Computer Science and Artificial Intelligence Laboratory

Technical Report

MIT-CSAIL-TR-2012-022 October 7, 2012

Patch complexity, finite pixel correlations and optimal denoising

Anat Levin, Boaz Nadler, Fredo Durand, and William T. Freeman

massachusetts institute of technology, cambridge, ma 02139 usa — www.csail.mit.edu

Patch Complexity, Finite Pixel Correlations and Optimal Denoising

Anat Levin[1] Boaz Nadler[1] Fredo Durand[2] William T. Freeman[2]

[1]Weizmann Institute [2]MIT CSAIL

Abstract. Image restoration tasks are ill-posed problems, typically solved with priors. Since the optimal prior is the exact unknown density of natural images, actual priors are only approximate and typically restricted to small patches. This raises several questions: How much may we hope to improve current restoration results with future sophisticated algorithms? And more fundamentally, even with perfect knowledge of natural image statistics, what is the inherent ambiguity of the problem? In addition, since most current methods are limited to finite support patches or kernels, what is the relation between the patch complexity of natural images, patch size, and restoration errors? Focusing on image denoising, we make several contributions. First, in light of computational constraints, we study the relation between denoising gain and sample size requirements in a non parametric approach. We present a law of diminishing return, namely that with increasing patch size, rare patches not only require a much larger dataset, but also gain little from it. This result suggests novel adaptive variable-sized patch schemes for denoising. Second, we study absolute denoising limits, regardless of the algorithm used, and the converge rate to them as a function of patch size. Scale invariance of natural images plays a key role here and implies both a strictly positive lower bound on denoising and a power law convergence. Extrapolating this parametric law gives a ballpark estimate of the best achievable denoising, suggesting that some improvement, although modest, is still possible.

1 Introduction

Characterizing the properties of natural images is critical for computer and human vision [18, 13, 20, 16, 7, 23]. In particular, low level vision tasks such as denoising, super resolution, deblurring and completion, are fundamentally ill-posed since an infinite number of images x can explain an observed degraded image y. Image priors are crucial in reducing this ambiguity, as even approximate knowledge of the probability $p(x)$ of natural images can rule out unlikely solutions.

This raises several fundamental questions. First, at the most basic level, what is the inherent ambiguity of low level image restoration problems? i.e., can they be solved with zero error given perfect knowledge of the density $p(x)$? More practically, how much can we hope to improve current restoration results with future advances in algorithms and image priors?

Clearly, more accurate priors improve restoration results. However, while most image priors (parametric, non-parametric, learning-based) [2, 14, 20, 16, 23] as well as studies

on image statistics [13, 7] are restricted to local image patches or kernels, little is known about their dependence on patch size. Hence another question of practical importance is the following: What is the potential restoration gain from an increase in patch size? and, how is it related to the "patch complexity" of natural images, namely their geometry, density and internal correlations.

In this paper we study these questions in the context of the simplest restoration task: image denoising [18, 20, 6, 11, 9, 15, 10, 23]. We build on prior attempts to study the limits of natural image denoising [17, 3, 8]. In particular, on the non-parametric approach of [14], which estimated the optimal error for the class of patch based algorithms that denoise each pixel using only a finite support of noisy pixels around it. A major limitation of [14], is that computational constraints restricted it to relatively small patches. Thus, [14] was unable to predict the best achievable denoising of algorithms that are allowed to utilize the entire image support. In other words, an absolute PSNR bound, independent of patch size restrictions, is still unknown.

We make several theoretical contributions with practical implications, towards answering these questions. First we consider non-parametric denoising with a finite external database and finite patch size. We study the dependence of denoising error on patch size. Our main result is a *law of diminishing return*: when the window size is increased, the difficulty of finding enough training data for an input noisy patch directly correlates with diminishing returns in denoising performance. That is, not only is it easier to increase window size for smooth patches, they also benefit more from such an increase. In contrast, textured regions require a significantly larger sample size to increase the patch size, while gaining very little from such an increase. From a practical viewpoint, this analysis suggests an *adaptive strategy* where each pixel is denoised with a variable window size that depends on its local patch complexity.

Next, we put computational issues aside, and study the fundamental limit of denoising, with an infinite window size and a perfectly known $p(x)$ (i.e., an infinite training database). Under a simplified image formation model we study the following question: What is the absolute lower bound on denoising error, and how fast do we converge to it, as a function of window size. We show that the *scale invariance* of natural images plays a key role and yields a power law convergence curve. Remarkably, despite the model's simplicity, its predictions agree well with empirical observations. Extrapolating this parametric law provides a ballpark prediction on the best possible denoising, suggesting that current algorithms may still be improved by about $0.5 - 1$ dB.

2 Optimal Mean Square Error Denoising

In image denoising, given a noisy version $y = x + n$ of a clean image x, corrupted by additive noise n, the aim is to estimate a cleaner version $\hat{x}$. The common quality measure of denoising algorithms is their mean squared error, averaged over all possible clean and noisy x, y pairs, where x is sampled from the density $p(x)$ of natural images

$$\text{MSE} = \mathbb{E}[\|\hat{x} - x\|^2] = \int p(x) \int p(y|x)\|x - \hat{x}\|^2 dy dx \tag{1}$$

It is known, e.g. [14], that for a single pixel of interest x_c the estimator minimizing Eq. (1) is the conditional mean:

$$\hat{x}_c = \mu(y) = \mathbb{E}[x_c|y] = \int \frac{p(y|x)}{p(y)} p(x) x_c dx. \tag{2}$$

Inserting Eq. (2) into Eq. (1) yields that the minimum mean squared error (MMSE) per pixel is the conditional variance

$$\text{MMSE} = \mathbb{E}_y[\mathbb{V}[x_c|y]] = \int p(y) \int p(x|y) \left(x_c - \mu(y)\right)^2 dxdy. \tag{3}$$

The MMSE measures the *inherent ambiguity* of the denoising problem and the statistics of natural images, as any natural image x within the noise level of y may have generated y. Since Eq. (2) depends on the exact unknown density $p(x)$ of natural images (with full image support), it is unfortunately not possible to compute. Nonetheless, by definition it is the theoretically optimal denoising algorithm, and in particular outperforms all other algorithms, even those that detect the class of a picture and then use class-specific priors [3], or those which leverage internal patch repetition [6, 22]. That said, such approaches can yield significant practical benefits when using a finite data.

Finally, note that the density $p(x)$ plays a *dual* role. According to Eq. (1), it is needed for evaluating *any* denoising algorithm, since the MSE is the average over natural images. Additionally, it determines the optimal estimator $\mu(y)$ in Eq. (2).

Finite support: First, we consider algorithms that only use information in a window of d noisy pixels around the pixel to be denoised. When needed, we denote by x_{w_d}, y_{w_d} the restriction of the clean and noisy images to a d-pixel window and by x_c, y_c the pixel of interest, usually the central one with $c = 1$. As in Eq. (3), the optimal MMSE_d of any denoising algorithm restricted to a d pixels support is also the conditional variance, but computed over the space of natural patches of size d rather than on full-size images.

By definition, the optimal denoising error may only decrease with window size d, since the best algorithm seeing $d + 1$ pixels can ignore the last pixel and provide the answer of the d pixels algorithm. This raises two critical questions: *how does MMSE_d decrease with d, and what is MMSE_∞, namely the best achievable denoising error of* **any** *algorithm (not necessarily patch based) ?*

Non-Parametric approach with a finite training set: The challenge in evaluating MMSE_d is that the density $p(x)$ of natural images is unknown. To bypass it, a non-parametric study of MMSE_d for small values of d was made in [14], by approximating Eq. (2) with a discrete sum over a large dataset of clean d-dimensional patches $\{x_i\}_{i=1}^N$.

$$\hat{\mu}_d(y) = \frac{\frac{1}{N}\sum_i p(y_{w_d}|x_{i,w_d}) x_{i,c}}{\frac{1}{N}\sum_i p(y_{w_d}|x_{i,w_d})} \tag{4}$$

where, for iid zero-mean Gaussian noise n with variance σ^2,

$$p(y_{w_d}|x_{w_d}) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|x_{w_d} - y_{w_d}\|^2}{2\sigma^2}}. \tag{5}$$

An interesting conclusion of [14] was that for small patches or high noise levels, existing denoising algorithms are close to the optimal MMSE_d.

For Eq. (4) to be an accurate estimate of $\mu_d(y)$, the given dataset must contain many clean patches at distance $(d\sigma^2)^{1/2}$ from y_{w_d}, which is the expected distance of the original patch, $\mathbb{E}[\|x_{w_d} - y_{w_d}\|^2] = d\sigma^2$. As a result, non-parametric denoising requires a larger training set at low noise levels σ where the distance $d\sigma^2$ is smaller, or at larger patch sizes d where clean patch samples are spread further apart. This curse of dimensionality restricted [14] to small values of d.

In contrast, in this paper we are interested in the best achievable denoising of *any* algorithm, without restrictions on support size, namely MMSE_∞. We thus generalize [14] by studying how MMSE_d decreases as a function of d, and as a result provide a novel prediction of MMSE_∞ (see Section 4).

Note that MMSE_∞ corresponds to an infinite database of all clean images, which in particular also includes the original image x. However, this does not imply that $\text{MMSE}_\infty = 0$, since this database also includes many slight variants of x, with small spatial shifts or illumination changes. Any of these variants may have generated the noisy image y, making it is impossible to identify the correct one with zero error.

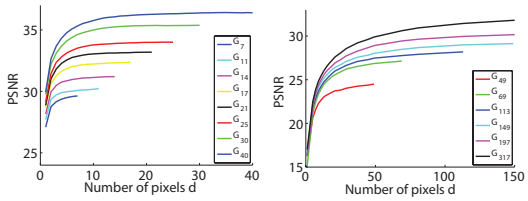
3 Patch Size, Complexity and PSNR Gain

Increasing the window size provides a more accurate prior as it considers the information of distant pixels on the pixel of interest. However, in a non-parametric approach, this requires a much larger training set and it is unclear how substantial the PSNR gain might be. This section shows that this tradeoff depends on "patch complexity", and presents a *law of diminishing return*: patches that require a large increase in database size also benefit little from a larger window. This gain is governed by the statistical dependency of peripheral pixels and the central one: weakly correlated pixels provide little information while leading to a much larger spread in patch space, and thus require a significantly larger training data.

3.1 Empirical study

To understand the dependence of PSNR on window size, we present an empirical study with $M = 10^4$ clean and noisy pairs $\{(x_j, y_j)\}_{j=1}^M$ and $N = 10^8$ samples taken from the LabelMe dataset, as in [14]. We compute the non-parametric mean (Eq. (4)) at varying window sizes d. For each noisy patch we determine the largest d at which estimation is still reliable by comparing the results with different clean subsets[1].

[1] We divide the N clean samples into 10 groups, compute the non-parametric estimator $\hat{\mu}_d(y_j)$ on each group separately, and check if the variance of these 10 estimators is much smaller than σ^2. For small d, samples are dense enough and all these estimators provide consistent results. For large d, sample density is insufficient, and each estimator gives a very different result.



1D signals, $\sigma = 10$ 2D signals, $\sigma = 50$

Fig. 1: *For patch groups G_ℓ of varying complexity, we present PSNR vs. number of pixels d in window w_d, where $d = 1, \ldots, \ell$. Higher curves correspond to smooth regions, which flatten at larger patch dimensions. Textured regions correspond to lower curves which not only run out of samples sooner, but also their curves flatten earlier.*

We divide the M test patches into groups G_ℓ based on the largest window size ℓ at which the estimate is still reliable. For each group, Fig. 1 displays the empirical PSNR averaged over the group's patches as a function of window size d, for $d = 1, \ldots, \ell$ (that is, up to the maximal window size $d = \ell$ at which estimation is reliable), where:

$$\text{PSNR}(G_\ell | w_d) = -10 \log_{10} \left(\frac{1}{|G_\ell|} \sum_{j \in G_\ell} (x_{j,c} - \hat{\mu}_d(y_j))^2 \right)$$

We further compute for each group its mean gradient magnitude, $\|\nabla y_{w_\ell}\|$, and observe that groups with smaller support size ℓ, which run more quickly out of training data, include mostly patches with large gradients (texture). These patches correspond to PSNR curves that are lower and also flatten earlier (Fig. 1). In contrast, smoother patches are in groups that run out of examples later (higher ℓ) and also gain more from an increase in patch width: the higher curves in Fig. 1 flatten later. The data in Fig. 1 demonstrates an important principle: *When an increase in patch width requires many more training samples, the performance gain due to these additional samples is relatively small.*

To understand the relation between patch complexity, denoising gain, and required number of samples, we show that the statistical dependency between adjacent pixels is broken when large gradients are observed. We sample rows of 3 consecutive pixels from clean x and noisy y natural images (Fig. 2(a)), discretize them into 100 intensity bins, and estimate the conditional probability $p(x_1, x_3 | y_1, y_2)$ by counting occurrences in each bin. When the gradient $|y_2 - y_1|$ is high with respect to the noise level, x_1, x_3 are approximately independent, $p(x_1 = i, x_3 = j | |y_1 - y_2| \gg \sigma) \approx p_1(i) p_3(j)$, see Fig. 2(d,f). In contrast, small gradients don't break the dependency, and we observe a much more elongated structure, see Fig. 2(b,c,e). For reference, Fig. 2(g) shows the unconditional joint distribution $p(x_1, x_3)$, without seeing any y. Its diagonal structure implies that while the pixels (x_1, x_3) are by default dependent, the dependency is broken in the presence of a strong edge between them. From a practical perspective, if $|y_1 - y_2| \gg \sigma$, adding the pixel y_3 does not contribute much to the estimation of x_1. If the gradient $|y_1 - y_2|$ is small there is still dependency between x_3 and x_1, so adding y_3 does further reduce the reconstruction error. A simple explanation for this phenomenon

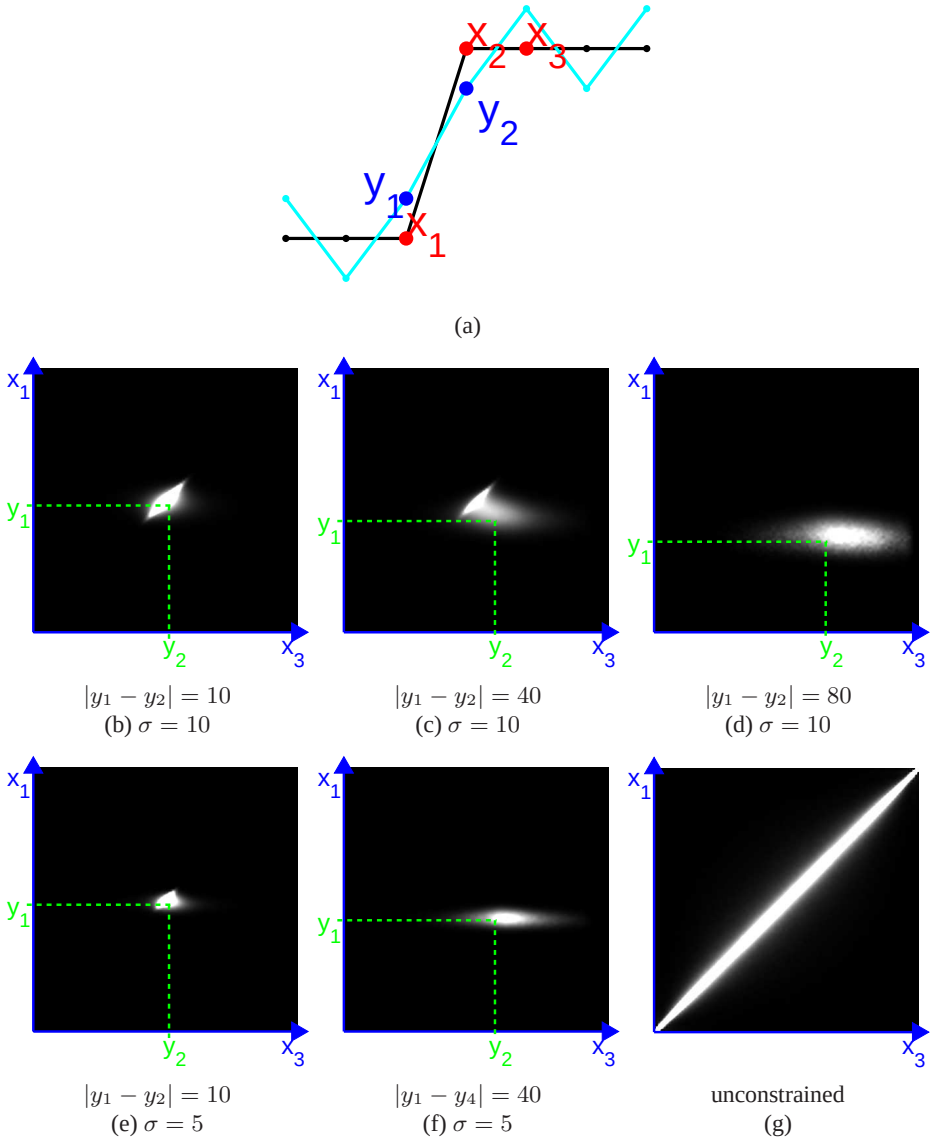
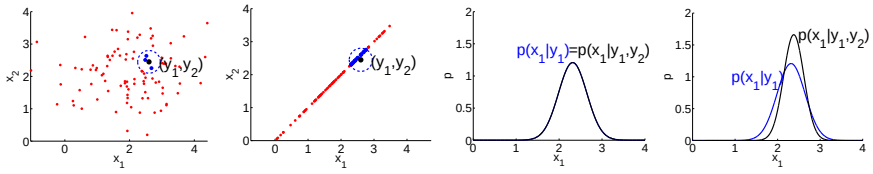


Fig. 2: *(a) A clean and noisy 1D signal. (b-g) Joint distribution tables. (b-f)* $p(x_1, x_3|y_1, y_2)$ *at two noise levels. (g)* $p(x_1, x_3)$*, before any observation. While neighboring pixels are dependent in default, the dependency is broken when the observed gradient is high with respect to the noise(d,f).*

is to think of adjacent objects in an image. As objects can have independent colors, the color of one object tells us nothing about its neighbor on the other side of the edge.



(a) Independent (b) Fully Dependent (c) Independent (d) Fully Dependent

Samples from joint distributions Marginal distributions

Fig. 3: *A toy example of 2D sample densities.*

3.2 Theoretical Analysis

Motivated by Fig. 1 and Fig. 2, we study the implications of partial statistical dependence between pixels, both on the performance gain expected by increasing the window size, and on the requirements on sample size.

2D Gaussian case: To gain intuition, we first consider a trivial scenario where patch size is increased from 1 to 2 pixels and distributions are Gaussians. In Fig. 3(a), x_1 and x_2 are independent, while in Fig. 3(b) they are fully dependent and $x_1 = x_2$. Both cases have the same marginal distribution $p(x_1)$ with equal denoising performance for a 1-pixel window. We draw $N = 100$ samples from $p(x_1, x_2)$ and see how many of them fall within a radius σ around a noisy observation (y_1, y_2). In the uncorrelated case (Fig. 3(a)), the samples are spread in the 2D plane and therefore only a small portion of them fall near (y_1, y_2). In the second case, since the samples are concentrated in a significantly smaller region (a 1-D line), there are many more samples near (y_1, y_2). Hence, in the fully correlated case a non parametric estimator requires a significantly smaller dataset to have a sufficient number of clean samples in the vicinity of y.

To study the accuracy of restoration, Fig. 3(c,d) shows the marginal distributions $p(x_1|y_1, y_2)$. When x_1, x_2 are independent, increasing window size to take y_2 into account provides no information about x_1, and $p(x_1|y_1) = p(x_1|y_1, y_2)$. Worse, denoising performance decreases when the window size is increased because we now have fewer training patches inside the relevant neighborhood. In contrast, in the fully correlated case, adding y_2 provides valuable information about x_1, and the variance of $p(x_1|y_1, y_2)$ is half of the variance given y_1 alone. This illustrates how high correlation between pixels yields a significant decrease in error without requiring a large increase in sample size. Conversely, weak correlation gives only limited gain while requiring a large increase in training data.

General derivation: We extend our 2D analysis to d dimensions. The following claim, proved in the appendix, provides the leading error term of the non-parametric estimator $\hat{\mu}_d(y)$ of Eq.(4) as a function of training set size N and window size d. It is similar to results in the statistics literature on the MSE of the Nadaraya-Watson estimator.

Claim. Asymptotically, as $N \to \infty$, the expected non-parametric MSE with a window of size d pixels is

$$\mathbb{E}_N[\mathit{MSE}_d(y)] = \mathit{MMSE}_d(y) + \tfrac{1}{N}\mathcal{V}_d(y) + o\left(\tfrac{1}{N}\right) \tag{6}$$

$$\mathcal{V}_d \approx \frac{\mathbb{V}[x_1|y_{w_d}]\,|\Phi_d|}{\sigma^{2d}}, \tag{7}$$

with $\mathbb{V}[x_1|y_{w_d}]$ the conditional variance of the central pixel x_1 given a window w_d from y, and $|\Phi_d|$ is the determinant of the local $d \times d$ covariance matrix of $p(y)$,

$$|\Phi_d|^{-1} = \left| -\frac{\partial^2 \log p(y_{w_d})}{\partial^2 y_{w_d}} \right|. \tag{8}$$

The expected error is the sum of the fundamental limit $\text{MMSE}_d(y)$ and a variance term that accounts for the finite number of samples N in the dataset. As in Monte-Carlo sampling, it decreases as $\frac{1}{N}$. When window size increases, $\text{MMSE}_d(y)$ decreases, but the variance $\mathcal{V}_d(y)$ might increase. The tension between these two terms determines whether for a constant training size N increasing window size is beneficial.

The variance $\mathcal{V}_d$ is proportional to the volume of $p(y_{w_d})$, as measured by the determinant $|\Phi_d|$ of the local covariance matrix. When the volume of the distribution is larger, the N samples are spread over a wider area and there are fewer clean patches near each noisy patch y. This is precisely the difference between Fig. 3(a) and Fig. 3(b).

For the error to be close to the optimal MMSE_d, the term $\mathcal{V}_d/N$ in Eq. (6) must be small. Eq. (7) shows that $\mathcal{V}_d$ depends on the volume $|\Phi_d|$ and we expect this term to grow with dimension d, thus requiring many more samples N. Both our empirical data and our 2D analysis show that the required increase in sample size is a function of the statistical dependencies of the central pixel with the added one.

To understand the required increase in training size N when window size is increased by one pixel from $d-1$ to d, we analyze the ratio of variances $\mathcal{V}_d/\mathcal{V}_{d-1}$. Let $g_d(y)$ be the gain in performance (for an infinite dataset), which according to Eq. (3) is given by:

$$g_d(y) = \frac{\text{MMSE}_{d-1}(y)}{\text{MMSE}_d(y)} = \frac{\mathbb{V}[x_1|y_1, \dots y_{d-1}]}{\mathbb{V}[x_1|y_1, \dots y_d]} \tag{9}$$

We also denote by $g_d^*(y)$ the ideal gain if x_d and x_1 were perfectly correlated, i.e. $r = cor(x_1, x_d \,|\, y_1, \dots, y_{d-1}) = 1$. The following claim shows that when $\text{MMSE}_d(y)$ is most improved, sampling is not harder since the volume and variance $\mathcal{V}_d$ do not grow. For simplicity, we prove the claim in the Gaussian case.

Claim. Let $p(y)$ be Gaussian. When increasing the patch size from $d-1$ to d, the variance ratio and the performance gain of the estimators are related by:

$$\frac{\mathcal{V}_d}{\mathcal{V}_{d-1}} = \frac{g_d^*}{g_d} \geq 1. \tag{10}$$

That is, the ratio of variances equals the ratio of optimal denoising gain to the achievable gain. When x_1, x_d are perfectly correlated, $g_d = g_d^*$, we get $\mathcal{V}_d/\mathcal{V}_{d-1} = 1$, and a larger window gives improved restoration results without increasing the required dataset size. In contrast, if x_d, x_1 are weakly correlated, increasing window size requires a bigger dataset to keep $\mathcal{V}_d/N$ small, and yet the PSNR gain is small.

Proof. Let C be the 2×2 covariance of x_1, x_d given $y_1, \ldots, y_{d-1}$ (before seeing y_d)

$$C = Cov(x_1, x_d | y_1, \ldots y_{d-1}) = \begin{pmatrix} c_1 & c_{12} \\ c_{12} & c_2 \end{pmatrix} \tag{11}$$

and let $r = c_{12}/\sqrt{c_1 c_2}$ be the correlation between x_1, x_d.

Assuming that the distribution is locally Gaussian, upon observing y_d, the marginal variance of x_1 decreases from c_1 to the following expression (see Eq. 2.73 in [5]),

$$\mathbb{V}[x_1 | y_1, \ldots, y_d] = c_1 - \frac{c_{12}^2}{c_2 + \sigma} = c_1 \left(1 - \frac{c_{12}^2/c_1}{c_2 + \sigma^2}\right) = c_1 \frac{c_2(1 - r^2) + \sigma^2}{c_2 + \sigma^2}. \tag{12}$$

Hence the contribution to performance gain of the additional pixel y_d is

$$g_d = \frac{\mathbb{V}[x_1 | y_1, \ldots y_{d-1}]}{\mathbb{V}[x_1 | y_1, \ldots y_d]} = \frac{c_2 + \sigma^2}{c_2(1 - r^2) + \sigma^2}. \tag{13}$$

When $r = 1$, the largest possible gain from y_d is $g_d^* = (c_2 + \sigma^2)/\sigma^2$. The ratio of best possible gain to achieved gain is

$$\frac{g_d^*}{g_d} = \frac{c_2(1 - r^2) + \sigma^2}{\sigma^2}. \tag{14}$$

Next, let us compute the ratio $\mathcal{V}_d/\mathcal{V}_{d-1}$. For Gaussian distributions, according to Eq. 2.82 in [5], the conditional variance of y_d given $y_1, \ldots, y_{d-1}$ is independent of the specific observed values. Further, since $p(y_1, \ldots, y_d) = p(y_1, \ldots, y_{d-1}) p(y_d | y_1, \ldots y_{d-1})$, we obtain that

$$|\Phi_d| = \mathbb{V}(y_d | y_1, \ldots y_{d-1}) |\Phi_{d-1}| \tag{15}$$

This implies that

$$\frac{\mathcal{V}_d}{\mathcal{V}_{d-1}} = \frac{\mathbb{V}(y_d | y_1, \ldots y_{d-1})}{\sigma^2} \frac{\mathbb{V}[x_1 | y_1, \ldots y_d]}{\mathbb{V}[x_1 | y_1, \ldots y_{d-1}]} \tag{16}$$

Next, since $y_d = x_d + n_d$ with $n_d \sim N(0, \sigma^2)$ independent of $y_1, \ldots, y_{d-1}$, then $\mathbb{V}(y_d | y_1, \ldots y_{d-1}) = c_2 + \sigma^2$. Thus,

$$\frac{\mathcal{V}_d}{\mathcal{V}_{d-1}} = \frac{c_2 + \sigma^2}{\sigma^2} \frac{c_2(1 - r^2) + \sigma^2}{c_2 + \sigma^2} = \frac{g_d^*}{g_d}.$$

□

To understand the growth of $\mathcal{V}_d$, consider two extreme cases, similar to Fig. 3. First, consider a signal whose pixels are all independent with variance γ. In this case $r =$

σ	20	35	50	75	100
Optimal Fixed	32.4	30.1	28.7	27.2	26.0
Adaptive	33.0	**30.5**	**29.0**	**27.5**	**26.4**
BM3D	**33.2**	30.3	28.6	26.9	25.6

Table 1: *Adaptive and fixed window denoising results in PSNR.*

0 and $c_2 = \gamma$ (since independence implies that seeing $y_1, \ldots y_{d-1}$ does not reduce the variance of x_d), hence for every additional dimension d, $g_d^*/g_d = (\gamma + \sigma^2)/\sigma^2$. That is, $\mathcal{V}_d \propto ((\gamma + \sigma^2)/\sigma^2)^d$ increases exponentially with the patch dimension, and thus, to control $\mathcal{V}_d/N$, there is also an exponential increase in the required number of samples N. However, if the pixels are independent there is no point in increasing the patch size as additional pixels provide no information on x_1. At the other extreme, of a perfectly correlated signal, $\mathcal{V}_d$ is constant independent of d. Moreover, increasing the patch dimension is very informative and can be done without any further increase in N. In the intermediate case of partial correlation between x_1, x_d (that is $0 < r < 1$), increasing the patch dimension provides limited reduction in error and requires some increase in sample size. As the error reduction is inversely proportional to the required number of samples, weak correlation not only leads to small gains, but also requires a large number of samples.

3.3 Adaptive Denoising

Our findings above motivate an *adaptive* denoising scheme [12] where each pixel is denoised with a variable patch size that depends on the local image complexity around it. To test this idea, we devised the following scheme. Given a noisy image, we denoise each pixel using several patch widths and multiple disjoint clean samples. As before, we compute the variance of all these different estimates, and select the largest width for which the variance is still below a threshold. Table 1 compares the PSNR of this adaptive scheme to fixed window size non-parametric denoising using the optimal window size at each noise level, and to BM3D [9], a state-of-the-art algorithm. We used $M = 1000$ test pixels and $N = 7 \cdot 10^9$ clean samples. At all considered noise levels, the adaptive approach significantly improves the fixed patch approach, by about $0.3 - 0.6$dB. At low noise levels, sample size N is too small, and adaptive denoising is worse than BM3D[2]. At higher noise levels it increasingly outperforms BM3D.

Fig. 4 visualizes the difference between the adaptive and fixed patch size approaches, at noise level $\sigma = 50$. When patch size is small, noise residuals are highly visible in the flat regions. With a large patch size, one cannot find good matches in the textured regions, and as a result noise is visible around edges. Both edges and flat regions are handled properly by the adaptive approach. Moreover, under perceptual error metrics

[2] The reason is that at this finite N, with $\sigma = 20$ our non-parametric approach uses 5×5 patches at textured regions. In contrast, BM3D uses 8×8 ones, with additional algorithmic operations which allow it to better generalize from a limited number of samples.

(a)Original (b)Noisy input (c)Adaptive (d)Fixed $k = 5$ (e)Fixed $k = 6$ (f)Fixed $k = 10$

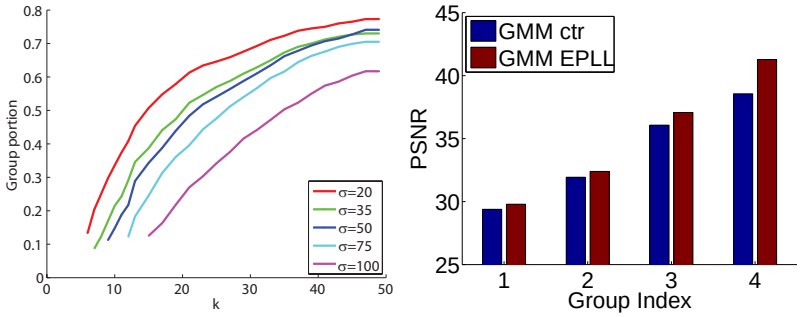
Fig. 4: *Visual comparison of adaptive vs. fixed patch size non parametric denoising (optimal fixed size results obtained with $k = 6$). A fixed patch has noise residuals either in flat areas(d,e), or in textured areas(f).*

such as SSIM [21], decreasing the error in the smooth regions is more important, thus underscoring the potential benefits of an adaptive approach.

Note that this adaptive non-parametric denoising is not a practical algorithm, as Fig. 4 required several days of computation. Nonetheless, these results suggest that adaptive versions to existing denoising algorithms such as [11, 9, 15, 10, 23] and other low-level vision tasks are a promising direction for future research.

Window size and noise variance: Another interesting question is the relation between the optimal window size and the noise level. Fig. 5(a) shows, for several noise levels, the percentage of test examples for which the adaptive approach selected a square patch of width smaller than k. Unsurprisingly, with the same number of samples N, when the noise level is high, larger patches are used since the non parametric approach essentially averages all samples within a Gaussian window of variance σ^2 around the noisy observation, so for large noise the neighborhood definition is wider and includes more samples. This property is implicitly used by other denoising algorithms. For example, BM3D [9] uses 8×8 windows at noise s.t.d below 40 and 12×12 windows at higher noise levels. Similarly, Bilateral filtering denoising algorithms [4] estimate a pixel as an adaptive average of its neighbors, where the neighbor weight is significantly reduced when an intensity discontinuity is observed. However, the discontinuity measure is relative to the noise level and only differences above the noise standard deviation actually reduce the neighbor weight. Thus, effectively, at higher noise levels Bilateral filtering averages over a wider area.

Our analysis suggests that this is not only an issue of sample density but an inherent property of the statistics of natural images. At high noise levels larger patches are indeed useful, while at low noise level increasing the patch size provides less information. One way to see this is to reconsider the conditional distribution tables of Fig. 2. For low noise a smaller gradient is sufficient to make the x_1, x_3 independent. e.g., we display conditional distribution tables for 2 noise levels $\sigma = 5$ and $\sigma = 10$. A gradient of $|y_1 - y_2| = 40$ was enough to make the distribution independent at $\sigma = 5$ but not yet at $\sigma = 10$. This is because the amount of noise limits the minimal contrast at which an edge is identified – gradients whose contrast is below the noise standard deviation can be explained as noise and not as real edges between different segments. As a result, the optimal denoising does average the values from the other side of a low contrast



(a) (b)

Fig. 5: *(a) Cumulative histogram of the portion of test examples using patch size below k, for varying noise levels. The patch size was selected automatically by the algorithm. When the noise variance is high larger patches are used. (b) The average gain of the EPLL algorithm in 4 groups of varying complexity (flatness). Most improvement is in the flat patches of group 4.*

edge. This implies that optimal denoising takes into account pixels from the other side of weak edges and thus, at high noise levels wider regions are useful. This is also the case in Bilateral filtering, which averages neighbors from the other side of edges whose contrast is below the noise standard deviation.

Denoising of smooth regions in previous works An interesting outcome of our analysis is that patch based denoising can be improved mostly in flat areas and less in textured ones. We now show that this property is implicit in several recent denoising papers.

Patch complexity and the EPLL algorithm: One interesting approach to analyze in this context is the EPLL algorithm of [23]. The authors learned a mixture of Gaussians prior over 8×8 image patches, but instead of denoising each patch independently, they then apply an optimization process to improve the Expected Patch Log Likelihood of all overlapping patches in the image. What is the actual source of improvement of the EPLL algorithm? To test that we divided $M = 1000$ test examples (x_i, y_i) to 4 groups according to the corresponding maximal patch width in our adaptive non-parametric approach. Effectively, groups 1 and 2 contained mostly textured and edge patches, whereas groups 3 and 4 contained mostly smooth and flat patches.

We denoised each test example y_i with the direct GMM prior applied to the 8×8 patch around it, and compared that with the result after the additional EPLL optimization aiming to achieve agreement between overlapping patches. In accordance with our analysis, Fig. 5(b) shows that the gain from the EPLL step is larger at flat regions, and almost insignificant at highly textured ones.

The local patch search: In [22] Zontak and Irani explore the relation between internal and external patch searches. In particular they observe that for simple flat patches, de-

noising results of non local means [6] with a small 5×5 window (which are far from optimal), can be improved if the internal patch search is not performed over the entire image, but is restricted to a local neighborhood around the pixel of interest. The explanation of [22] is that for textured areas the probability of finding relevant neighbors within the local neighborhood patches is too low.

Our analysis provides an alternative explanation for these findings. In textured regions there is inherently far less statistical dependency among local pixels, as compared to flat regions. The local patch search can be interpreted as a way to use information from a wider window around the pixel of interest. In flat regions denoising is approximately equivalent to averaging the pixel values over the whole region. Clearly, averaging over a wider flat region reduces the error, which is precisely what is implicitly achieved by restricting the patch search to a local image neighborhood.

Image dependent optimal denoising: In [8] the authors derived, under some simplifying assumptions, image-specific lower bounds on the optimal possible denoising. Comparing these lower bounds to the results of existing algorithms, [8] concluded that for textured natural images existing algorithms are close to optimal, whereas for synthetic piecewise constant images there is still a large room for improvement. These findings are consistent with our analysis, that in flat regions a large support can improve denoising results. Thus, current algorithms, tuned to perform well on textured regions, and working with fixed small patch sizes, can be improved considerably in smooth image regions.

4 The Convergence and Limits of Optimal Denoising

In this section, we put computational and database size issues aside, and study the behavior of optimal denoising error as window size increases to infinity. Fig. 1 shows that optimal denoising yields a diminishing return beyond a window size that varies with patches. Moreover, patches that plateau at larger window sizes also reach a higher PSNR. Fig. 2 shows that strong edges break statistical correlation between pixels. Combining the two suggests that each pixel has a finite compact region of informative pixels. Intuitively, the size distribution of these regions must directly impact both denoising error vs. window size and its limit with an infinite window.

We make two contributions towards elucidating this question. First we show that a combination of the simplified *dead leaves* image formation model, together with *scale invariance* of natural images implies both a *power-law* convergence, $\mathrm{MMSE}_d \sim e + c/d$, as well as a strictly positive lower bound on the optimal denoising with infinite window, $\mathrm{MMSE}_\infty = e > 0$. Next, we present empirical results showing that despite the simplicity of this model, its conclusions match well the behavior of real images.

4.1 Scale-invariance and Denoising Convergence

We consider a *dead leaves* image formation model, e.g. [1], whereby an image is a random collection of piecewise constant segments, whose size is drawn from a scale-invariant distribution and whose intensity is drawn i.i.d. from a uniform distribution. This yields perfect correlation between pixels in the same region, as in Fig. 3(b).

To further simplify the analysis, we conservatively assume an edge oracle which gives the exact locations of edges in the image. The optimal denoising is then to average all observations in a segment. For a pixel belonging to segment of size s pixels, the MMSE is σ^2/s. Overall the expected reconstruction error with infinite-sized windows is

$$\text{MMSE} = \int p(s)\frac{\sigma^2}{s}ds \tag{17}$$

where $p(s)$ is the probability that a pixel belongs to a segment with s pixels. The optimal error is strictly larger than zero if the probability of finite segments is larger than zero. Without the edge-oracle, the error is even higher.

Scale invariance: A short argument [1] which we review below for completeness, shows that the probability that a random image pixel belongs to a segment of size s is of the form $p(s) \propto 1/s$. In a Markov model, in contrast, $p(s)$ decays exponentially fast with s [19].

Claim. Let $p(s)$ denote the probability that a uniformly sampled pixel belongs to a segment of size s pixels in a scale invariant distribution. Then

$$p(s) \propto \frac{1}{s}. \tag{18}$$

Proof. Let

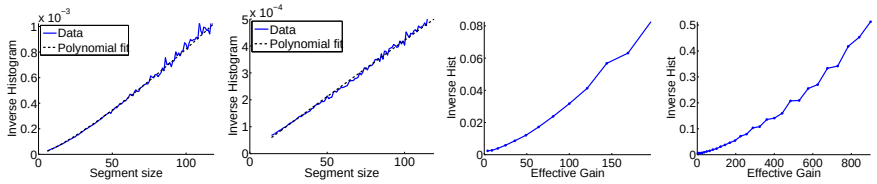
$$F(t_1, t_2) = \int_{t_1}^{t_2} p(s)ds \tag{19}$$

denote the probability of a pixel belonging to an object of size $t_1 \leq s \leq t_2$. Scale invariance implies that this probability does not change when the image is scaled, hence for every a, t_1, t_2 $F(t_1, t_2) = F(at_1, at_2)$. This implies that

$$\int_{t_1}^{t_2} p(s)ds = \int_{at_1}^{at_2} p(s)ds \tag{20}$$

and hence $p(s) = ap(as)$. The only distribution satisfying this property is $p(s) \propto 1/s$, since, e.g. by substituting $a = 1/s$ we get that $p(s) = 1/s \cdot p(1)$. □

The power law distribution of segment sizes was also previously used [19] to argue that Markov models cannot capture the distribution of natural images, since in a Markov model the probability of observing a uniform segment should decay exponentially fast. To see this, consider 1D signals and let $p(x_i \approx x_{i-1}) = a$ for some constant a. In a first



$T = 5, h(d) \propto d^{-1.2}$ $T = 20, h(d) \propto d^{-1}$ $\sigma = 35$ $\sigma = 75$

(a) Inverse histogram of segment lengths (b) Inverse histogram of reconstruction error

Fig. 6: *(a) Inverse histograms of segment lengths follow a scale invariant distribution. (b) Inverse histograms of $\sigma^2/(x_i - \hat{y}_i)^2$ exhibit a power law, similar to the distribution of segment sizes.*

order Markov process the probability of observing a segment of length d is proportional to

$$\Pi_{i=1}^{d} p(x_i \approx x_{i-1}) = a^d, \tag{21}$$

since the memory-less definition of a Markov model implies that the probability of the i'th pixel depends only on pixel $i-1$ and not on any of the previous ones. Thus, the distribution of segment sizes in a Markov model decays exponentially. This result is not restricted to the case of a first order Markov model and one can show that the exponential decay holds for a Markov model of any order. However, empirically the distribution of segment areas in natural images decays only polynomially and not exponentially fast.

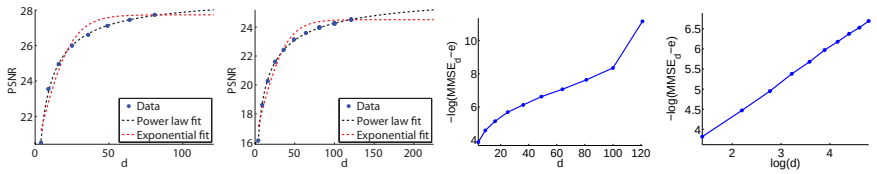
To get a sense of the empirical size distribution of nearly-constant-intensity regions in natural images, we perform a simple experiment inspired by [1]. For a random set of pixels $\{x_i\}$, we compute the size $d(i)$ of the connected region whose pixel values differ from x_i by at most a threshold T: $d(i) = \#\{x_j || x_j - x_i| \leq T\}$. The empirical histogram $h(d)$ of region sizes follows a power law behavior $h(d) \propto d^{-\alpha}$ with $\alpha \approx 1$, as shown in Fig. 6(a,b), which plots $1/h(d)$.

Optimal denoising as a function of window size: We now compute the optimal denoising for the dead leaves model with the scale invariance property. Since $1/s$ is not integrable, scale invariance cannot hold at infinitely large scales. Assuming it holds up to a maximal size $D \gg 1$, gives the normalized probability

$$p_D(s) = \frac{s^{-1}}{\int_1^D s^{-1} ds} = \frac{1}{\ln D}\frac{1}{s}. \tag{22}$$

We compute the optimal error with a window of size $d \ll D$ pixels. Given the edge oracle, every segment of size $s \leq d$ attains its optimal denoising error of σ^2/s, whereas if $s > d$ we obtain only σ^2/d. Splitting the integral in (17) into these two cases gives

$$\begin{aligned} \text{MMSE}_d &= \int_1^d \tfrac{\sigma^2}{s} p_D(s) ds + \int_d^D \tfrac{\sigma^2}{d} p_D(s) ds \\ &= \int_1^D \tfrac{\sigma^2}{s} p_D(s) ds + \sigma^2 \int_d^D \left(\tfrac{1}{d} - \tfrac{1}{s}\right) p_D(s) ds \\ &= \text{MMSE}_D + \tfrac{\sigma^2}{d}\left(1 - \tfrac{\ln d + 1}{\ln D}\right) + \tfrac{\sigma^2}{D \ln D} \approx \text{MMSE}_D + \frac{\sigma^2}{d} \end{aligned} \tag{23}$$



(a) $\sigma = 50$ (b) $\sigma = 100$ (c) Log plot (d) Log log plot

Fig. 7: *PSNR vs. patch dimension. A power law fits the data well, whereas an exponential law fits poorly. Panels (c) and (d) show* $|\log(MMSE_d - e)|$ *v.s.* d *or* $\log(d)$*. An exponential law should be linear in the first plot, a power law linear in the second.*

For this model, $\text{MMSE}_\infty = \text{MMSE}_D$. Thus, *the dead leaves model with scale invariance property implies a power law* $1/d$ *convergence to a strictly positive* $MMSE_\infty$.

4.2 Empirical validation and optimal PSNR

While dead leaves is clearly an over-simplified model, it captures the salient properties of natural images. Even though real images are not made of piecewise constant segments, the results of Sec. 3, and Fig. 6 suggest that each image pixel has a finite "informative region", whose pixel values are most relevant for denoising it. While for real images, correlations may not be perfect inside this region and might not fully drop to zero outside it, we now show that empirically, optimal denoising in natural images indeed follows a power law similar to that of the dead-leaves model.

To this end, we apply the method of [14] and compute the optimal patch based MMSE_d for several small window sizes d. Fig. 7(a-b) show that consistent with the dead leaves model, we obtain an excellent fit to a power law $\text{MMSE}_d = e + \frac{c}{d^\alpha}$ with $\alpha \approx 1$. In contrast, we get a poor fit to an exponential law, $\text{MMSE}_d = e + cr^{-d}$, implied by the common Markovian assumption [19]. In addition, Fig. 7(c,d) show log and log-log plots of $(\text{MMSE}_d - e)$, with the best fitted e in each case. The linear behavior in the log-log plot (Fig. 7(d)) further supports the power law.

As an additional demonstration of the scale-invariance of natural images in the denoising context, we evaluate the distribution of denoising error over pixels. For a large collection of image pixels $\{x_i\}$ we compute the histogram of $\sigma^2/(x_i - \hat{y}_i)^2$. Fig. 6(c,d) shows that the resulting inverted histogram approximately follows a polynomial curve. Recall that in the idealized dead-leaves model, a perfectly uniform segment of size ℓ yields an error of σ^2/ℓ. Hence, under scale invariance, we expect a linear fit to the histograms of Fig. 6(c,d). While in real natural images, denoising is not simply an average over the pixels in each segment, interestingly, the inverse histogram is almost linear, matching the prediction of the dead-leaves model.

Predicting Optimal PSNR: For small window sizes, using a large database and Eq. (4), we can estimate the optimal patch-based denoising MMSE_d. Fig. 7 shows that the curve of MMSE_d is accurately fitted by a power law $\text{MMSE}_d = e + c/d^\alpha$, with $\alpha \approx 1$. To

σ	35	50	75	100
Extrapolated bound	30.6	28.8	27.3	26.3
KSVD [11]	28.7	26.9	25.0	23.7
BM3D [9]	30.0	28.1	26.3	25.0
EPLL [23]	29.8	28.1	26.3	25.1

Table 2: *Extrapolated optimal denoising in PSNR, and the results of recent algorithms. A modest room for improvement exists.*

fit the curve MMSE_d robustly, for each d value we split the N samples to 10 different groups, compute $PSNR_d$ from each of them, and compute the variance in the estimation η_d^2. We used gradient descent optimization to search for e, c, α minimizing

$$\sum_d w_d \frac{(-10log10(e + c/d^\alpha) - PSNR_d)^2}{\eta^2} \tag{24}$$

where the weights w_d account for the fact that the sample of d values is not uniform as we have evaluated only d values of the form $d = k^2$ (squared patches).

Given the fitted parameters, the curve $\text{MMSE}_d = e + c/d^\alpha$, can be extrapolated and we can predict the value of MMSE_∞, which is the best possible error of *any* denoising algorithm (not necessarily patch based). Since the power law is only approximate, this extrapolation should be taken with a grain of salt. Nonetheless, it gives an interesting ballpark estimate on the amount of further achievable gain by any future algorithmic improvements. Table 2 compares the PSNR of existing algorithms to the predicted PSNR_∞, over $M = 20,000$ patches using the power law fit based on $N = 10^8$ clean samples[3]. The comparison suggests that current methods may still be improved by $0.5 - 1$dB. While the extrapolated value may not be exact, our analysis does suggest that there are inherent limits imposed by the statistics of natural images, which cannot be broken, no matter how sophisticated future denoising algorithms will be.

5 Discussion

In this paper we sted both computational and information aspects of image denoising. Our analysis revealed an intimate relation between denoising performance and the scale invariance of natural image statistics. Yet, only few approaches account for it [18]. Our findings suggest that scale invariance can be an important cue to explore in the development of future natural image priors. In addition, adaptive patch size approaches are a promising direction to improve current algorithms, such as [11, 9, 15, 10, 23].

Our work also highlights the relation between the frequency of occurrence of a patch, local pixel correlations, and potential denoising gains. This concept is not restricted to the denoising problem, and may have implications in other fields.

[3] The numerical results in Tables 1,2 are not directly comparable, since Table 1 was computed on a small subset of only $M = 1,000$ test examples, but with a larger sample size N.

Acknowladgments: We thank ISF, BSF, ERC, Intel, Quanta and NSF for funding.

References

1. L. Alvarez, Y. Gousseau, and J. Morel. The size of objects in natural images, 1999.
2. S. Arietta and J. Lawrence. Building and using a database of one trillion natural-image patches. *IEEE Computer Graphics and Applications*, 2011.
3. S. Baker and T. Kanade. Limits on super-resolution and how to break them. *PAMI*, 2002.
4. D. Barash and D. Comaniciu. A common framework for nonlinear diffusion, adaptive smoothing, bilateral filtering and mean shift. *Image Vision Comput.*, 2004.
5. C. M. Bishop. *Pattern recognition and machine learning*. Springer, 2006.
6. A. Buades, B. Coll, and J. Morel. A review of image denoising methods, with a new one. *Multiscale Model. Simul.*, 2005.
7. D. Chandler and D. Field. Estimates of the information content and dimensionality of natural scenes from proximity distributions. *J. Opt. Soc. Am.*, 2007.
8. P. Chatterjee and P. Milanfar. Is denoising dead? *IEEE Trans Image Processing*, 2010.
9. K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans Image Processing*, 2007.
10. W. Dong, X. Li, L. Zhang, and G. Shi. Sparsity-based image denoising vis dictionary learning and structural clustering. In *CVPR*, 2011.
11. M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans Image Processing*, 2006.
12. C. Kervrann and J. Boulanger. Optimal spatial adaptation for patch-based image denoising. *ITIP*, 2006.
13. A. Lee, K. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *IJCV*, 2003.
14. A. Levin and B. Nadler. Natural image denoising: optimality and inherent bounds. In *CVPR*, 2011.
15. J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman. Non-local sparse models for image restoration. In *ICCV*, 2009.
16. S. Osindero and G. Hinton. Modeling image patches with a directed hierarchy of markov random fields. *NIPS*, 2007.
17. J. Polzehl and V. Spokoiny. Image denoising: Pointwise adaptive approach. *Annals of Statistics*, 31:30–57, 2003.
18. J. Portilla, V. Strela, M. Wainwright, and E. Simoncelli. Image denoising using scale mixtures of gaussians in the wavelet domain. *IEEE Trans Image Processing*, 2003.
19. X. Ren and J. Malik. A probabilistic multi-scale model for contour completion based on image statistics. In *ECCV*, 2004.
20. S. Roth and M.J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, 2005.
21. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. on Image Processing*, 2004.
22. M. Zontak and M. Irani. Internal statistics of a single natural image. In *CVPR*, 2011.
23. D. Zoran and Y. Weiss. From learning models of natural image patches to whole image restoration. In *ICCV*, 2011.

6 Appendix

Claim. The error of a non parametric estimator in a $k \times k$ patch, can be expressed as

$$MSE^{NP}(y) = \mathbb{V}[x|y] + \frac{1}{N}\mathcal{V}(y) + o\left(\frac{1}{N}\right) \tag{25}$$

with

$$\mathcal{V}(y) = \frac{p_{\sigma^*}(y)}{(4\pi\sigma^2)^{k^2/2} p_\sigma(y)^2}\left(\mathbb{V}_{\sigma^*}[x_c|y] + (\mathbb{E}_\sigma[x_c|y] - \mathbb{E}_{\sigma^*}[x_c|y])^2\right) \tag{26}$$

Where y is a shorten notation for $y_{w_{k\times k}}$, $\sigma^* = \sigma/\sqrt{2}$, and $p_\sigma(\cdot), p_{\sigma^*}(\cdot), \mathbb{E}_\sigma[\cdot], \mathbb{E}_{\sigma^*}[\cdot]$ denote probability and expectation of random variables with noise variance σ, σ^* respectively.

Proof. The non parametric estimator is defined as

$$\hat{\mu}(y) = \frac{\frac{1}{N}\sum_i p(y|x_i)x_{i,c}}{\frac{1}{N}\sum_i p(y|x_i)}, \tag{27}$$

For a particular set of N samples $\{x_i\}$, its error is

$$\mathbb{E}_\sigma\left[(x_c - \hat{\mu}(y))^2|y\right] = \mathbb{E}_\sigma[x_c^2|y] - 2\mathbb{E}_\sigma[x_c|y]\hat{\mu}(y) + \hat{\mu}(y)^2 \tag{28}$$

In expectation over all possible sequences of N samples from $p(x)$ the estimator error is

$$\begin{aligned} MSE^{NP}(y) &= \mathbb{E}_N\left[\mathbb{E}_\sigma\left[(x_c - \hat{\mu}(y))^2|y\right]\right] \\ &= \mathbb{E}_\sigma[x_c^2|y] - 2\mathbb{E}_\sigma[x_c|y]\mathbb{E}_N[\hat{\mu}(y)] + \mathbb{E}_N[\hat{\mu}(y)^2] \end{aligned} \tag{29}$$

We thus have to compute what is the expected value of $\mathbb{E}_N[\hat{\mu}(y)], \mathbb{E}_N[\hat{\mu}(y)^2]$. For ease of notation, we will sometimes drop the $_N$, $_\sigma$ subscripts.

We denote by $A(\sigma, k) = (4\pi\sigma^2)^{-k^2/2}$ and use the following equalities

$$\begin{array}{ll} \mathbb{E}[p(y|x)] & = \int p(x)p(y|x)dx = p_\sigma(y) \\ \mathbb{E}[p(y|x)x_c] & = p_\sigma(y)\mathbb{E}_\sigma[x_c|y] \\ \mathbb{E}[p(y|x)x_c^2] & = p_\sigma(y)\mathbb{E}_\sigma[x_c^2|y] \\ \mathbb{E}[p(y|x)^2] & = \int p(x)\frac{e^{-\|x-y\|^2/\sigma^2}}{(2\pi\sigma^2)^{k^2}}dx = A(\sigma,k)p_{\sigma*}(y) \\ \mathbb{E}[p(y|x)^2 x_c] & = A(\sigma,k)p_{\sigma*}(y)\mathbb{E}_{\sigma^*}[x_c|y] \\ \mathbb{E}[p(y|x)^2 x_c^2] & = A(\sigma,k)p_{\sigma*}(y)\mathbb{E}_{\sigma^*}[x_c^2|y] \end{array} \tag{30}$$

The two expressions in Eq.(30) are nothing but the mean of the denominator and numerator of $\hat{\mu}(y)$, respectively. We thus rewrite the term $\hat{\mu}(y)$ as

$$\hat{\mu}(y) = \frac{p_\sigma(y)\mathbb{E}[x_c|y]\left(1 + \frac{1}{N}\sum \frac{p(y|x_i)x_{i,c} - p_\sigma(y)\mathbb{E}[x_c|y]}{p_\sigma(y)\mathbb{E}[x_c|y]}\right)}{p_\sigma(y)\left(1 + \frac{1}{N}\sum \frac{p(y|x_i) - p_\sigma(y)}{p_\sigma(y)}\right)} \tag{31}$$

Next, we assume $N \gg 1$ and that the patch y is not too rare, such that

$$\frac{1}{N}\sum \frac{p(y|x_i) - p_\sigma(y)}{p_\sigma(y)} \ll 1$$

Then, using a Taylor expansion for small ϵ,

$$\frac{1}{1+\epsilon} = 1 - \epsilon + \epsilon^2 + O(\epsilon^3)$$

we obtain the following asymptotic expansion for $\hat{\mu}(y)$,

$$\begin{aligned}\hat{\mu}(y) \approx \mathbb{E}[x_c|y] &\left(1 + \frac{1}{N}\sum_i \frac{p(y|x_i)x_{i,c} - p(y)\mathbb{E}[x_c|y]}{p(y)E[x_c|y]}\right) \\ \cdot &\left(1 - \frac{1}{N}\sum_i \frac{p(y|x_i) - p(y)}{p(y)} + \left(\frac{1}{N}\sum_i \frac{p(y|x_i) - p(y)}{p(y)}\right)^2\right)\end{aligned} \tag{32}$$

We now take the expectation of Eq. (32) over x samples. We use the fact that $\mathbb{E}[p(y|x) - p(y)] = 0,\ \mathbb{E}[p(y|x)x_c - p(y)\mathbb{E}[x_c|y]] = 0$. We also neglect all $O(1/N^2)$ terms.

$$\begin{aligned}\mathbb{E}_N[\hat{\mu}(y)] = \mathbb{E}[x_c|y] &\left(1 + \frac{1}{N}\frac{\mathbb{E}[p(y|x)^2]}{p(y)^2}\right. \\ &\left. - \frac{1}{N}\frac{E[p(y|x)^2 x_c]}{p(y)^2\mathbb{E}[x_c|y]} + o(1/N)\right)\end{aligned} \tag{33}$$

Using the identities of Eq. (30) we can express this as

$$\mathbb{E}_N[\hat{\mu}(y)] = \mathbb{E}[x_c|y] + \zeta_y(\mathbb{E}[x_c|y] - \mathbb{E}_{\sigma^*}[x_c|y]) + o(1/N) \tag{34}$$

With

$$\zeta_y = \frac{1}{N}\frac{A(\sigma,k)p_{\sigma^*}(y)}{p(y)^2} \tag{35}$$

We now move to computing $\mathbb{E}_N[\hat{\mu}(y)^2]$. Using the identities in Eq. (30), we rewrite the term $\hat{\mu}(y)^2$ as

$$\begin{aligned}\hat{\mu}(y)^2 &= \mathbb{E}[x_c|y]^2 \frac{\left(1 + \frac{1}{N}\sum \frac{p(y|x_i)x_{i,c} - p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}\right)^2}{\left(1 + \frac{1}{N}\sum \frac{p(y|x_i) - p_\sigma(y)}{p_\sigma(y)}\right)^2} \\ &= E[x_c|y]^2 \\ &\cdot \frac{\left(1 + \frac{2}{N}\sum \frac{p(y|x_i)x_{i,c} - p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]} + \left(\frac{1}{N}\sum \frac{p(y|x_i)x_{i,c} - p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}\right)^2\right)}{\left(1 + \frac{2}{N}\sum \frac{p(y|x_i) - p_\sigma(y)}{p_\sigma(y)} + \left(\frac{1}{N}\sum \frac{p(y|x_i) - p_\sigma(y)}{p_\sigma(y)}\right)^2\right)} \\ &\approx E[x_c|y]^2 \\ &\cdot \left(1 + \frac{2}{N}\sum \frac{p(y|x_i)x_{i,c} - p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]} + \left(\frac{1}{N}\sum \frac{p(y|x_i)x_{i,c} - p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}\right)^2\right) \\ &\cdot \left(1 - \frac{2}{N}\sum \frac{p(y|x_i) - p_\sigma(y)}{p_\sigma(y)} + \frac{3}{N}\left(\sum \frac{p(y|x_i) - p_\sigma(y)}{p_\sigma(y)}\right)^2\right)\end{aligned} \tag{36}$$

Taking expectations over all sequences of N samples, and omitting $O(1/N^2)$ terms, we get:

$$\begin{aligned}\mathbb{E}_N[\hat{\mu}(y)^2] = \mathbb{E}[x_c|y]^2 \Bigg(1 - \frac{4}{N}\frac{\mathbb{E}[p(y|x)^2 x_c]}{p(y)^2\mathbb{E}[x_c|y]} + \frac{1}{N}\frac{\mathbb{E}[p(y|x)^2 x_c^2]}{p(y)^2\mathbb{E}[x_c|y]^2} \\ + \frac{3}{N}\frac{\mathbb{E}[p(y|x)^2]}{p(y)^2} \; + \; o(1/N) \Bigg)\end{aligned} \tag{37}$$

Using Eqs. (30) and (35) we can simplify Eq. (37) to

$$\begin{aligned}&\mathbb{E}_N[\hat{\mu}(y)^2] = \mathbb{E}[x_c|y]^2 \\ &+ \zeta_y \left(-4\mathbb{E}_{\sigma^*}[x_c|y]E[x_c|y] + \mathbb{E}_{\sigma^*}[x_c^2|y] + 3\mathbb{E}[x_c|y]^2\right) + o(1/N)\end{aligned} \tag{38}$$

We now substitute the terms from Eqs. (34) and (38) in Eq. (29), resulting in

$$\begin{aligned}MSE^{NP} \approx & \mathbb{E}[x_c^2|y] - \mathbb{E}[x_c|y]^2 + & (39)\\ & \zeta_y \left(\mathbb{E}_{\sigma^*}[x_c^2|y] + \mathbb{E}[x_c|y]^2 - 2\mathbb{E}[x_c|y]\mathbb{E}_{\sigma^*}[x_c|y]\right) \\ = & \mathbb{V}[x_c|y] + & (40)\\ & \zeta_y \left(\mathbb{V}_{\sigma^*}[x_c|y] + \left(\mathbb{E}[x_c|y] - \mathbb{E}_{\sigma^*}[x_c|y]\right)^2\right) \\ = & \mathbb{V}[x_c|y] + \frac{1}{N}\mathcal{V}(y) & (41)\end{aligned}$$

□

Claim. For a Gaussian distribution, the average non parametric variance is

$$\bar{\mathcal{V}} = \mathbb{E}_y[\mathcal{V}(y)] = \frac{|\Phi|\mathbb{V}[x_c|y]}{\sigma^{2d}} \tag{42}$$

Proof. We denote by $\Phi, \Phi^*, \Gamma, \Psi, \Psi_*$ the covariance matrices of $p_\sigma(y), p_{\sigma^*}(y), p(x), p_\sigma(x|y), p_{\sigma^*}(x|y)$ respectively.

We denote by

$$B(y) = \mathbb{V}_{\sigma^*}[x_c|y] + \left(\mathbb{E}_\sigma[x_c|y] - \mathbb{E}_{\sigma^*}[x_c|y]\right)^2 \tag{43}$$

We would like to compute

$$\begin{aligned}&\mathbb{E}_y[\mathcal{V}(y)] \\ &= \int p_\sigma(y)\frac{p_{\sigma^*}(y)}{(4\pi\sigma^2)^{d/2}p_\sigma(y)^2}B(y)dy \\ &= \int \frac{p_{\sigma^*}(y)}{(4\pi\sigma^2)^{d/2}p_\sigma(y)}B(y)dy \\ &= \frac{1}{(4\pi\sigma^2)^{d/2}}\frac{(2\pi)^{d/2}|\Phi|^{1/2}}{(2\pi)^{d/2}|\Phi^*|^{1/2}}\int e^{-\frac{1}{2}y^T(\Phi^{*-1}-\Phi^{-1})y}B(y)dy\end{aligned} \tag{44}$$

Denoting:

$$\Theta^{-1} = \Phi^{*-1} - \Phi^{-1} \tag{45}$$

$$Z^{-1} = \frac{|\Phi|^{1/2}|\Theta|^{1/2}}{(2\sigma^2)^{d/2}|\Phi^*|^{1/2}} \tag{46}$$

We can write

$$\begin{aligned}&\mathbb{E}_y[\mathcal{V}(y)]\\&= \frac{1}{Z}\frac{1}{(2\pi)^{d/2}|\Theta|^{1/2}}\int e^{-\frac{1}{2}y^T\Theta^{-1}y}B(y)dy\\&= Z^{-1}\mathbb{E}_\Theta[B(y)]\end{aligned} \tag{47}$$

We now note that for Gaussian distributions we can express the relation between the various covariance matrices as

$$\begin{aligned}\Phi &= \Gamma + \sigma^2\mathbf{I}_d\\ \Phi^* &= \Gamma + \sigma^{*2}\mathbf{I}_d\\ \Psi &= \left(\Gamma^{-1} + \tfrac{1}{\sigma^2}\mathbf{I}_d\right)^{-1}\\ \Psi_* &= \left(\Gamma^{-1} + \tfrac{1}{\sigma^{*2}}\mathbf{I}_d\right)^{-1}\end{aligned} \tag{48}$$

Since all this matrices are obtained from Γ or Γ^{-1} by adding a scalar matrix, they are all diagonal in the same basis. We denote by $\{\phi_\ell\}, \{\phi^*{}_\ell\}, \{\gamma_\ell\}, \{\psi_\ell\}, \{\psi_{*\ell}\}, \{\theta_\ell\}$ the eigenvalues of $\Phi, \Phi^*, \Gamma, \Psi, \Psi_*, \Theta$, and by $\{u_1, \dots u_d\}$ the joint eigenvectors basis. We can express the eigenvalues relations as

$$\begin{aligned}\phi_\ell &= \gamma_\ell + \sigma^2\\ \phi^*{}_\ell &= \gamma_\ell + \sigma^{*2}\\ \psi_\ell &= \left(\gamma_\ell^{-1} + \tfrac{1}{\sigma^2}\right)^{-1}\\ \psi_{*\ell} &= \left(\gamma_\ell^{-1} + \tfrac{1}{\sigma^{*2}}\right)^{-1}\\ \theta_\ell &= \left(\tfrac{1}{\gamma_\ell+\sigma^{*2}} - \tfrac{1}{\gamma_\ell+\sigma^2}\right)^{-1} = \tfrac{(\gamma_\ell+\sigma^{*2})(\gamma_\ell+\sigma^2)}{\sigma^{*2}}\end{aligned} \tag{49}$$

We can now express Z^{-1} as a product of eigenvalues

$$Z^{-1} = \left(\Pi_\ell \frac{\phi_\ell\theta_\ell}{2\sigma^2\phi^*{}_\ell}\right)^{1/2} \tag{50}$$

By substituting the terms from Eq. (49) in Eq. (50) we get

$$Z^{-1} = \Pi_\ell \frac{\phi_\ell}{\sigma^2} = \frac{|\Phi|}{\sigma^{2d}} \tag{51}$$

We now want to compute the term $\mathbb{E}_\Theta[B(y)]$. We denote with $\tilde{x}, \tilde{y}$ the transformation of the signals x, y to the eigenvectors basis

$$\tilde{x}_\ell = u_\ell^T x, \quad \tilde{y}_\ell = u_\ell^T y \tag{52}$$

We can express

$$\mathbb{E}_\sigma[x_c|y] = \sum_\ell u_{\ell,c}\mathbb{E}_\sigma[\tilde{x}_\ell|\tilde{y}] \tag{53}$$

$$\mathbb{E}_{\sigma^*}[x_c|y] = \sum_\ell u_{\ell,c}\mathbb{E}_{\sigma^*}[\tilde{x}_\ell|\tilde{y}] \tag{54}$$

$$\tag{55}$$

where $u_{\ell,c}$ denote the c entry of the eigenvector u_ℓ. In the diagonal basis

$$\mathbb{E}_\sigma[\tilde{x}_\ell|\tilde{y}] = \frac{\psi_\ell}{\sigma^2}\tilde{y}_\ell, \quad \mathbb{E}_{\sigma^*}[\tilde{x}_\ell|\tilde{y}] = \frac{\psi_{*\ell}}{\sigma^{*2}}\tilde{y}_\ell \tag{56}$$

If we take expectation over y, when $y \sim N(0, \Theta)$ and recall that Θ is also diagonal in the basis $\{u_\ell\}$, we get

$$E_\Theta\left[\left(\mathbb{E}_\sigma[x_c|y] - \mathbb{E}_{\sigma^*}[x_c|y]\right)^2\right] = \sum_\ell u_{\ell,c}^2 \left(\frac{\psi_\ell}{\sigma^2} - \frac{\psi_{*\ell}}{\sigma^{*2}}\right)^2 \theta_\ell \tag{57}$$

We also use the diagonal basis to express

$$\mathbb{V}_{\sigma^*}[x_c|y] = \sum_\ell u_{\ell,c}^2 \psi_{*\ell} \tag{58}$$

Thus

$$\mathbb{E}_\Theta[B(y)] = \sum_\ell u_{\ell,c}^2 \left(\left(\frac{\psi_\ell}{\sigma^2} - \frac{\psi_{*\ell}}{\sigma^{*2}}\right)^2 \theta_\ell + \psi_{*\ell}\right) \tag{59}$$

Substituting the terms from equation Eq. (49) plus some algebraic manipulations provides that

$$\left(\frac{\psi_\ell}{\sigma^2} - \frac{\psi_{*\ell}}{\sigma^{*2}}\right)^2 \theta_\ell + \psi_{*\ell} = \psi_\ell \tag{60}$$

Hence

$$\mathbb{E}_\Theta[B(y)] = \sum_\ell u_{\ell,c}^2 \psi_\ell = \mathbb{V}_\sigma[x_c|y] \tag{61}$$

Combining Eqs. (51) and (61) into Eq. (47) we get

$$\mathbb{E}_y[\mathcal{V}(y)] = \frac{|\Phi|\mathbb{V}[x_c|y]}{\sigma^{2d}} \tag{62}$$

□