{\varepsilon}_{\mathsf{P}_j} - \boldsymbol{\varepsilon}_{\mathsf{P}_i} \sim \mathcal{N}(0, 2\sigma^2 \mathbf{I}_{m_{\mathsf{P}}})$, we have $\|\mathbf{y}_{\mathsf{P}_j} - \mathbf{y}_{\mathsf{P}_i}\|_2^2 \sim 2\sigma^2 \chi^2_{m_{\mathsf{P}}}(\|\mathbf{f}_{\mathsf{P}_j} - \mathbf{f}_{\mathsf{P}_i}\|_2^2/(2\sigma^2))$, with

$$\|\mathbf{f}_{\mathsf{P}_j} - \mathbf{f}_{\mathsf{P}_i}\|_2^2 = \sum_{x_k \in \mathsf{P}_0} (f(x_k + x_j) - f(x_k + x_i))^2.$$

If $x_j \in \Omega$ with $\mathsf{P}_j \subset \Omega$, then

$$\|\mathbf{f}_{\mathsf{P}_j} - \mathbf{f}_{\mathsf{P}_i}\|_2^2 = \sum_{x_k \in \mathsf{P}_0} (f_\Omega(x_k + x_i) - f_\Omega(x_k + x_j))^2 \tag{8.25}$$

$$\leq m_{\mathsf{P}} C_0^2 \|x_i - x_j\|_2^2 \leq m_{\mathsf{P}} C_0^2 h^2, \tag{8.26}$$

since f_Ω is C_0-Lipschitz. By Lemma 8.11 and the fact that $m_{\mathsf{P}} C_0^2 h^2/\sigma^2 = o(1)$, we conclude that

$$\max_j \|\mathbf{y}_{\mathsf{P}_j} - \mathbf{y}_{\mathsf{P}_i}\|_2^2 \leq 2\sigma^2 m_{\mathsf{P}} + \zeta_\chi, \quad \zeta_\chi := 6\sigma^2 \sqrt{C m_{\mathsf{P}} \log m_h}, \tag{8.27}$$

with probability at least $1-m^{1-C/2}$, where the maximum is over j such that $x_j \in B(x_i,h)$ and $\mathsf{P}_j \subset \Omega \setminus \mathsf{P}_i$. Let E_i be this event.

If $x_j \in \Omega^c$, then there is a point $x \in B(x_i,h) \cap \partial\Omega$. Let $Q_j = \{x_k \in \mathsf{P}_0 : x_k + x_j \in \Omega^c\}$. For $x_k \in Q_j$, we use the decomposition

$$\begin{aligned} f(x_k+x_j) - f(x_k+x_i) = f_{\Omega^c}(x_k+x_j) - f_{\Omega^c}(x) \\ + f_{\Omega^c}(x) - f_\Omega(x) + f_\Omega(x) - f_\Omega(x_k+x_i), \end{aligned}$$

with the first and third differences bounded by $C_0 h$ in absolute value, and the second bounded from below by μ in absolute value. We therefore have

$$\delta_{ij}^2 := \|\mathbf{f}_{\mathsf{P}_j} - \mathbf{f}_{\mathsf{P}_i}\|_2^2 \geq \sum_{x_k \in Q_j} (f_\Omega(x_k+x_i) - f_{\Omega^c}(x_k+x_j))^2 \tag{8.28}$$

$$\geq |Q_j|(\mu - 2C_0 h)^2 \geq \delta^2 := m_{\mathsf{P}}(2C_0)^{-d}\mu^2/2, \tag{8.29}$$

where we used Lemma 8.2 to bound $|Q_j|$ from below and the fact that $\mu \asymp 1$ while $h = o(1)$. Since $\|\mathbf{y}_{\mathsf{P}_j} - \mathbf{y}_{\mathsf{P}_i}\|_2^2 \sim 2\sigma^2 \chi^2_{m_{\mathsf{P}}}(\delta_{ij}^2/(2\sigma^2))$ and $\delta_{ij} \geq \delta$, with Lemma 8.11 we see that

$$\min_j \|\mathbf{y}_{\mathsf{P}_j} - \mathbf{y}_{\mathsf{P}_i}\|_2^2 \geq \delta^2/4 + 2\sigma^2 m_{\mathsf{P}} - \zeta_\chi, \tag{8.30}$$

with probability at least $1-m^{1-C/2}$, where the minimum is over j such that $x_j \in \Omega^c \cap B(x_i,h)$. Let F_i denote this event.

Assuming (8.27) and (8.30) hold, we see that we need to choose h_y such that

$$2\sigma^2 m_{\mathsf{P}} + \zeta_\chi \leq h_y^2 < m_{\mathsf{P}}(2C_0)^{-d}\mu^2/8 + 2\sigma^2 m_{\mathsf{P}} - \zeta_\chi. \tag{8.31}$$

The lower bound is to ensure that all the points $x_j \in B(x_i,h)$ such that $\mathsf{P}_j \subset \Omega$ and $\mathsf{P}_j \cap \mathsf{P}_i = \emptyset$ are included in the neighborhood of x_i (meaning $\omega_{ij} = 1$), while the upper bound is to ensure that no points $x_j \in \Omega^c$ are included (under E_i). For all other points $x_j \in \Omega \cap B(x_i,h)$, they may or may not be included, depending on how large that intersection is. Note that there is an h_y satisfying (8.31) if, and only if,

$$m_{\mathsf{P}}(2C_0)^{-d}\mu^2/8 > 2\zeta_\chi \Leftrightarrow m_{\mathsf{P}} > C_1 \sigma^4 \log n,$$

for a constant C_1 which depends only on d, C_0, μ. Assuming m_{P} is that large, (8.31) is satisfied when $h_y^2 = 2(1+\eta)\sigma^2 m_{\mathsf{P}}$ with η sufficiently small. In any case, we assume that (8.31) holds and the rest of the proof is identical to the one for NLM-average.

Lower bound (heuristics). We discuss here the lower bound and where the issues are. Consider the important case where σ is fixed and assume that $f = \mathbb{1}_{\{\Omega\}}$, where $\Omega = (0,1/2) \times (0,1)^{d-1}$. Consider direct neighbors (*i.e.,* points with distance $1/n^d$) $x_i \in \Omega$ and $x_j \in \Omega^c$. For x_k such that $\mathsf{P}_k \cap (\mathsf{P}_i \cup \mathsf{P}_j) = \emptyset$, we use (8.19) to arrive at

$$\overline{y}_{\mathsf{P}_k} - \overline{y}_{\mathsf{P}_i} \sim \mathcal{N}(\lambda_k, 2\sigma^2/m_{\mathsf{P}}),$$

and

$$\overline{y}_{\mathsf{P}_k} - \overline{y}_{\mathsf{P}_j} \sim \mathcal{N}(\lambda'_k, 2\sigma^2/m_{\mathsf{P}}),$$

where $|\lambda_k - \lambda'_k| = |\mathsf{P}_i \setminus \mathsf{P}_j| = {m_\mathsf{P}}^{-1/d}$. When $d \leq 2$, the difference in means ${m_\mathsf{P}}^{-1/d}$ is of order at most that of the standard deviation ${m_\mathsf{P}}^{-1/2}$, so that these two distributions cannot be effectively separated. *Heuristically*, this indicates that if the photometric kernel of NLM-average includes x_k in the neighborhood of x_i, it also includes it in the neighborhood of x_j with non-negligible probability. This is evidence that the squared bias is of order 1 at these points. Since there are order $(nh)^{d-1}$ such sample points, averaging over them yields a lower bound on the squared bias (and therefore on the MSE) of order $O(1/n)$. The same heuristics could be applied to NLM.

The story changes when $d \geq 3$. In fact, for any f in the cartoon model with similar foreground, NLM-average — and NLM too — achieve a much better risk. To see this, fix $x_i \in \Omega$. We already know that NLM-average behaves well when $\mathsf{P}_i \subset \Omega$; therefore assume that $\mathsf{P}_i \cap \Omega \neq \emptyset$. If $x_j - x_i$ (as a vector) is not parallel to $\partial\Omega$, then $|\overline{f}_{\mathsf{P}_j} - \overline{f}_{\mathsf{P}_i}| \geq {m_\mathsf{P}}^{-1/d}$, so that under (8.19),

$$|\overline{y}_{\mathsf{P}_k} - \overline{y}_{\mathsf{P}_j}| \geq {m_\mathsf{P}}^{-1/d} - \zeta \geq {m_\mathsf{P}}^{-1/3} - \zeta.$$

Noting that $\zeta \asymp \sqrt{\log(n)/m_\mathsf{P}} = o({m_\mathsf{P}}^{-1/d})$, if we choose $h_y \asymp {m_\mathsf{P}}^{-2/5}$, then with high probability, the neighborhood of x_i only includes those $x_j \in B(x_i, h)$ such that $x_j - x_i$ is parallel to $\partial\Omega$, perhaps excluding those such that $\mathsf{P}_j \cap \mathsf{P}_i = \emptyset$. There are order $(nh)^{d-1}$ such x_j's, which drives the variance of the local polynomial estimator at x_i. This applies to all x_i with $\mathsf{P}_i \cap \Omega \neq \emptyset$, and there are order $n^d h$ such x_i's. The MSE over these points yields an MSE of order

$$\frac{1}{n^d}(n^d h)\left(h^{2\alpha} + \frac{\sigma^2}{(nh)^{d-1}}\right) = h^{2\alpha+1} + (nh^2)\frac{\sigma^2}{(nh)^d}.$$

We know that the MSE over the points away from the discontinuity is of order

$$h^{2\alpha} + \frac{\sigma^2}{(nh)^d},$$

so the overall MSE is of order

$$h^{2\alpha} + (nh^2 \vee 1)\frac{\sigma^2}{(nh)^d}.$$

Minimizing over h yields a lower bound of $(\sigma^2/n^{d-1})^{2\alpha/(2\alpha+d-2)} \vee (\sigma^2/n^d)^{2\alpha/(2\alpha+d)}$, which is the MO rate if $d \geq 2\alpha$.

Non-local versions. We quickly argue that, without spatial localization, YF, NLM, and NLM-average do not perform that well, unless the underlying function is a polynomial (of degree at most r) or all jumps are greater than h_y. Let's look at what the methods do on noiseless data. For a given photometric bandwidth h_y, consider the function $f = h_y \mathbb{1}_{\{\Omega\}}$, where $\Omega = (0, 1/2) \times (0, 1)^{d-1}$. Then both YF and NLM-average output a constant estimator equal everywhere to the local polynomial estimator applied to the whole image. Hence, the MSE is at least $h_y^2/4$. Given that we take h_y relatively large, this leads to a large MSE (of order 1).

8.2.6. Proof of Theorem 5.1. The only difference with the cartoon model is in the behavior of local polynomial regression. Fix a point $x_i \in \Omega$. By Lemma 8.4 (scaled by h), $\mathrm{Vol}(B(x_i, h) \cap \Omega) \asymp h^{d_0} a^{d-d_0}$. (In fact, this is slightly easier here since Ω is a band around the graph of a function.) Therefore, by Lemma 8.3 (scaled by h), we see that, $\#\{j : x_j \in B(x_i, h) \cap \Omega\} \asymp n^d h^{d_0} a^{d-d_0}$. This is the number of observations we are "averaging" over.

For LF, we prove a lower bound of order 1 for the squared bias at x_i; we proceed as in Lemma 8.7 with only cosmetic adjustments.

For BO, we apply LPR to the sample points x_j belonging to the largest ball centered at x_i which is contained in Ω. Since we only consider $x_i \in \Omega \setminus B(\partial\Omega, a/C)$, then this ball of radius at least a/C. We then conclude using the same argument bounding the risk of BO in the cartoon model detailed in Section 8.2.3.

For MO, and its mimickers YF and NLM, we need to refine Lemma 8.5 because, in the case where Ω is a thin band, the largest ball within it is not representative of the sample size used in the local polynomial fit — which is what drives the variance. We explain how to adapt the proof of Lemma 8.5 to show that

$$\mathrm{Var}(\widehat{f}_i) \le \frac{C\sigma^2}{n^d h^{d_0} a^{d-d_0}},$$

for a constant C. Let $u_1, \dots, u_m$ be a maximal a-packing of $B(x_i, h) \cap \Omega$, with $m \asymp (h/a)^{d_0}$. Then

$$\bigsqcup_{k=1}^{m} B(u_k, a) \subset \Omega.$$

Using the notation introduced in the proof of Lemma 8.5, we have

$$\mathbf{Z}^T \mathbf{Z} \succ \sum_{k=1}^{m} \mathbf{Z}_k^T \mathbf{Z}_k,$$

where $\mathbf{Z}_k = (z_j^s : x_j \in B(u_k, a), |s| \le r)$. We can then use (8.8) to obtain

$$\lambda_{\min}(\mathbf{Z}_k^T \mathbf{Z}_k) \ge \frac{1}{C}(na)^d,$$

implying

$$\lambda_{\min}(\mathbf{Z}^T \mathbf{Z}) \succ \frac{1}{C} m (na)^d.$$

This gives the upper bound on the variance, and the bias behaves as expected, meaning that Lemma 8.6 holds. It is now straightforward to deduce that MO at i has a squared bias of order $O(h^{2\alpha})$ and a variance of $O(\sigma^2 n^{-d} h^{-d_0} a^{-d+d_0})$. Given that $h = h^{\mathrm{MO}}$ and $a = o(h^{\mathrm{MO}})$, the variance dominates and may be expressed as $(h/a)^{d-d_0} O(\sigma^2/(nh)^d)$, with $O(\sigma^2/(nh)^d)$ being the order of magnitude of the point risk of MO under the cartoon model.

YF is still able to perfectly mimic MO under the conditions of Theorem 4.4 (same exact arguments).

For points $x_i \in \Omega$ with $\mathrm{dist}(x_i, \partial\Omega) > h_{\mathsf{P}}^{\mathrm{NLM}}$, the analysis for NLM is again exactly the same, the difference here being in the number of j's such that $\mathsf{P}_j \subset \Omega$, which is of order $n^d h^{d_0} a^{d-d_0}$. The rest is the same.

8.2.7. Proof of Proposition 5.1. Here Ω and Ω^c are interchangeable, so we focus on the former WLOG and fix point $x_i \in \Omega$. Again, the only difference with the cartoon model is in the behavior of local linear regression and we need a stronger version of Lemma 8.5 in the case where Ω is a repeated pattern. Using the notation introduced in the proof of Lemma 8.5, we have

$$\mathbf{Z}^T\mathbf{Z} = \sum_v \mathbf{Z}_v^T\mathbf{Z}_v,$$

where $\mathbf{Z}_v = (z_j^s : x_j \in B(x_i,h) \cap (\Xi + v), |s| \le r)$. Note that we may restrict the sum to those $v \in a\mathbb{Z}^d$ such that $B(x_i,h) \cap (\Xi + v) \neq \emptyset$, and there are order $(h/a)^d$ such v's. Since they are all translates of each other, let us focus on Ξ, that is, $v = 0$.

We again express $\mathbf{Z}_0^T\mathbf{Z}_0$ as a sum of matrices by partitioning the d-dimensional subgrid $\{x_j \in \Xi\}$ into discrete 1D grids of the form

$$L_{j_1,\dots,j_{d-1}} := \{((j_1 - 1/2)/n, \dots, (j_{d-1} - 1/2)/n, (j_d - 1/2)/n) \in \Xi : j_d = 1, \dots, [na]\},$$

where $j_1, \dots, j_{d-1} \in \{1, \dots, [na]\}$. We therefore have

$$\mathbf{Z}_0^T\mathbf{Z}_0 = \sum_{j'} \mathbf{Z}_{(j')}^T\mathbf{Z}_{(j')},$$

where $\mathbf{Z}_{(j')} := (z_k^s : x_k \in L_{j'} \cap \Omega, |s| \le r)$ for $j' \in \{1, \dots, [na]\}^{d-1}$.

Since $N_\Omega \ge (1/C)N_{\Omega^c}$, we also have that $N_\Xi \ge (1/C)N_{(0,a)^d \setminus \Xi}$, so that Ξ contains at least the fraction $1/(C+1)$ of the sample points in $(0,a)^d$ and therefore

$$\sum_{j' \in \{1,\dots,[na]\}^{d-1}} |L_{j'}| \ge \frac{[na]^d}{C+1}. \tag{8.32}$$

Let

$$J' := \{j' \in \{1, \dots, [na]\}^{d-1} : |L_{j'}| \ge [na]/(2C+2)\}.$$

Since $|L_{j'}| \le [na]$, we have

$$\sum_{j' \in \{1,\dots,[na]\}^{d-1}} |L_{j'}| \le [na]|J'| + \frac{[na]}{2C+2}([na]^{d-1} - |J'|),$$

so that $|J'| \ge [na]^{d-1}/(2C+1)$ by (8.32). We focus on $\mathbf{Z}_{(j')}$ with $j' \in J'$. Notice that this reduces the analysis to the one-dimensional case.

Lemma 8.12. *There is a numeric constant $C > 0$ such that any polynomial regression matrix of the form $\mathbf{U} = ((k/m)^s : 0 \le s \le r; k \in K)$, with $K \subset \{-m, \dots, m\}$ and $|K| \ge r+1$, satisfies $\lambda_{\min}(\mathbf{U}^T\mathbf{U}) \ge |K|(|K|/m)^{2r}/C$.*

Proof. Let $k_1 < \cdots < k_q$ be the elements of K. Define $\ell_0 = [q/(r+2)]$ and for $\ell = 1, \dots, \ell_0 - 1$, let $K_\ell = \{k_\ell, k_{\ell+\ell_0}, \dots, k_{\ell+(r+1)\ell_0}\}$. Note that $|K_\ell| = r+1$ and $k_{\ell+(j+1)\ell_0} - k_{\ell+j\ell_0} \ge \ell_0$. Now, the matrix $\mathbf{U}_\ell = ((k/m)^s : 0 \le s \le r; k \in K_\ell)$ is a Vandermonde $(r+1) \times (r+1)$ matrix. It is well-known that $\mathbf{U}_\ell$ is invertible, and more precisely, the main result in [18] says that

$$\|\mathbf{U}_\ell^{-1}\|_\infty = \max_{1\le i\le r+1} \prod_{j\in\{1,\dots,r+1\}\setminus\{i\}} \frac{1+|k_{\ell+j\ell_0}|/m}{|k_{\ell+j\ell_0}/m - k_{\ell+i\ell_0}/m|},$$

where $\|(a_{ij})\|_\infty := \max_i \sum_j |a_{ij}|$. Hence,

$$\|\mathbf{U}_\ell^{-1}\|_2 \le \sqrt{r+1}\|\mathbf{U}_\ell^{-1}\|_\infty \le \sqrt{r+1}(2m/\ell_0)^r,$$

where $\|\cdot\|_2$ is the usual Euclidean operator norm. Hence,

$$\lambda_{\min}(\mathbf{U}_\ell^T\mathbf{U}_\ell) \ge \|\mathbf{U}_\ell^{-1}\|_2^{-2} \ge (\ell_0/(2m))^{2r}/(r+1).$$

Since the index sets K_ℓ do not overlap, we have

$$\lambda_{\min}(\mathbf{U}^T\mathbf{U}) \ge \sum_{\ell=1}^{\ell_0} \lambda_{\min}(\mathbf{U}_\ell^T\mathbf{U}_\ell).$$

When r is fixed, $\ell_0 \asymp q$, so the RHS $\asymp q(q/m)^{2r}$. ■

Let C_1 denote the constant of Lemma 8.12 and let $C_2 = C_1(2C+1)^{2r+1}$. Applying this result, we find that $\lambda_{\min}(\mathbf{Z}_{(j')}^T\mathbf{Z}_{(j')}) \ge [na]/C_2$ for all $j' \in J'$.

Moving up, we have

$$\lambda_{\min}(\mathbf{Z}_0^T\mathbf{Z}_0) \ge (\#J')[na]/C_2 \ge \frac{[na]^d}{C_2(2C+2)},$$

and then

$$\lambda_{\min}(\mathbf{Z}^T\mathbf{Z}) \ge (h/a)^d\lambda_{\min}(\mathbf{Z}_0^T\mathbf{Z}_0) \asymp (nh)^d.$$

With this established, the bias behaves as in the cartoon model, and the rest of the analysis for MO and YF is exactly as before.

For NLM, some additional arguments are required. Specifically, we need to compare P_i with other the patches is centered at $x_j \in B(x_i, h)$. First, suppose that $x_j - x_i \in a\mathbb{Z}^d$. Then, by the periodicity of Ω, $x_j \in \Omega$ too, and also $x_k + x_i \in \Omega$ if, and only if, $x_k + x_j \in \Omega$, for all $x_k \in \mathsf{P}_0$. Hence,

$$\begin{aligned}\|\mathbf{f}_{\mathsf{P}_j} - \mathbf{f}_{\mathsf{P}_i}\|_2^2 &= \sum_{x_k\in\mathsf{P}_0\cap\Omega} (f_\Omega(x_k+x_j) - f_\Omega(x_k+x_i))^2 \\ &\quad + \sum_{x_k\in\mathsf{P}_0\cap\Omega^c} (f_{\Omega^c}(x_k+x_j) - f_{\Omega^c}(x_k+x_i))^2 \\ &\le m_{\mathsf{P}} C_0^2\|x_i - x_j\|^2 \le m_{\mathsf{P}} C_0^2 h^2.\end{aligned}$$

This is the equivalent of (8.26).

Suppose now that $x_j \in \Omega^c$. Using the fact that f_Ω and f_{Ω^c} are C_0-Lipschitz, we have

$$\mathbf{f}_{\mathsf{P}_i} = f_\Omega(x_i)\mathbb{1}(\mathsf{P}_i \cap \Omega) + f_{\Omega^c}(x_i)\mathbb{1}(\mathsf{P}_i \cap \Omega^c) + O(h),$$

and similarly,

$$\mathbf{f}_{\mathsf{P}_j} = f_\Omega(x_i)\mathbb{1}(\mathsf{P}_j \cap \Omega) + f_{\Omega^c}(x_i)\mathbb{1}(\mathsf{P}_j \cap \Omega^c) + O(h),$$

since $f_\Omega(x_i) - f_\Omega(x_j) = O(h)$ and $f_{\Omega^c}(x_i) - f_{\Omega^c}(x_j) = O(h)$. Hence,

$$\begin{aligned}\|\mathbf{f}_{\mathsf{P}_j} - \mathbf{f}_{\mathsf{P}_i}\|_2^2 &\geq (f_{\Omega^c}(x_i) - f_\Omega(x_i))^2 \\ &\quad \times \|\mathbb{1}(\mathsf{P}_j \cap \Omega) - \mathbb{1}(\mathsf{P}_i \cap \Omega)\|_2^2 + O(m_\mathsf{P} h^2) \\ &\geq \mu^2 m_\mathsf{P}/C' + O(m_\mathsf{P} h^2)\end{aligned}$$

by (5.1). This is the equivalent of (8.29).

Arguing as the in the proof of Theorem 4.5, we see that, with high probability, the regression neighborhood of x_i includes all x_j such that $x_j - x_i \in a\mathbb{Z}^d$, $x_j \in B(x_i, h)$ and $\mathsf{P}_j \cap \mathsf{P}_i = \emptyset$ — those x_j's are in Ω like x_i — and excludes all $x_j \in \Omega^c$ such that $\mathsf{P}_j \cap \mathsf{P}_i = \emptyset$. In principle, an additional argument would be needed to exclude those $x_j \in \Omega^c$ such that $\mathsf{P}_j \cap \mathsf{P}_i \neq \emptyset$, since in that case $\|\boldsymbol{\varepsilon}_{\mathsf{P}_j} - \boldsymbol{\varepsilon}_{\mathsf{P}_i}\|_2^2$ is not chi-square as before. However, it is not hard to see that even if these are included in the regression neighborhood, it does not change things much since their number is small — of order $O(\log n)$.

Acknowledgements. The second author would like to thanks Michaël Chichignoud for fruitful comments that helped improving this work.

REFERENCES

[1] Alvarez, L., Mazorra, L.: Signal and image restoration using shock filters and anisotropic diffusion. SIAM J. Numer. Anal. **31**(2), 590–605 (1994)
[2] Arias-Castro, E., Donoho, D.L.: Does median filtering truly preserve edges better than linear filtering? Ann. Statist. **37**(3), 1172–1206 (2009)
[3] Awate, S.P., Whitaker, R.T.: Unsupervised, information-theoretic, adaptive image filtering for image restoration. IEEE Trans. Pattern Anal. Mach. Intell. **28**(3), 364–376 (2006)
[4] Azzabou, N., Paragios, N., Guichard, F.: Image denoising based on adapted dictionary computation. In: ICIP, pp. 109–112 (2007)
[5] Buades, A.: Image and movie denoising by non local means. Ph.D. thesis, Universitat de les Illes Balears (2006)
[6] Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. Multiscale Model. Simul. **4**(2), 490–530 (2005)
[7] Chatterjee, P., Milanfar, P.: Patch-based near-optimal image denoising. submitted (2011)
[8] Criminisi, A., Pérez, P., Toyama, K.: Region filling and object removal by exemplar-based image inpainting. IEEE Trans. Image Process. **13**(9), 1200–1212 (2004)
[9] Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.O.: Image denoising by sparse 3-D transform-domain collaborative filtering. IEEE Trans. Image Process. **16**(8), 2080–2095 (2007)
[10] Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.O.: BM3D image denoising with shape-adaptive principal component analysis. In: Proc. Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS'09) (2009)
[11] Deledalle, C.A., Duval, V., Salmon, J.: Anisotropic non-local means with spatially adaptive patch shapes. In: SSVM (2011)
[12] Deledalle, C.A., Duval, V., Salmon, J.: Non-local methods with shape-adaptive patches (NLM-SAP). J. Math. Imaging Vis. pp. 1–18 (2011)
[13] Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)

[14] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Wavelet shrinkage: asymptopia? J. Roy. Statist. Soc. Ser. B **57**(2), 301–369 (1995)
[15] Duval, V., Aujol, J.F., Gousseau, Y.: A bias-variance approach for the nonlocal means. SIAM J. Imaging Sciences **4**(2), 760–788 (2011)
[16] Efros, A.A., Leung, T.: Texture synthesis by non-parametric sampling. In: ICCV, pp. 1033–1038 (1999)
[17] Fan, J., Gijbels, I.: Local polynomial modelling and its applications, *Monographs on Statistics and Applied Probability*, vol. 66. Chapman & Hall, London (1996)
[18] Gautschi, W.: On inverses of Vandermonde and confluent vandermonde matrices. Numerische Mathematik **4**, 117–123 (1962)
[19] Gilboa, G., Osher, S.: Nonlocal linear image regularization and supervised segmentation. Multiscale Model. Simul. **6**(2), 595–630 (2007)
[20] Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning, second edn. Springer Series in Statistics. Springer, New York (2009)
[21] Johnstone, I.M.: Oracle inequalities and nonparametric function estimation. In: Proceedings of the International Congress of Mathematicians, Vol. III (Berlin, 1998), Extra Vol. III, pp. 267–278 (electronic) (1998)
[22] Katkovnik, V.: A new method for varying adaptive bandwidth selection. IEEE Trans. Image Process. **47**(9), 2567–2571 (1999)
[23] Katkovnik, V., Egiazarian, K.O., Astola, J.T.: Adaptive window size image de-noising based on intersection of confidence intervals (ICI) rule. J. Math. Imaging Vis. **16**(3), 223–235 (2002)
[24] Katkovnik, V., Foi, A., Egiazarian, K.O., Astola, J.T.: Directional varying scale approximations for anisotropic signal processing. In: EUSIPCO, pp. 101–104 (2004)
[25] Kervrann, C., Boulanger, J.: Optimal spatial adaptation for patch-based image denoising. IEEE Trans. Image Process. **15**(10), 2866–2878 (2006)
[26] Korostelëv, A.P., Tsybakov, A.B.: Minimax theory of image reconstruction, *Lecture Notes in Statistics*, vol. 82. Springer-Verlag, New York (1993)
[27] Lee, J.S.: Digital image smoothing and the sigma filter. Computer Vision, Graphics, and Image Processing **24**(2), 255–269 (1983)
[28] Lepski, O.V., Mammen, E., Spokoiny, V.G.: Optimal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selectors. Ann. Statist. **25**(3), 929–947 (1997)
[29] Levin, A., Nadler, B.: Natural image denoising: Optimality and inherent bounds. In: CVPR (2011)
[30] Mahmoudi, M., Sapiro, G.: Fast image and video denoising via nonlocal means of similar neighborhoods. IEEE Signal Process. Lett. **12**, 839–842 (2005)
[31] Mairal, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Non-local sparse models for image restoration. In: ICCV, pp. 2272–2279 (2009)
[32] Maleki, A., Narayan, M., Baraniuk, R.G.: Anisotropic nonlocal means (2011). Submitted to *Applied and Computational Harmonic Analysis*
[33] Maleki, A., Narayan, M., Baraniuk, R.G.: Suboptimality of nonlocal means for images with sharp edges (2011). Submitted to *Applied and Computational Harmonic Analysis*
[34] Mallat, S.: A wavelet tour of signal processing. Elsevier/Academic Press, Amsterdam (2009). The sparse way, With contributions from Gabriel Peyré
[35] Müller, H.G., Stadtmüller, U.: Variable bandwidth kernel estimators of regression curves. Ann. Statist. **15**(1), 182–201 (1987)
[36] Nadaraya, E.A.: On estimating regression. Theory of Probability and its Applications **9**(1), 141–142 (1964)
[37] Perona, P., Malik, J.: Scale space and edge detection using anisotropic diffusion. IEEE Trans. Pattern Anal. Mach. Intell. **12**, 629–639 (1990)
[38] Polzehl, J., Spokoiny, V.G.: Adaptive weights smoothing with applications to image restoration. J. R. Stat. Soc. Ser. B Stat. Methodol. **62**(2), 335–354 (2000)
[39] Polzehl, J., Spokoiny, V.G.: Image denoising: pointwise adaptive approach. Ann. Statist. **31**(1), 30–57 (2003)
[40] Portilla, J., Strela, V., Wainwright, M., Simoncelli, E.P.: Image denoising using scale mixtures of gaussians in the wavelet domain. IEEE Trans. Image Process. **12**(11), 1338–1351 (2003)

[41] Salmon, J.: On two parameters for denoising with Non-Local Means. IEEE Signal Process. Lett. **17**, 269–272 (2010)
[42] Salmon, J., Strozecki, Y.: Patch reprojections for Non Local methods. Signal Processing **In press** (2011)
[43] Singer, A., Shkolnisky, Y., Nadler, B.: Diffusion interpretation of nonlocal neighborhood filters for signal denoising. SIAM J. Imaging Sci. **2**(1), 118–139 (2009)
[44] Smith, S.M., Brady, J.M.: Susan-a new approach to low level image processing. International Journal of Computer Vision **23**(1), 45–78 (1997)
[45] Spokoiny, V.G.: Estimation of a function with discontinuities via local polynomial fit with an adaptive window choice. Ann. Statist. **26**(4), 1356–1378 (1998)
[46] Starck, J.L., Candès, E.J., Donoho, D.L.: The curvelet transform for image denoising. IEEE Trans. Image Process. **11**(6), 670–684 (2002)
[47] Szlam, A.D., Maggioni, M., Coifman, R.R.: Regularization on graphs with function-adapted diffusion processes. J. Mach. Learn. Res. **9**, 1711–1739 (2008)
[48] Tasdizen, T.: Principal neighborhood dictionaries for nonlocal means image denoising. IEEE Trans. Image Process. **18**(12), 2649–2660 (2009)
[49] Tomasi, C., Manduchi, R.: Bilateral filtering for gray and color images. In: ICCV, pp. 839–846 (1998)
[50] Tsybakov, A.B.: Introduction to nonparametric estimation. Springer Series in Statistics. Springer, New York (2009)
[51] Van De Ville, D., Kocher, M.: SURE-based Non-Local Means. IEEE Signal Process. Lett. **16**, 973–976 (2009)
[52] Van De Ville, D., Kocher, M.: Non-local means with dimensionality reduction and sure-based parameter selection. IEEE Trans. Image Process. **99**(99), 1–1 (2011)
[53] Watson, G.S.: Smooth regression analysis. Sankhya: The Indian Journal of Statistics, Series A **26**(4), 359–372 (1964)
[54] Weissman, T., Ordentlich, E., Seroussi, G., Verdú, S., Weinberger, M.J.: Universal discrete denoising: known channel. IEEE Trans. Inf. Theory **51**(1), 5–28 (2005)
[55] Yaroslavsky, L.P.: Digital picture processing, *Springer Series in Information Sciences*, vol. 9. Springer-Verlag, Berlin (1985)
[56] Zewail, A.H., Thomas, J.M.: 4D electron microscopy: imaging in space and time. Imperial College Pr (2009)