CONVERGENCE THEOREMS FOR THE NON-LOCAL MEANS FILTER

QIYU JIN
School of mathematical science, Inner Mongolia University
No.235 Daxuexilu Road, 010021 Hohhot, Inner Mongolia, China

ION GRAMA
UMR 6205, Laboratoire de Mathématiques de Bretagne Atlantique
Université de Bretagne-Sud
Campus de Tohannic, BP 573, 56017 Vannes, France

QUANSHENG LIU
UMR 6205, Laboratoire de Mathématiques de Bretagne Atlantique
Université de Bretagne-Sud
Campus de Tohannic, BP 573, 56017 Vannes, France
School of Computer and Software
Nanjing University of Information Science and Technology
Nanjing 210044, China

(Communicated by Jean-Michel Morel)

Abstract. We introduce an oracle filter for removing the Gaussian noise with weights depending on a similarity function. The usual Non-Local Means filter is obtained from this oracle filter by substituting the similarity function by an estimator based on similarity patches. When the sizes of the search window are chosen appropriately, it is shown that the oracle filter converges with the optimal rate. The same optimal convergence rate is preserved when the similarity function has suitable errors-in-measurements. We also provide a statistical estimator of the similarity which converges at a convenient rate. Based on our convergence theorems, we propose some simple formulas for the choice of the parameters. Simulation results show that our choice of parameters improves the restoration quality of the filter compared with the usual choice of parameters in the original algorithm.

1. Introduction. We deal with the additive Gaussian noise model

\[ v(x) = u(x) + \epsilon(x), \quad x \in I, \]

where \( I \) is the uniform \( N \times N \) grid of pixels on the unit square

\[ I = \left\{ \frac{1}{N}, \frac{2}{N}, \ldots, \frac{N-1}{N}, 1 \right\}^2, \quad N > 1, \]

\( v = (v(x))_{x \in I} \) is the observed image brightness, \( u : [0, 1]^2 \to \mathbb{R}_+ \) is the original image (unknown target regression function) and \( \epsilon = (\epsilon(x))_{x \in I} \) are independent and identically distributed (i.i.d.) Gaussian random variables with mean 0 and standard deviation \( \sigma > 0 \).

2010 Mathematics Subject Classification. 62H35.
Key words and phrases. Image denoising, oracle filter, Non-Local Means filter, weighted means, convergence, mean squared error.

* Corresponding author: quansheng.liu@univ-ubs.fr.
Important denoising techniques for the model (1) have been developed in recent years. A very significant step in these developments was the introduction of the Non-Local Means filter by Buades, Coll and Morel [5]. For related works, see for example [36, 27, 7, 25, 32].

The basic idea of the filters by weighted means is to estimate the unknown image \( u(x_0) \) by a weighted average of observations \( v(x) \) of the form

\[
\tilde{u}_w(x_0) = \sum_{x \in N_{x_0,D}} w(x) v(x),
\]

where for each \( x \) and \( D \) an odd integer, \( N_{x_0,D} \) denotes a square window with center \( x_0 \) and size \( D \times D \):

\[
N_{x_0,D} = \left\{ x \in I : \| x - x_0 \|_\infty \leq \frac{D-1}{2N} \right\},
\]

with \( \| z \|_\infty = \max\{ |z_1|, |z_2| \} \) for \( z = (z_1, z_2) \), \( w(x) \) are some non-negative weights satisfying \( \sum_{x \in N_{x_0,D}} w(x) = 1 \). The choice of the weight \( w(x) \) is usually based on two criteria: a spatial criterion so that \( w(x) \) is a decreasing function of the distance between \( x \) and \( x_0 \), and a similarity criterion so that \( w(x) \) is also a decreasing function of the brightness difference \( |v(x) - v(x_0)| \) which measures the similarity of the observed image at the pixels \( x \) and \( x_0 \). We refer, for example to the bilateral filter [40] which uses these two criteria. In the Non-Local Means filter, \( D \) can be chosen relatively large, and the weights \( w(x) \) are calculated according to the similarity between data patches

\[
v(N_{x,d}) = (v(y) : y \in N_{x,d})
\]

(which is identified as a vector whose components are ordered lexicographically) and \( v(N_{x_0,d}) = (v(y) : y \in N_{x_0,d}) \), instead of the similarity between just the pixels \( x \) and \( x_0 \) (this is the main difference with the bilateral filter). Here \( N_{x,d} \) is a square window with center \( x \) and size \( d \times d \) defined exactly as \( N_{x_0,D} \) with \( x_0 \) and \( D \) replaced by \( x \) and \( d \) respectively. In the sequel \( N_{x,D} \) will be called search window, while \( N_{x,d} \) will be called similarity patch.

The Non-Local Means filter was further enhanced for speed in subsequent works such as [33, 3, 23, 42]. For some other improvements we refer to [26, 27, 8, 6, 9, 10, 34, 35, 41, 30, 43, 28, 29, 4, 11]. Katkovnik et al. [25] review the evolution of the non-parametric regression modeling in imaging from the local Nadaraya-Watson kernel estimate to the Non-Local Means filter and further to transform-domain filtering based on non-local block-matching. In [2, 18, 39, 31] some theoretical aspects of the Non-Local Means filter have been investigated.

The goal of this paper is to give some new mathematical insights to explain further how the Non-Local Means filter works in theory and to highlight its potential and properties. We use the statistical estimation and optimization techniques to give a justification of the filter, and to suggest the order of magnitudes of the search window and similarity patches. Our main idea is to optimize (up to the rate of convergence) a tight upper bound of the \( L^2 \) risk

\[
R(\tilde{u}_w(x_0)) = \mathbb{E}(\tilde{u}_w(x_0) - u(x_0))^2
\]

by changing the widths of the search window and similarity patches. We first introduce an oracle filter \( u^* \) with weights depending on the unknown image function \( u \). The oracle \( u^* \) is shown to converge at the optimal rate and to have high performance in numerical simulations. To mimic the oracle \( u^* \), we estimate \( u^* \) by some
Convergence of the Non-Local Means filter

adaptive weights $\hat{w}$ based on the observed image $v$. We thus deduce the Non-Local Means filter and we then prove that it converges at the optimal rate when the sizes of the search window and similarity patches are chosen properly. In particular our result suggests to take the similarity patch size $d$ larger than the search window size $D$. This is confirmed by our simulations, cf. for instance Fig. 4.

In the proofs of our main results we use the bias variance approach as developed in [20, 22, 21]. We mention that the bias-variance approach was also used in [13] to evaluate locally the bandwidth parameter $H$. Their approach is very different from ours, since we are computing the optimal rate of convergence of some estimators and we consider the choice of the search window and the similarity patch sizes $D$ and $d$.

The paper is organized as follows. In Section 2, we introduce an oracle estimator and prove its convergence. In Section 3 we use the oracle estimator to deduce the Non-Local Means filter by substituting the non-observable quantities by statistical estimators based on similarity patches. To justify the use of the Non-Local Means filter we show that measuring the similarity function with specified precision does not change the rate of convergence of the oracle filter and show that the corresponding statistical estimators attain the required precision. In Section 4 we propose simple expressions for the choice of parameters, and present our simulation results with a brief analysis. The proofs of theorems are given in Section 5. Section 6 concludes the paper.

2. Oracle filter and its convergence. Consider the function

$$\rho_{x_0}(x) \equiv |u(x) - u(x_0)|$$

which characterizes the similarity of the image brightness at the pixel $x$ with respect to the pixel $x_0$. In the sequel we shall call $\rho_{x_0}$ similarity function. Introduce a neighborhood filter in the Yaroslavski sense (see e.g. [44]) as follows: for any $x_0 \in I$ set

$$u^*(x_0) = \sum_{x \in N(x_0,D)} w^*(x)u(x),$$

where the weights are defined by

$$w^*(x) = e^{-\frac{\rho_{x_0}^2(x)}{2H^2}} \sum_{y \in N(x_0,D)} e^{-\frac{\rho_{x_0}^2(y)}{2H^2}}, \quad x \in N(x_0,D)$$

with $H > 0$ a parameter. The weights satisfy

$$\sum_{x \in N(x_0,D)} w^*(x) = 1 \quad \text{and} \quad w^*(x) \geq 0.$$

Since the similarity function $\rho_{x_0}$ depends on the unknown image $u$, the filter $u^*$ is in fact not computable and is called oracle (for details on this concept see Donoho and Johnstone (1994 [12])). A computable filter will be obtained in the next section from the oracle filter by replacing the unknown similarity function $\rho_{x_0}$ by an estimator based on similarity patches.

We shall prove the convergence of the oracle filter under the local Hölder condition: for any $x_0 \in I$ there are some constants $\beta \in (0,1]$ and $L > 0$ such that

$$|u(x) - u(y)| \leq L\|x - y\|_\infty^\beta, \quad \forall x, y \in N_{x_0,D+d}.$$
This hypothesis seems reasonable especially when the Hölder regularity is low. In particular it is allowed that $\beta$ is very small so that the image can be rather irregular. For example, the Brownian motion, which is considered to be rather irregular, has the Hölder regularity $\beta = 1/2$. Moreover the Hölder regularity is assumed to be satisfied locally, so the image can be irregular at some locations and smooth in other locations. Note also that the Hölder condition is much weaker than the Lipschitz condition (which corresponds to the case $\beta = 1$) assumed in some other papers, see e.g. Buades et al. [5].

Let
\begin{equation}
\Delta = \frac{D - 1}{2N} \quad \text{and} \quad n = N^2
\end{equation}
be respectively the width (half of the edge size) of the search window and the number of pixels of the image. In the following theorem we prove that when the width of the search window $\Delta$ is properly chosen, the Mean Squared Error of the oracle estimator $u^*(x_0)$ converges at the rate $n^{-\beta/2\beta+1}$.

**Theorem 2.1.** Assume that the function $u$ satisfies the local Hölder condition (9). Suppose also that $\Delta = \left(\frac{\sigma^2}{9L^2}\right)^{\frac{1}{2\beta+1}} n^{-\frac{1}{2\beta+2}}$ and $H \geq L\Delta^\beta$. Let $u^*(x_0)$ be given by (6). Then
\begin{equation}
\mathbb{E} (u^*(x_0) - u(x_0))^2 \leq c n^{-\frac{\beta}{2\beta+1}} \quad \text{with} \quad c = \frac{9^{\frac{\beta+2}{2\beta+1}} \sigma^{\frac{2\beta}{2\beta+1}} L^{\frac{2}{2\beta+1}}}{2^{\frac{2}{2\beta+1}}}
\end{equation}

The proof of this theorem is deferred to Section 5.2.

Let us comment briefly on this result. There are well-known classical results in the statistical estimation theory, where Hölder smooth functions are estimated. From this general theory the rate $n^{-\beta/2\beta+1}$ is known to be optimal for a given Hölder smoothness $\beta > 0$ (see e.g. Fan and Gijbels (1996 [14])). Here we obtain the same rate of convergence for the oracle estimator $u^*(x_0)$ which depends on the unobserved image function $u$ though the similarity function $\rho_{x_0}$. To obtain a computable filter, in Section 3 we will replace the similarity function $\rho_{x_0}$ by a patch based estimator. Moreover, we will show that when the similarity function $\rho_{x_0}$ gets perturbed by a suitable error, the rate of convergence of the mean squared error of the corresponding estimator remains the optimal one $n^{-\beta/2\beta+1}$ (see Theorem 3.1). To the best of our knowledge, our result is the first one for determining the rate of convergence of a patch based estimator in image denoising.

Our simulations show that the difference between the oracle $u^*(x_0)$ and the true value $u(x_0)$ is extremely small (see Table 1), which suggests that a good filter can be obtained if we have a consistent estimator of the similarity function $\rho_{x_0}^2$.

**3. Non-Local Means filter via the oracle filter.** We now deduce the Non-Local Means filter [5] via the oracle filter by giving a statistical estimator of the similarity function $\rho_{x_0}^2$.

For any $x_0 \in I$ and any $x \in \mathcal{N}_{x_0,D}$, we measure the similarity between the data patches $v(\mathcal{N}_{x,d})$ and $v(\mathcal{N}_{x_0,d})$ by the quantity
\begin{equation}
\|v(\mathcal{N}_{x,d}) - v(\mathcal{N}_{x_0,d})\|_2^2 = \frac{1}{d^2} \sum_{y \in \mathcal{N}_{x_0,d}} (v(T_{x_0,x}y) - v(y))^2,
\end{equation}
TABLE 1. The PSNR when the oracle estimator \( u^* \) is applied with different values of \( D \).

| Image | Lena | Barbara | Boats | House | Peppers |
|-------|------|---------|-------|-------|---------|
| Size  | 512 × 512 | 512 × 512 | 512 × 512 | 256 × 256 | 256 × 256 |
| \( \sigma / \text{PSNR} \) | 10/28.12db | 10/28.12db | 10/28.12db | 10/28.11db | 10/28.11db |
| 4 × 9 | 32.96db | 31.33db | 32.92db | 31.34db | 32.89db |
| 11 × 11 | 40.12db | 38.49db | 38.89db | 40.04db | 38.85db |
| 13 × 13 | 41.09db | 39.55db | 39.78db | 40.98db | 39.64db |
| 15 × 15 | 41.92db | 40.45db | 40.63db | 41.77db | 40.39db |
| 17 × 17 | 42.64db | 41.23db | 41.39db | 42.40db | 41.00db |
| 19 × 19 | 43.29db | 41.93db | 42.06db | 43.06db | 41.58db |
| 21 × 21 | 43.88db | 42.57db | 42.67db | 43.61db | 42.14db |
| \( \sigma / \text{PSNR} \) | 20/22.11db | 20/22.11db | 20/22.11db | 20/28.12db | 20/28.12db |
| 9 × 9 | 33.61db | 31.91db | 32.32db | 33.72db | 32.62db |
| 11 × 11 | 34.78db | 33.20db | 33.49db | 34.92db | 33.65db |
| 13 × 13 | 35.80db | 34.28db | 34.49db | 35.98db | 34.51db |
| 15 × 15 | 36.69db | 35.22db | 35.40db | 36.80db | 35.26db |
| 17 × 17 | 37.48db | 36.05db | 36.20db | 37.48db | 35.89db |
| 19 × 19 | 38.17db | 36.74db | 36.90db | 38.07db | 36.45db |
| 21 × 21 | 38.80db | 37.40db | 37.54db | 38.67db | 36.98db |
| \( \sigma / \text{PSNR} \) | 30/18.60db | 30/18.60db | 30/18.60db | 30/18.61db | 20/28.12db |
| 9 × 9 | 30.65db | 28.89db | 29.29db | 30.90db | 29.51db |
| 11 × 11 | 31.86db | 30.25db | 30.45db | 31.90db | 30.51db |
| 13 × 13 | 32.85db | 31.33db | 31.49db | 32.92db | 31.34db |
| 15 × 15 | 33.74db | 32.27db | 32.37db | 33.76db | 32.68db |
| 17 × 17 | 34.50db | 33.09db | 33.16db | 34.48db | 32.74db |
| 19 × 19 | 35.20db | 33.81db | 33.85db | 35.13db | 33.32db |
| 21 × 21 | 35.79db | 34.46db | 34.48db | 35.71db | 33.85db |

where \( T_{x_0,x} \) is the translation which maps \( N_{x_0,d} \) to \( N_{x,d} \):

\[
T_{x_0,x}y = x + (y - x_0).
\]

We call \( \| v(N_{x,d}) - v(N_{x_0,d}) \|_2 \) similarity norm. Note that in the definition (12) of the similarity norm we have included the normalization factor \( d^2 \), which makes the parameter \( H \) in (16) independent of \( d \). Since \( E(\| v(T_{x_0,x}y) - v(y) \|_2^2 = (\| u(T_{x_0,x}y) - u(y) \|_2^2 + 2\sigma^2) \) for all \( y \in N_{x_0,d} \), we have

\[
E\| v(N_{x,d}) - v(N_{x_0,d}) \|_2^2 - 2\sigma^2 = \frac{1}{d^2} \sum_{y \in N_{x_0,d}} (u(T_{x_0,x}y) - u(y))^2
\]

\[
\approx \| u(x) - u(x_0) \|_2^2 = \rho_{x_0}^2(x),
\]

where the approximation will be shown to have a good convergence rate, provided that \( u \) satisfies the local Hölder condition (9): cf. Lemma 5.2, Section 5.3. This suggests the following asymptotically unbiased estimator of \( \rho_{x_0}^2(x) \):

\[
\hat{\rho}_{x_0}^2(x) = \| v(N_{x,d}) - v(N_{x_0,d}) \|_2^2 - 2\sigma^2 \approx \rho_{x_0}^2(x).
\]

Actually we will prove that \( \hat{\rho}_{x_0}^2(x) \) converges to \( \rho_{x_0}^2(x) \) in probability with a good convergence rate: see Theorem 3.2 below. In the oracle filter (6) replacing \( \hat{\rho}_{x_0}^2(x) \) by its estimator \( \hat{\rho}_{x_0}^2(x) \), we obtain the following adaptive filter, which is just the Non-Local Means filter:

\[
\hat{u}(x_0) = \sum_{x \in N_{x_0,v}} \hat{w}(x)v(x),
\]
where

\[
\hat{w}(x) = e^{-\frac{\rho_{\hat{x}_0}(x) + \epsilon}{2n_H^2}} \sum_{y \in N_{x_0,D}} e^{-\frac{\rho_{\hat{x}_0}(y)}{2n_H^2}}
\]

(16)

\[
= e^{-\frac{\|v(N_{x_0,D}) - v(N_{x_0,D})\|_2^2}{2n_H^2}} \sum_{y \in N_{x_0,D}} e^{-\frac{\|v(N_{x_0,D}) - v(N_{x_0,D})\|_2^2}{2n_H^2}}
\]

Note that in the obtained formula for the weights, the variance \(\sigma^2\) of the noise is not involved.

We shall prove that when \(\rho_{\hat{x}_0}^2(x)\) is perturbed, say replaced by

(17)

\[
\tilde{\rho}_{\hat{x}_0}^2(x) = \rho_{\hat{x}_0}^2(x) + \epsilon_n,
\]

where \(\epsilon_n\) is a suitable error term, then we still have the convergence of the corresponding filter with the optimal convergence rate. For any \(x_0 \in \mathbf{I}\), set

(18)

\[
\tilde{w}^*(x_0) = \sum_{x \in N_{x_0,D}} \tilde{w}^*(x)v(x),
\]

where the weights are defined by

(19)

\[
\tilde{w}^*(x) = e^{-\frac{\tilde{\rho}_{\hat{x}_0}^2(x)}{2n_H^2}} \sum_{y \in N_{x_0,D}} e^{-\tilde{\rho}_{\hat{x}_0}^2(y)}, \quad x \in N_{x_0,D}
\]

with \(H > 0\) a parameter. Then as an extension of Theorem 2.1 we have:

Theorem 3.1. Assume that the function \(u\) satisfies the local Hölder condition (9).
Suppose also that \(\Delta = \left(\frac{\sigma^2}{g_D^2}\right)^{\frac{1}{3\alpha+1}} n^{-\frac{1}{2\alpha+1}}, H \geq L\Delta^\beta\). Let \(\tilde{\rho}_{\hat{x}_0}^2(x)\) be defined by
(17) with \(\epsilon_n \to 0\) as \(n \to \infty\), and let \(\tilde{w}^*(x_0)\) be given by (18). Then

(20)

\[
\mathbb{E}(\tilde{w}^*(x_0) - u(x_0))^2 \leq 2ce^{2|\epsilon_n|/H^2} n^{-\frac{\beta}{2\alpha+1} + 2|\epsilon_n|},
\]

with \(c\) defined by (11). In particular, if \(\epsilon_n = O(n^{-\left(\frac{1}{2} - \alpha\right)} \ln^b n)\) for some \(0 < \alpha < \frac{1}{2(\beta + 1)}\) and \(b > 0\), then

(21)

\[
\mathbb{E}(\tilde{w}^*(x_0) - u(x_0))^2 = O(n^{-\frac{\beta}{2\alpha+1}}).
\]

This theorem shows that when the error term \(\epsilon_n\) is small enough, then the filter \(\tilde{w}^*(x_0)\) converges at the optimal rate \(O(n^{-\frac{\beta}{2\alpha+1}})\), just as in the case of the oracle filter (6).

In practice, we will take \(\tilde{\rho}_{\hat{x}_0}^2(x)\) as \(\hat{\rho}_{\hat{x}_0}^2(x)\) defined in (14). In fact, the following theorem proves that the estimator \(\tilde{\rho}_{\hat{x}_0}^2(x)\) of the similarity function \(\rho_{\hat{x}_0}^2(x)\) converges in probability, uniformly in \(x \in N_{x_0,D}\) and \(x_0 \in \mathbf{I}\), with the error term \(\epsilon_n\) satisfying

\[
\epsilon_n = O(n^{-\left(\frac{1}{2} - \alpha\right)} \ln^b n)\text{ for } b = 1/2.
\]

Recall that \(\Delta\) and \(n\) are respectively the width of the search window and the number of pixels of the image defined by (10). Let \(\delta\) be the width of the similarity patches:

(22)

\[
\delta = \frac{d - 1}{2N}.
\]

Theorem 3.2. Assume that the image \(u\) satisfies the local Hölder condition (9).
Suppose also that \(\Delta = \left(\frac{\sigma^2}{g_D^2}\right)^{\frac{1}{3\alpha+1}} n^{-\frac{1}{2\alpha+1}}\) and that \(\delta = c_0 n^{-\alpha}\) for some \(\frac{1}{2(\beta + 1)} <
α < \frac{1}{2} and c_0 > 0. Let \hat{\rho}_{x_0}^2 be given by (14). Then there is a constant c > 0 depending only on L, β, α and c_0 such that

\begin{equation}
\max_{x_0 \in I} \max_{x \in \mathcal{N}_{x_0,D}} \lim_{n \to \infty} P \left( \left| \hat{\rho}_{x_0}^2(x) - \rho_{x_0}^2(x) \right| \geq cn^{-\left(\frac{1}{2} - \alpha\right)} \sqrt{\ln n} \right) = 0.
\end{equation}

The proof will be given in Section 5.3.

Theorem 3.2 shows that, when we replace \rho_{x_0}^2(x) by its estimator \hat{\rho}_{x_0}^2(x), the error term \varepsilon_n satisfies \varepsilon_n = O(n^{-\left(\frac{1}{2} - \alpha\right)} \sqrt{\ln n}) with probability close to 1. Therefore, from Theorem 3.1, we see that if we choose \Delta = c_1 n^{-\left(\frac{1}{2} - \alpha\right)} and \delta = c_0 n^{-\alpha} with 0 < \alpha < \frac{1}{2(\beta+1)} , we obtain the optimal convergence rate \sqrt{\frac{\ln n}{n}} . This suggests that choosing \delta > \Delta would give good denoising results. This will be confirmed by our simulations in Section 4.

Define the following smoothed version of the distance (12):

\begin{equation}
\|v(N_{x,d}) - v(N_{x_0,d})\|_{2,\kappa}^2 = \frac{\sum_{y \in \mathcal{N}_{x_0,d}} \kappa(y) (v(T_{x_0,xy}) - v(y))^2}{\sum_{y \in \mathcal{N}_{x_0,d}} \kappa(y)},
\end{equation}

where \kappa(y) is some kernel defined on \mathcal{N}_{x_0,d}. The reason for introducing this smoothed version is that using it instead of the distance (12) allows to improve the denoising results significantly. With the rectangular kernel \kappa = \kappa_r defined by

\begin{equation}
\kappa_r(y) = \begin{cases} 1 & \text{if } y \in \mathcal{N}_{x_0,d}, \\ 0 & \text{otherwise,} \end{cases}
\end{equation}

the distance (24) coincides with (12). Other smoothing kernels usually used in simulations are the Gaussian kernel

\begin{equation}
\kappa_g(y) = e^{-\frac{N^2\|y-x_0\|^2}{2h_g^2}},
\end{equation}

where \( h_g \) is the bandwidth parameter, and the kernel \( \kappa_0 \) is defined as follows:

\begin{equation}
\kappa_0(y) = \sum_{k=\max(1,j)}^{(d-1)/2} \frac{1}{(2k+1)^2}.
\end{equation}

if \( \|y-x_0\|_{\infty} = \frac{j}{N} \) for some \( j \in \{0, 1, \cdots, (d-1)/2\} \), for \( y \in \mathcal{N}_{x_0,d} \).

Following the original algorithm by Buades et al. [5], in the simulations all over this paper we shall use the smoothed version (24) of the distance between the patches using the kernel \( \kappa(y) = \kappa_0(y) \).

To avoid the undesirable border effects, we extend the image outside the image limits symmetrically with respect to the border. At the corners, the image is extended symmetrically with respect to the corner pixels.

In our simulations we will use the following algorithm for the Non-Local Means filter (15) of Buades et al. [5].

---

**Algorithm** NL-means (Buades et al. [5])

Let \{H, D, d\} be the parameters.

Repeat for each \( x_0 \in I \) :

1. compute \( \|v(N_{x,d}) - v(N_{x_0,d})\|_{2,\kappa}^2 \) by (24) (with \( \kappa = \kappa_0 \) defined by (27)) for \( x \neq x_0 \), and set

\[ \|v(N_{x_0,d}) - v(N_{x_0,d})\|_{2,\kappa}^2 = \min \{\|v(N_{x,d}) - v(N_{x_0,d})\|_{2,\kappa}^2 : x \neq x_0, x \in \mathcal{N}_{x_0,D}\} \]
compute \( w(x) = \exp(-\|v(N_{x,d}) - v(N_{x_0,d})\|_2^2/(2H^2)) \subseteq \sum_{x' \in N_{x_0,D}} \exp(-\|v(N_{x',d}) - v(N_{x_0,d})\|_2^2/(2H^2)) \),
compute \( \hat{u}(x_0) = \sum_{x \in N_{x_0,D}} w(x)v(x) \).

In the algorithm, to avoid that the central pixel \( x_0 \) of the search window has a too important effect of the filtering, the value of its similarity norm is taken to be the minimum among the values of the similarity norms of all other pixels.

4. Choice of parameters and simulation results. In this section, based on simulation results, we first propose a simple expression for the choice of the parameters \( H, d, D \) in the Non-Local Means filter. Then we compare the performance of the Non-Local Means filter computed using the parameters proposed in this paper and those proposed in Buades et al. [5]. The results are measured by the usual

| Image  | Lena 512 × 512 | Barbara 512 × 512 | Boats 512 × 512 | House 256 × 256 | Peppers 256 × 256 |
|--------|----------------|-----------------|-----------------|-----------------|------------------|
| \( \sigma/\text{PSNR} \) | 10/28.12db | 10/28.12db | 10/28.12db | 10/28.14db | 10/28.14db |
| PSNR/Buades et al. [6] | 34.99db | 33.82db | 32.85db | 35.50db | 33.13db |
| PSNR/Ours | 35.22db | 33.55db | 33.00db | 35.35db | 33.16db |
| \( \Delta \text{PSNR} \) | 0.23db | -0.27db | 0.15db | -0.15db | 0.03db |

| Image  | Lena 512 × 512 | Barbara 512 × 512 | Boats 512 × 512 | House 256 × 256 | Peppers 256 × 256 |
|--------|----------------|-----------------|-----------------|-----------------|------------------|
| \( \sigma/\text{PSNR} \) | 20/22.11db | 20/22.11db | 20/22.11db | 20/22.11db | 20/22.12db |
| PSNR/Buades et al. [6] | 31.51db | 30.38db | 29.32db | 32.51db | 29.73db |
| PSNR/Ours | 32.41db | 30.62db | 30.02db | 32.57db | 30.30db |
| \( \Delta \text{PSNR} \) | 0.82db | 0.24db | 0.70db | 0.08db | 0.57db |

| Image  | Lena 512 × 512 | Barbara 512 × 512 | Boats 512 × 512 | House 256 × 256 | Peppers 256 × 256 |
|--------|----------------|-----------------|-----------------|-----------------|------------------|
| \( \sigma/\text{PSNR} \) | 30/18.60db | 30/18.60db | 30/18.60db | 30/18.61db | 30/18.61db |
| PSNR/Buades et al. [6] | 28.86db | 27.05db | 27.38db | 29.17db | 27.67db |
| PSNR/Ours | 30.20db | 28.06db | 28.60db | 30.49db | 28.28db |
| \( \Delta \text{PSNR} \) | 1.34db | 0.41db | 1.22db | 1.32db | 0.61db |

Table 3. Comparison of the Non-Local Means filter with our choice of parameters and other algorithms. By * we mark the algorithms for which the results were reported by their authors.
Figure 1. Approximation of $H = \sqrt{a_0 \sigma^2 + b_0}$ (red line) by $H = a\sigma + b$ (black line) with $a = 0.4$, $b = 2$, $a_0 = 0.2$, $b_0 = 10$.

Figure 2. The evolution of the PSNR as a function of the parameter $H$.

Figure 3. The evolution of the PSNR as a function of the size of similarity patches $d$. 
Figure 4. The restored image (left) and its square error (right) with different similarity patch sizes $d = 7, 9, 21, 41$ and the same search window size $D = 13$. The original image Lena was polluted by a Gaussian noise with $\sigma = 20$. 

(a) $D = 13$ and $d = 7$, PSNR = 30.8974

(b) $D = 13$ and $d = 9$, PSNR = 31.4142

(c) $D = 13$ and $d = 21$, PSNR = 32.4103

(d) $D = 13$ and $d = 41$, PSNR = 32.5274
Figure 5. The restored image (left) and its square error (right) with different similarity patch sizes $d = 7, 9, 21, 41$ and the same search window size $D = 13$. The original image Boat was polluted by a Gaussian noise with $\sigma = 20$. 
Figure 6. The restored image (left) and its square error (right) with different similarity patch sizes $d = 7, 9, 21, 41$ and the same search window size $D = 13$. The original image Peppers was polluted by a Gaussian noise with $\sigma = 20$. 
Peak Signal-to-Noise Ratio (PSNR) in decibels (db) defined as

$$PSNR = 10 \log_{10} \frac{255^2}{MSE},$$

with

$$MSE = \frac{1}{\text{card} I} \sum_{x \in I} (u(x_0) - \hat{u}(x))^2,$$

where $u$ is the original image and $\hat{u}$ the estimated one.

In our simulations we have used a set of images available at http://www.cs.tut.fi/~foi/GCF-BM3D/: “Lena”, “Barbara”, “Boat”, “House” and “Peppers”. We assume that the parameters $\{\sigma, H, d, D\}$ take values in the set $\mathcal{P} = A_1 \times A_2 \times A_3 \times A_4$, where

- $A_1 = \{5, 10, 15, \ldots, 50\}$,
- $A_2 = \{2, 4, 6, \ldots, 40\}$,
- $A_3 = \{3, 5, 7, \ldots, 41\}$,
- $A_4 = \{3, 5, 7, \ldots, 41\}$.

For each $\sigma \in A_1$, we maximize the PSNR as a function of $H, d$ and $D$, and we obtain simple formulas for the $H, d, D$ (in function of $\sigma$) where the maximum is attained. To do this, for each $\sigma$, we let $H, d$ and $D$ run over the set $A_2 \times A_3 \times A_4$ and we record the maximal values of the PSNR and the maximizers $H, d$ and $D$. In our simulations the standard deviation $\sigma$ is assumed to be known. It can be estimated if it is not known, for example, by the method proposed in [19].

In the following, we find expressions to approximate the maximizers $H, d$ and $D$ in function of $\sigma$ using the image “Lena”, but the validity of the expressions are confirmed by experiments on other images (cf. Tables 2 and 3).

First we propose to choose the parameter $H$ as a function of $\sigma$. The parameter $H$ determines the contribution of the similarity norm in the weights (16). As the similarity norm (12) is stable with respect to $d$ (due to the factor $d^2$ in its definition), we naturally assume that $H$ does not depend on $d$. Since the similarity norm (12) is also independent of $D$ we can also assume that $H$ is independent of $D$. On the contrary, $H$ depends on the noise deviation $\sigma$. In fact, by (13) the mean square value of the similarity norm $\mathbb{E}[\|v(\mathcal{N}_{x,d}) - v(\mathcal{N}_{x_0,d})\|_2^2]$ is about $2\sigma^2$ plus a quantity which does not depend on $\sigma$. Therefore, as a reasonable approximation, we can

![Figure 7](image-url)
Figure 8. The restored image (left) and its square error (right) with the same similarity patch size $d = 21$ and different search window sizes $D = 9, 13, 17, 21$. The original image Lena was polluted by a Gaussian noise with $\sigma = 20$. 

(a) $D = 9$ and $d = 21$, PSNR = 32.2814

(b) $D = 13$ and $d = 21$, PSNR = 32.4103

(c) $D = 17$ and $d = 21$, PSNR = 32.3557

(d) $D = 21$ and $d = 21$, PSNR = 32.2473
Figure 9. The restored image (left) and its square error (right) with the same similarity patch size $d = 21$ and different search window sizes $D = 9, 13, 17, 21$. The original image Boat was polluted by a Gaussian noise with $\sigma = 20$. 

(a) $D = 9$ and $d = 21$, PSNR = 29.2133

(b) $D = 13$ and $d = 21$, PSNR = 29.4280

(c) $D = 17$ and $d = 21$, PSNR = 29.5726

(d) $D = 21$ and $d = 21$, PSNR = 29.6052
Figure 10. The restored image (left) and its square error (right) with the same similarity patch size $d = 21$ and different search window sizes $D = 9, 13, 17, 21$. The original image Peppers was polluted by a Gaussian noise with $\sigma = 20$. 
assume that $H^2$ is an affine function of $\sigma^2$; $H^2 = a_0 \sigma^2 + b_0$, with some constants $a_0 > 0$ and $b_0 > 0$. It is easy to see that, for $\sigma$ in the interval $[5, 50]$, the linear relation $H = a \sigma + b$ with some constants $a, b > 0$ provides a good approximation for $H = \sqrt{a_0 \sigma^2 + b_0}$, see Fig. 1 where we display both functions with $a = 0.4, b = 2$ and $a_0 = 0.2, b_0 = 10$.

When $\sigma = 10$ the maximal PSNR is attained for $H = 6, d = 17, D = 11$. In Fig. 2 (a) we display the values of PSNR as a function of $H$ for $d = 17, D = 11$. We do the same for $\sigma = 20$ and $\sigma = 30$: when $\sigma = 20$ (resp. $\sigma = 30$) the maximal PSNR is attained for $H = 10, d = 21, D = 13$ (resp. $H = 14, d = 21, D = 15$); in Fig. 2 (b) (resp. (c)) we display the values of PSNR as a function of $H$ for $d = 21, D = 13$ (resp. $d = 21, D = 15$). Actually we have $H = 6, 10, 14$ for $\sigma = 10, 20, 30$, respectively, while the other parameters are maximizers. These numerical results show that the values $H = H(\sigma)$ that realize the maximum PSNR for $\sigma = 10, 20, 30$ are related by the formula

$$H = 0.4\sigma + 2.$$  

It turns out that the proposed formula is also valid for other values of $\sigma$, in the sense that with (28) we still obtain the optimal or near optimal PSNR.

Buades et al. [5] suggested the empirical choice $H = \frac{c\sigma}{\sqrt{d}}$ with $c \in [10, 15]$, for $d = 7, 9$ (actually they proposed $h = c\sigma$ with $c \in [10, 15]$; with their notation we have the correspondence $h^2 = 2d^2 H^2$). Note that (28) has the advantage that the value of $H$ is stable with respect to $d$.

Next we deal with the choice of the size of similarity patches $d$. In Fig. 3 we display the PSNR in function of $d$ with different levels of the Gaussian noise $\sigma = 10, 20, 30$, with search window size $D = 11, 13, 15$, respectively, and $H$ given by (28). These values of $H$ and $D$ are maximizers of the PSNR. For $\sigma = 10$, Fig. 3 (a) shows that the PSNR increases with $d$ when $d \leq 15$ and decreases when $d \geq 20$, so that it reaches the peak value for $d$ about 17. For $\sigma = 20$ and $\sigma = 30$, graphs (b) and (c) in Fig. 3 illustrate that the PSNR increases with $d$. We note that the growth of the PSNR is not significant for $d \geq 21$. Since a larger width of similarity patches increases the amount of computation, we take the similarity patch size $d = 21$ for $\sigma = 20$ and $\sigma = 30$.

Fig. 4 displays the denoised image Lena and its square error image (the square of the difference between the denoised image and the original one) when the image is contaminated by Gaussian noise with $\sigma = 20$, which verifies the conclusions we have just come to. Fig. 4(c) with $d = 21$ is not only significantly better than Fig. 4(a) with $d = 7$ and Fig. 4(b) with $d = 9$ in image visual quality, but also has almost the same image visual quality as Fig. 4(d) with $d = 41$. Figs. 5 and 6 show the same performance as in Fig. 4 with the images Boat and Peppers. The above mentioned analysis leads us to the choice

$$d = \begin{cases} 17, & \text{if } \sigma < 15, \\ 21, & \text{if } \sigma \geq 15. \end{cases}$$  

Now we turn to the choice of the size of the search window $D$. We choose the parameters $H$ and $d$ by (28) and (29) as functions of $\sigma$. In Fig. 7 we show the PSNR in function of $D$ for $\sigma = 10, 20, 30$. We note that the PSNR a concave function not increasing in $D$. In term of image visual effect, take Lena with $\sigma = 20$ as an example (see Fig. 8). Fig. 8(b) with $D = 13$ has the best image visual quality compared with those with $D = 9, 17$ and 21, but the change is not significant. Figs. 9 and 10 confirm this for the images Boat and Peppers. This is consistent with Theorem
Figure 11. Denoising results with the “Lena” 512 × 512 image.
**Figure 12.** Denoising results with the “Boat” $512 \times 512$ image.
Figure 13. Denoising results with the “Pepper” 256 × 256 image.
which ensures that there is an optimal choice of $D$ which gives the optimal rate of convergence. Since the maximal value of PSNR is attained when $D$ is located in the interval $[9, 21]$, we display only the PSNR corresponding to these values. From the graphs in Fig. 7 we conclude that the PSNR is close to the maximum when

$$D = 2 \times \lfloor 0.75 \times \sqrt{\sigma} + 2.55 \rfloor + 1,$$

where $\lfloor x \rfloor$ denotes the least integer exceeding $x$. For instance we have $D = 11$ when $\sigma = 10$, $D = 13$ when $\sigma = 20$, and $D = 15$ when $\sigma = 30$; these values of $D$ fit well the points of maximums in Fig. 7. The empirical choice (30) is confirmed by further simulation results (not presented here) with other values of $\sigma$. We mention that if we replace $\lfloor x \rfloor$ by $\lceil x \rceil$ (the integer part of $x$) in (30) the restoration results are very close.

From Figs. 3-10 we also draw the conclusion that the effect of parameter $d$ on denoising is significantly greater than that of parameter $D$. Furthermore, a larger $D$ does not give necessarily better denoising results, while a larger $d$ improves always the denoising quality.

The performance of the Non-Local Means filter with the choice of parameters (28), (29) and (30) is illustrated in Figs. 11-13, using the 512 $\times$ 512 images Lena and Boat and the 256 $\times$ 256 image Peppers corrupted by an additive Gaussian white noise with $\sigma = 20$. The original image and the noisy one (with PSNR= 22.10db) are given in Fig. 11(a) and (b). The denoised image (with PSNR= 32.41db) using our choice of parameters is given in Fig. 11(c), where we can see that the noise is reduced in a natural manner and significant geometric features, fine textures, and original contrasts are visually well recovered with no undesirable artifacts.

To better appreciate the accuracy of the restoration with the choose of parameters (28), (29) and (30), in Figs. 11-13, we show the denoised image and its square error image. The dark values correspond to high-confidence estimates. As expected, pixels with a low level of confidence are located in the neighborhood of image discontinuities.

For comparison, in Fig. 11(d) and (f), we give the image denoised by the Non-Local Means filter with parameters $D = 21$ and $d = 9$ as proposed in Buades et al. [5], as well as its square error image. We note that with our choice of parameters ($D = 13$ and $d = 21$ for $\sigma = 20$) the PSNR is 32.41db, which is better than that obtained with the parameters proposed in Buades et al. [5], PSNR= 31.51db. We can see clearly that the overall visual impression and the numerical results are improved using the new choice of the parameters. Figs. 12 and 13 bring us to the same conclusion.

The fact that the formulas (28), (29), (30) are valid also for the images Lena, Barbara, Boat, House and Peppers is confirmed by the PSNR values presented in Tables 2 and 3.

In Table 2, we show a comparison of the PSNR values of the Non-Local Means filter computed with parameters proposed in Buades et al. [5] and with those proposed in our paper. The difference in the visual quality rises noticeably as the standard deviation $\sigma$ increases. There is no significant improvement for $\sigma = 10$; the average PSNR gain is 0.50db for $\sigma = 20$ and 0.98db for $\sigma = 30$.

The comparison with several other filters is given in Table 3. The PSNR values show that the Non-Local Means filter with proper parameters is as good as more sophisticated methods, like [17, 27, 16, 1], and better than the filters proposed in [38, 24, 15, 37, 17, 2, 41]. The proposed approach gives a denoising quality which is competitive with that of the state of the art methods [9, 29].
5. Proofs of the main results.

5.1. Proof of Theorem 2.1. By the definition (6) of $u^*(x_0)$ we have

$$u^*(x_0) - u(x_0) = \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x) (v(x) - u(x_0))$$

(31)

$$= \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x) (u(x) - u(x_0) + \varepsilon(x)).$$

The usual bias-variance decomposition of the mean squared error gives

$$\mathbb{E}(u^*(x_0) - u(x_0))^2 = \left( \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x) (u(x) - u(x_0)) \right)^2 + \sigma^2 \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x)^2$$

(32)

$$\leq \left( \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x) |u(x) - u(x_0)| \right)^2 + \sigma^2 \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x)^2.$$

The inequality (32) combined with (5) implies the following upper bound:

$$\mathbb{E}(u^*(x_0) - u(x_0))^2 \leq g(w^*),$$

(33)

where

$$g(w^*) = \left( \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x) \rho_{x_0}(x) \right)^2 + \sigma^2 \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x)^2.$$

The assertion of Theorem 2.1 follows from (46) and the following lemma.

Lemma 5.1. We have

$$g(w^*) \leq \frac{81}{2} L^2 \Delta^{2\beta} = cn^{-\frac{\beta}{\pi r}}$$

(35)

with

$$c = \frac{9^{\frac{\beta}{\pi r}} \sigma^{\frac{2\beta}{\pi r}} L}{2^{\frac{2\beta}{\pi r}} \beta^{\frac{2\beta}{\pi r}}}.$$

Proof. Note that $x \in \mathcal{N}_{x_0,D}$ is equivalent to $\|x - x_0\|_\infty \leq \Delta$. Denoting for brevity

$$I_1 = \left( \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x) \rho_{x_0}(x) \right)^2 = \left( \sum_{\|x - x_0\|_\infty \leq \Delta} e^{-\frac{\rho_{x_0}(x)}{2\kappa}} \rho_{x_0}(x) \right)^2,$$

(36)

and

$$I_2 = \sigma^2 \sum_{x \in \mathcal{N}_{x_0,D}} (w^*(x))^2 = \sigma^2 \left( \sum_{\|x - x_0\|_\infty \leq \Delta} e^{-\frac{\rho_{x_0}(x)}{2\kappa}} \right)^2,$$

(37)

we have

$$g(w^*) = I_1 + I_2.$$
Note that $te^{-t^2/2H^2}$, $t \in [0, H)$ is increasing, and $D = 2N\Delta + 1 \leq 3N\Delta$, we have

\[\sum_{\|x-x_0\|_{\infty} \leq \Delta} e^{-\frac{L^2|x-x_0|_{\infty}^{2\beta}}{2H^2}} L\|x-x_0\|_{\infty}^{\beta} \leq \sum_{\|x-x_0\|_{\infty} \leq \Delta} L\Delta^{\beta} e^{-\frac{L^2\Delta^{2\beta}}{2H^2}} \leq \sum_{\|x-x_0\|_{\infty} \leq \Delta} L\Delta^{\beta} \leq 9L\Delta^{\beta+2}n.\]  

(39)

Since $e^{-t^2/2H^2}$, is decreasing in $t \geq 0$, $e^{-x} \geq 1 - x$ for $x \geq 0$, and $L\Delta^{\beta}/H < 1$, we get,

\[\sum_{\|x-x_0\|_{\infty} \leq \Delta} \|x-x_0\|_{\infty} \leq \Delta e^{-\frac{L^2\Delta^{2\beta}}{2H^2}} \leq \sum_{\|x-x_0\|_{\infty} \leq \Delta} \|x-x_0\|_{\infty} \leq \Delta e^{-\frac{2\sigma^2}{\Delta^2n}} \leq \sum_{\|x-x_0\|_{\infty} \leq \Delta} \frac{1}{2} = \frac{D^2}{2} \geq 2\Delta^2n.\]  

(40)

The three inequalities (36), (39) and (40) imply that

\[I_1 \leq \frac{81}{4} L^2 \Delta^{2\beta}.\]  

(41)

Taking into account the inequality

\[\sum_{\|x-x_0\|_{\infty} \leq \Delta} e^{-\frac{L^2|x-x_0|_{\infty}^{2\beta}}{2H^2}} \leq \sum_{\|x-x_0\|_{\infty} \leq \Delta} 1 = D^2 \leq 9\Delta^2n,\]  

together with (37) and (40), we see that

\[I_2 \leq \frac{9}{4} \frac{\sigma^2}{\Delta^2n}.\]  

(42)

Combining (38), (41) and (42), we get

\[g(w^*) \leq \frac{81}{4} L^2 \Delta^{2\beta} + \frac{9}{4} \frac{\sigma^2}{\Delta^2n}.\]  

(43)

Let $\Delta$ minimize the right hand side of the above inequality. Then

\[18\beta L^2 \Delta^{2\beta-1} - \frac{2\sigma^2}{\Delta^3n} = 0,\]  

from which we infer that

\[\Delta = \left(\frac{\sigma^2}{9\beta L^2}\right)^{\frac{1}{\beta+2}} n^{-\frac{1}{\beta+2}}.\]  

(44)

Substituting (44) to (43) leads to

\[g(w^*) \leq \frac{81}{2} L^2 \Delta^{2\beta} = cn^{-\frac{\beta}{\beta+2}} \quad \text{with} \quad c = \frac{9^{\frac{\beta+1}{\beta+2}} \sigma^{\frac{2\beta}{\beta+2}} L^{\frac{2}{\beta+2}}}{2^{\frac{\beta}{\beta+2}}},\]  

(45)

Which finishes the proof of the lemma.
5.2. Proof of Theorem 3.1. The inequality (32) with \( w^* \) replaced by \( \tilde{w}^* \) implies the following upper bound:

\[
\mathbb{E} (\tilde{w}^*(x_0) - u(x_0))^2 \leq g(\tilde{w}^*),
\]

where

\[
g(\tilde{w}^*) = \left( \sum_{x \in \mathcal{N}_{x_0,D}} \tilde{w}^*(x)\tilde{\rho}_{x_0}(x) \right)^2 + \sigma^2 \sum_{x \in \mathcal{N}_{x_0,D}} \tilde{w}^*(x)^2.
\]

From (17) we have \( \tilde{\rho}_{x_0}(x) \leq \rho_{x_0}(x) + |\varepsilon_n| \). Using this and the fact that \((a+b)^2 \leq 2a^2 + 2b^2\), we get

\[
\left( \sum_{x \in \mathcal{N}_{x_0,D}} \tilde{w}^*(x)\tilde{\rho}_{x_0}(x) \right)^2 \leq \left( \sum_{x \in \mathcal{N}_{x_0,D}} \tilde{w}^*(x)(\rho_{x_0}(x) + |\varepsilon_n|) \right)^2
\]

\[
\leq 2 \left( \sum_{x \in \mathcal{N}_{x_0,D}} \tilde{w}^*(x)\rho_{x_0}(x) \right)^2 + 2|\varepsilon_n|.
\]

Since \( \tilde{\rho}_{x_0}^2(x) \leq \rho_{x_0}^2(x) + |\varepsilon_n| \) and \( \tilde{\rho}_{x_0}^2(x) \geq \rho_{x_0}^2(x) - |\varepsilon_n| \), we have

\[
\tilde{w}^*(x) = \frac{e^{-\frac{\rho_{x_0}^2(x)}{2H^2}}}{\sum_{y \in \mathcal{N}_{x_0,D}} e^{-\frac{\rho_{x_0}^2(y)}{2H^2}}} \leq \frac{e^{-\frac{\rho_{x_0}^2(x) - |\varepsilon_n|}{2H^2}}}{\sum_{y \in \mathcal{N}_{x_0,D}} e^{-\frac{\rho_{x_0}^2(y) + |\varepsilon_n|}{2H^2}}} = e^{|\varepsilon_n|/H^2} w^*(x).
\]

Therefore

\[
\left( \sum_{x \in \mathcal{N}_{x_0,D}} \tilde{w}^*(x)\rho_{x_0}(x) \right)^2 \leq 2e^{2|\varepsilon_n|/H^2} \left( \sum_{x \in \mathcal{N}_{x_0,D}} w^*(x)\rho_{x_0}(x) \right)^2 + 2|\varepsilon_n|
\]

Consequently

\[
g(\tilde{w}^*) \leq 2e^{2|\varepsilon_n|/H^2} g(w^*) + 2|\varepsilon_n|.
\]

Therefore, by Lemma 5.1 (inequality (35)) we obtain

\[
\mathbb{E} (\tilde{w}^*(x_0) - u(x_0))^2 \leq 2c e^{2|\varepsilon_n|/H^2} n^{-\frac{d}{\beta+2}} + 2|\varepsilon_n|, \quad \text{with} \quad c = \frac{9^{d+2} 2^{2d+2} L_{\varepsilon_n}^2}{2 \beta \pi^{d+2}},
\]

which is just (20).

5.3. Proof of Theorem 3.2. For convenience, let

\[
\Lambda_{x_0, x}(y) = u(y) - u(T_{x_0, x}y)
\]

and

\[
\zeta(y) = \varepsilon(y) - \varepsilon(T_{x_0, x}y).
\]

Denote \( m = d^2 \). With these notations we get, using (1),

\[
\tilde{\rho}_{x_0}^2(x) = \| v(N_{x,d}) - v(N_{x_0,d}) \|^2 - 2\sigma^2
\]

\[
= \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} (v(y) - v(T_{x_0,x}y))^2 - 2\sigma^2
\]
\( \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} (u(y) - u(T_{x_0,x}y) + \varepsilon(y) - \varepsilon(T_{x_0,x}y))^2 \)

\( = \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} (\Lambda_{x_0,x}(y) + \zeta(y))^2 - 2\sigma^2 \)

\( = \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} \Lambda_{x_0,x}^2(y) + \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} \left( \zeta(y)^2 - 2\sigma^2 + 2\Lambda_{x_0,x}(y)\zeta(y) \right) \)

(51) \( = \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} \Lambda_{x_0,x}^2(y) + \frac{1}{m} S(x_0, x) \),

where

\( S(x_0, x) = \sum_{y \in \mathcal{N}_{x_0,d}} \left( \zeta(y)^2 - 2\sigma^2 + 2\Lambda_{x_0,x}(y)\zeta(y) \right) \).

We will need two lemmas for the estimation of the two sums in (51).

Lemma 5.2. Under the local Hölder condition (9), with \( \Delta \) and \( \delta \) defined by (10) and (22), we have

\[ \left| \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} \Lambda_{x_0,x}^2(y) - \rho_{x_0}^2(x) \right| \leq 4L^2\Delta^2\delta^2. \]

Proof. Clearly

\( \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} \Lambda_{x_0,x}^2(y) - \rho_{x_0}^2(x) = \frac{1}{m} \sum_{y \in \mathcal{N}_{x_0,d}} (\Lambda_{x_0,x}^2(y) - \rho_{x_0}^2(x)) \).

For each \( y \in \mathcal{N}_{x_0,d} \), we have

\( \Lambda_{x_0,x}^2(y) - \rho_{x_0}^2(x) = (\Lambda_{x_0,x}^2(y) - \rho_{x_0}^2(x)) \).

\( = (\|y - T_{x_0,x}(y)\|_\infty - \|x - \rho_{x_0}^2(x)\|_\infty) \leq \Delta \) and \( \|x - T_{x_0,x}(y)\|_\infty = \|y - x_0\|_\infty \leq \delta \),

from the local Hölder condition (9) we have

\( \|y - u(x_0,y) + u(x_0) - u(x)\|_\infty \leq 2L\|x - x_0\|_\infty \leq 2L\Delta^2 \)

and

\( \|u(y) - u(x_0) + u(x) - u(T_{x_0,x}(y))\|_\infty \leq 2L\|y - x_0\|_\infty \leq 2L\delta^2 \).

From (53)-(56) we obtain the conclusion of the lemma.

Lemma 5.3. There are two positive constants \( c_1 \) and \( c_2 \), depending only on \( L \) and \( \sigma \), such that for any \( 0 < z \leq c_1m^{1/2} \),

\[ P \left( |S(x_0, x)| \geq z\sqrt{m} \right) \leq c_2z^{-2}. \]

Proof. Note that the variables

\( X_y = \zeta(y)^2 - 2\sigma^2 + 2\Lambda_{x_0,x}(y)\zeta(y), \quad y \in \mathcal{N}_{x_0,d} \)

are identically distributed with \( E X_y = 0 \) and finite variance \( b^2 = E X_y^2 \). The point in handling the sum \( S(x_0, x) = \sum_{y \in \mathcal{N}_{x_0,d}} X_y \) is that the variables \( X_y, y \in \mathcal{N}_{x_0,d} \) are
not necessarily independent. Remark that $\zeta(y)$ and $\zeta(y')$ are correlated if and only if $y - y' = \pm (x_0 - x)$: indeed, it can be easily checked that

$$E(\zeta(y) \zeta(y')) = \begin{cases} -\sigma^2, & \text{if } y - y' = x_0 - x, \\ \sigma^2, & \text{if } y - y' = x - x_0, \\ 0, & \text{otherwise.} \end{cases}$$

Since $\zeta_{x_0,x}(y)$ and $\zeta_{x_0,x}(y')$ are Gaussian random variables, this implies that if $y - y' \neq \pm (x - x_0)$, then $\zeta_{x_0,x}(y)$ and $\zeta_{x_0,x}(y')$ are independent, so that $X_y$ and $X_{y'}$ are independent if $y - y' \neq \pm (x - x_0)$. Consequently

$$\operatorname{Var}(S(x_0, x)) = E(S(x_0, x)^2) = \sum_{y,y' \in N_{x_0,d}} E(X_y X_{y'})$$

By the Cauchy-Schwartz inequality $E(X_y X_{y'}) \leq b^2$. Hence

$$\operatorname{Var}(S(x_0, x)) \leq mb^2 + 2mb^2 = 3mb^2.$$ Therefore, by Chebyshev’s inequality

$$P\left(\frac{1}{m} |S(x_0, x)| \geq c \sqrt{\ln n} \right) \leq \frac{\operatorname{Var}(S(x_0, x))}{c^2 m} \leq \frac{3b^2}{c^2}.$$

Now we turn to the proof of Theorem 3.2. By equation (22) and the condition of $\delta$ in the Theorem 3.2, we have $m = d^2 \geq 4c_0^2 n^{1-2\alpha}$. Clearly, $c_1 m^{1/2} > \sqrt{\frac{1}{c_5} \ln n}$ for some constant $c_3 > 0$ independent of $n$. Applying Lemma 5.3 with $z = \sqrt{\frac{1}{c_5} \ln n}$, we see that

$$P\left(\frac{1}{m} |S(x_0, x)| \geq c \sqrt{\ln n} \right) \leq \frac{c_3}{\ln n}.$$ Taking $m = d^2 \geq 4c_0^2 n^{1-2\alpha}$, we obtain with $c_5 = 1/(2 \sqrt{c_0 c_3})$,

$$P\left(\frac{1}{m} |S(x_0, x)| \geq c_5 n^{\alpha - \frac{\beta}{2}} \sqrt{\ln n} \right) \leq \frac{c_3}{\ln n}.$$ By Lemma 5.2 and the conditions on $\Delta$ and $\delta$ we have, with $c_6 = c_0^\beta \left( \frac{\sigma^2}{\pi b L_7} \right)^{\frac{1}{2\beta}}$,

$$\left| \frac{1}{m} \sum_{y \in N_{x_0,d}} \lambda^2_{x_0,x}(y) - \rho^2_{x_0}(x) \right| \leq 4L^2 \Delta^2 \delta^3 \leq c_6 n^{-\frac{\beta}{2\beta + 2} - \alpha \beta}.$$ From (51) and (63) we obtain

$$|\hat{\rho}_{x_0}(x) - \rho_{x_0}(x)| \leq c_6 n^{-\frac{\beta}{2\beta + 2} - \alpha \beta} + \frac{1}{m} |S(x)|.$$ Combining (62) and (64), we get

$$P\left(\left| \hat{\rho}_{x_0}(x) - \rho_{x_0}(x) \right| \geq c_5 n^{\alpha - \frac{\beta}{2}} \sqrt{\ln n} + c_6 n^{-\frac{\beta}{2\beta + 2} - \alpha \beta} \right) \leq \frac{c_3}{\ln n}.$$ Therefore, since the condition $\frac{1}{2(\beta + 1)^2} < \alpha < \frac{1}{2}$ implies $\frac{\beta}{2\beta + 2} + \alpha \beta > \frac{1}{2} - \alpha > 0$, we obtain the inequality (23) with $c = c_5 + c_6$. 

Inverse Problems and Imaging Volume 12, No. 4 (2018), 853–881
6. Conclusion. We derived the Non-Local Means filter by introducing an oracle neighborhood filter and by giving a statistical estimator of the similarity function in the oracle filter. By studying the convergence of the oracle filter while the similarity function is perturbed with a suitable error term and the convergence of the statistical estimator of the similarity function, we showed that the Non-Local Means filter converges at an optimal rate under a proper choice of the search window size $D$ and the similarity patch size $d$. We also proposed simple formulas for the choice of the parameters $H, d, D$ as functions of $\sigma$, where $H$ is the bandwidth of the Gaussian kernel; in particular we found that choosing $d > D$ may improve restoration results, as suggested by the convergence theorems. Simulation results show that with the new proposed choice of the parameters, we improved the performance of the standard Non-Local Means filter of Buades, Coll and Morel [5] both numerically and visually.

Acknowledgments. We would like to thank Professor Jean-Michel Morel and the reviewer for very helpful comments, suggestions and remarks, which lead to a significant improvement of the initial manuscript. The work has benefited from the support of the Centre Henri Lebesgue (CHL, ANR-11-LABX-0020-01). Jin and Liu have been supported by the National Natural Science Foundation of China (Grant No. 61661039, No.61661040, No. 61661038, No. 11401590, No. 11571052 and No. 11731012), Natural Science Fund of Inner Mongolia Autonomous Region (Grant No. 2016MS0107) and Scientific Research Projection of Higher Schools of Inner Mongolia (Grant No. NJZY16017), and Hunan Provincial Natural Science Foundation of China (Grant No.2017JJ2271).

REFERENCES

[1] M. Aharon, M. Elad and A. Bruckstein, rmk-svd: An algorithm for designing overcomplete dictionaries for sparse representation, IEEE Trans. Signal Process., 54 (2006), 431–4322.
[2] E. Arias-Castro, J. Salmon and R. Willett, Oracle inequalities and minimax rates for non-local means and related adaptive kernel-based methods, Siam Journal on Imaging Sciences, 5 (2012), 944–992.
[3] R.C. Bilcu and M. Vehvilainen, Fast nonlocal means for image denoising, In Proc. of SPIE Conf. on Digital Photography III, 6502 (2007), 65020R.
[4] J. Boulanger, C. Kervrann, P. Bouthemy, P. Elbau, J. B. Sibarita and J. Salamero, Patch-based nonlocal functional for denoising fluorescence microscopy image sequences, IEEE Transactions on Medical Imaging, 29 (2010), 442–454.
[5] A. Buades, B. Coll and J. M. Morel, A review of image denoising algorithms, with a new one, Multiscale Model. Simul., 4 (2005), 490–530.
[6] A. Buades, B. Coll and J. M. Morel, The staircasing effect in neighborhood filters and its solution, IEEE Trans. Image Process., 15 (2006), 1499–1505.
[7] T. Buades, Y. Lou, J. M. Morel and Z. Tang, A note on multi-image denoising, In Int. workshop on Local and Non-Local Approximation in Image Processing, pages 1–15, August 2009.
[8] P. Chatterjee and P. Milanfar, A generalization of non-local means via kernel regression, In Proc. of SPIE Conf. on Computational Imaging, Citeseer, 6814 (2008), 6814Op.
[9] K. Dabov, A. Foi, V. Katkovnik and K. Egiazarian, Image denoising by sparse 3-D transform-domain collaborative filtering, IEEE Trans. Image Process., 16 (2007), 2080–2095.
[10] K. Dabov, A. Foi, V. Katkovnik and K. Egiazarian, Bm3d image denoising with shape-adaptive principal component analysis, In Proc. Workshop on Signal Processing with Adaptive Sparse Structured Representations (SPARS 09), volume 49, Citeseer, 2009.
[11] C. A. Deledalle, V. Duval and J. Salmon, Non-local methods with shape-adaptive patches (nlm-sap), Journal of Mathematical Imaging and Vision, 43 (2012), 103–120.
[12] D. L. Donoho and J. M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, Biometrika, 81 (1994), 425–455.
[13] V. Duval, J.-F. Aujol and Y. Gousseau, A bias-variance approach for the nonlocal means, *SIAM Journal on Imaging Sciences*, 4 (2011), 760–788.

[14] J. Q. Fan and I. Gijbels, *Local Polynomial Modelling and Its Applications*, Chapman & Hall, London, 1996.

[15] A. Foi, V. Katkovnik, K. Egiazarian and J. Astola, A novel anisotropic local polynomial estimator based on directional multiscale optimizations, In *Proc. 6th IMA Int. Conf. Math. in Signal Process*, pages 79–82, Citeseer.

[16] D. K. Hammond and E. P. Simoncelli, *Image modeling and denoising with orientation-adapted gaussian scale mixtures*, IEEE Trans. Image Process., 17 (2008), 2089–2101.

[17] K. Hirakawa and T. W. Parks, *Image denoising using total least squares*, IEEE Trans. Image Process., 15 (2006), 2730–2742.

[18] H. Hu, B. Li and Q. Liu, Removing mixture of gaussian and impulse noise by patch-based weighted means, *Journal of Scientific Computing*, 67 (2016), 103–129.

[19] J. Immerkaer, *Fast noise variance estimation*, Computer vision and image understanding, 64 (1996), 300–302.

[20] Q. Jin, I. Grama, C. Kervrann and Q. Liu, *Nonlocal means and optimal weights for noise removal*, SIAM Journal on Imaging Sciences, 10 (2017), 1878–1920.

[21] Q. Jin, I. Grama and Q. Liu, Removing gaussian noise by optimization of weights in non-local means, preprint, arXiv:1109.5640.

[22] Q. Jin, I. Grama and Q. Liu, *A new poisson noise filter based on weights optimization*, Journal of Scientific Computing, 58 (2014), 548–573.

[23] V. Karnati, M. Uliyar and S. Dey, *Fast non-local algorithm for image denoising*, In *IEEE International Conference on Image Processing (ICIP)*, 2009, 6th, pages 3873–3876, IEEE, 2009.

[24] V. Katkovnik, A. Foi, K. Egiazarian and J. Astola, Directional varying scale approximations for anisotropic signal processing, In *Proc. XII European Signal Proc. Conf., EUSIPCO 2004*, Vienna, pages 101–104, 2004.

[25] V. Katkovnik, A. Foi, K. Egiazarian and J. Astola, From local kernel to nonlocal multiple-model image denoising, *Int. J. Comput. Vis.*, 86 (2010), 1–32.

[26] C. Kervrann and J. Boulanger, *Optimal spatial adaptation for patch-based image denoising*, IEEE Trans. Image Process., 15 (2006), 2866–2878.

[27] C. Kervrann and J. Boulanger, *Local adaptivity to variable smoothness for exemplar-based image regularization and representation*, Int. J. Comput. Vis., 79 (2008), 45–60.

[28] M. Lebrun, A. Buades and J. M. Morel, Implementation of the “Non-Local Bayes” (NL-bayes) image denoising algorithm, *Image Processing On Line*, 2013 (2013), 1–42.

[29] M. Lebrun, A. Buades and J. M. Morel, *A nonlocal bayesian image denoising algorithm*, SIAM Journal on Imaging Sciences, 6 (2013), 1665–1688.

[30] A. Levin and B. Nadler, *Natural image denoising: Optimality and inherent bounds*, In *Computer Vision and Pattern Recognition*, pages 2833–2840, 2011.

[31] B. Li, Q. Liu, J. Xu and X. Luo, *A new method for removing mixed noises*, Science China Information Sciences, 54 (2011), 51–59.

[32] Y. Lou, X. Zhang, S. Osher and A. Bertozzi, *Image recovery via nonlocal operators*, J. Sci. Comput., 42 (2010), 185–197.

[33] M. Mahmoudi and G. Sapiro, *Fast image and video denoising via nonlocal means of similar neighborhoods*, IEEE Signal. Proc. Let., 12 (2005), 839–842.

[34] A. Maleki, M. Narayan and R. Baraniuk, *Suboptimality of Nonlocal Means on Images with Sharp Edges*, In Annual Allerton Conference on Communication, Control, and Computing, 2011.

[35] A. Maleki, M. Narayan and R. Baraniuk, *Anisotropic nonlocal means denoising*, Applied and Computational Harmonic Analysis, 35 (2013), 452–482.

[36] J. Polzehl and V. Spokoiny, *Propagation-separation approach for local likelihood estimation*, Probab. Theory Rel. Fields, 135 (2006), 335–362.

[37] S. Roth and M. J. Black, *Fields of experts*, Int. J. Comput. Vision, 82 (2009), 205–229.

[38] J. Salmon and E. Le Pennec, NL-means and aggregation procedures, In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2977–2980, IEEE, 2009.

[39] N. A. Thacker, P. A. Bromiley and J. V. Manjonb, *A quantitative theory of the non-local means algorithm*, In *Proc. MIUA 2008*, Dundee, Scotland, pages 174–178. Citeseer, 2008.

[40] C. Tomasi and R. Manduchi, *Bilateral filtering for gray and color images*, In *Proc. Int. Conf. Computer Vision*, pages 839–846, 1998.
Convergence of the Non-Local Means filter

[41] D. Van De Ville and M. Kocher, Non-local means with dimensionality reduction and sure-based parameter selection, IEEE Trans. Image Process., 20 (2010), 2683–2690.

[42] R. Vignesh, B. T. Oh and C. C. J. Kuo, Fast non-local means (nlm) computation with probabilistic early termination, IEEE Signal. Proc. Let., 17 (2010), 277–280.

[43] Y. Q. Wang, The implementation of sure guided piecewise linear image denoising, Image Processing On Line, 2013 (2013), 43–67.

[44] L. P. Yaroslavsky, Digital Picture Processing. An Introduction, In Springer-Verlag, Berlin, 1985.

Received March 2017; 1st revision November 2017, final revision December 2017.
E-mail address: qyjin2015@aliyun.com
E-mail address: ion.grama@univ-ubs.fr
E-mail address: quansheng.liu@univ-ubs.fr