Natural Image Denoising: Optimality and Inherent Bounds

Anat Levin and Boaz Nadler
Department of Computer Science and Applied Math
The Weizmann Institute of Science

Abstract

The goal of natural image denoising is to estimate a clean version of a given noisy image, utilizing prior knowledge on the statistics of natural images. The problem has been studied intensively with considerable progress made in recent years. However, it seems that image denoising algorithms are starting to converge and recent algorithms improve over previous ones by only fractional dB values. It is thus important to understand how much more can we still improve natural image denoising algorithms and what are the inherent limits imposed by the actual statistics of the data. The challenge in evaluating such limits is that constructing proper models of natural image statistics is a long standing and yet unsolved problem.

To overcome the absence of accurate image priors, this paper takes a non parametric approach and represents the distribution of natural images using a huge set of 10^{10} *patches. We then derive a simple statistical measure which provides a lower bound on the optimal Bayesian minimum mean square error (MMSE). This imposes a limit on the best possible results of denoising algorithms which utilize a fixed support around a denoised pixel and a generic natural image prior. Our findings suggest that for small windows, state of the art denoising algorithms are approaching optimality and cannot be further improved beyond* $\sim 0.1dB$ *values.*

1. Introduction

Natural image denoising is defined as the problem of estimating a clean version of a noise corrupted image, given a-priori knowledge that the original unknown signal is a natural image. The main idea in this setting is that image denoising can be performed using not only the noisy image itself but rather using a suitable prior on natural image statistics. Image denoising algorithms have drastically advanced over the last few decades [17, 7, 20, 21, 3, 10, 8]. Various image priors have been learned [20, 21, 26], and particularly impressing results have been obtained with non parametric techniques such as non local means [3] or BM3D [8], and sparse representation methods such as KSVD [10]. However, it seems that the performance of denoising algorithms is starting to converge. Recent techniques typically improve over previous ones by only fractional dB values. In some cases the difference between the results of competing algorithms is so small and inconclusive, that one actually has to successively toggle between images on a monitor to visually compare their denoising quality. This raises the question of whether the error rates of current denoising algorithms can be reduced much further, or whether there are inherent limitations imposed by the statistical structure of natural images? The goal of this paper is to derive a *lower bound* on the best possible denoising error under a well defined statistical framework. Such a bound can help us understand if there is hope to significantly improve the current state-of-the-art image denoising with even better algorithms, or whether we have nearly approached the fundamental limits.

Understanding the limits of natural image denoising is also important as an instance of a more fundamental computer and human vision challenge: modeling the statistics of natural images and understanding the inherent limits of their statistical power. Several works attempted to estimate the entropy of natural images [15, 4]. However, there is no direct relation between entropy and our ability to solve low level vision tasks. Furthermore, numerous research attempts have been devoted to the learning of natural image priors [27, 2, 21, 26, 18, 12]. However, it is not clear if these models actually capture the full statistical structure of natural images. In principle, a perfectly accurate natural image prior would allow us to compute optimal Bayesian estimators for low level vision tasks such as denoising, super-resolution, deconvolution and others. Therefore, understanding how far from optimality do current denoising algorithms stand, will give us an indication of how far are we from understanding and modeling the statistical structure of natural images.

Several works studied the limits of image denoising. Some methods focused mostly on SNR arguments [13, 24, 23] without taking into account the strength of natural image priors. Similarly, limits on the related problem of super-resolution have been derived in the past [1, 16], but they again account for the numerical stability of the linear system being inverted, and do not model the additional gain provided by natural image priors. In fact, in the same paper, Baker and Kanade [1] show that their super-resolution limits can be broken using a face class prior. In the statistical literature, sharp bounds on image denoising and edge detection have been obtained. However, these bounds assume over-simplified image models, such as piecewise constant regions separated by sharp edges with smooth boundaries [19, 14]. The non local means denoising algorithm [3] is proven to be asymptotically optimal given an infinitely large image. However, its expected error and deviation from optimality on realistic finite-size images is unclear. Recently, Chatterjee and Milanfar [5, 6] derived denoising bounds that do account for natural image statistics. Their model assumes that the image patches can be factorized into a small number of clusters, and each such cluster can be described using second order statistics. This, however, is a strong assumption which can largely affect the conclusions.

The lack of a more detailed analysis which accounts for the statistics of natural images is due to the fact that modeling the statistics of natural images is an extremely hard problem, with no agreeably good model suggested up to this day. Thus, any evaluation which would rely on one of the existing image priors will be strongly biased by the inaccuracy of the prior. In order to overcome this challenge, our aim in this paper is to estimate a lower bound on image denoising *without using any particular parametric model.*

Using the sum of square differences error metric, we derive a lower bound on the smallest possible reconstruction error of denoising algorithms which consider a $k \times k$ pixels support around a noisy pixel, and use a *generic* natural image prior on $k \times k$ patches. Thus, the possible gain by image specific priors [3, 10, 8], rather than generic ones, is beyond the scope of this analysis.

Our approach builds on the success of recent large image databases approaches in high level vision and graphics applications [11, 22], and represents the distribution of natural images in a non-parametric way using a huge set of 10^{10} patches. We use this patch set to approximate the optimal Bayesian minimum mean squared error (MMSE) estimator and its expected error. While every finite set of samples provides only an approximation for the actual MMSE, we derive statistical formulas which, under certain conditions, allow us to compute both a lower bound and an upper bound on the MMSE. We show that for small support size k or for large noise variance, the number of patches (10^{10}) is large enough. Hence, the lower and upper bounds coincide and we can estimate the actual MMSE. At the more challenging cases of very large patch sizes or very small noise levels, we only get a lower bound on the best possible denoising error.

Our calculations suggest that for the tested support sizes, the state of the art denoising results of BM3D [8] are already close to optimality, and cannot be further improved beyond 0.1 dB values. On the other hand, increasing the support size does carry some potential for improved denoising performance. However, this increase should probably require switching to parametric approaches, since non-parametric approaches are inherently limited by the density of nearest neighbors, which drastically decreases with window size. Developing parametric models which can achieve state of the art denoising results is also useful since they can be applied to other low-level vision tasks, and more importantly, because their generalization power can improve our understanding of natural images.

2. Bounding denoising performance

2.1. Problem formulation

In image denoising, one is given a noisy version y of a clean image x

$$y = x + n \tag{1}$$

and the aim is to estimate a cleaner version $\hat{y}$ from y. In this paper, we consider random noise vectors n whose entries are distributed according to a zero mean i.i.d. Gaussian with variance σ^2.[1] We restrict the discussion to denoising algorithms which denoise each pixel separately using a $k \times k$ pixels support around it. Equivalently, we assume that x, y are $k \times k$ patches and focus on the estimation of the central pixel in each patch x, which we denote by x_c.

The success of image denoising algorithms is usually evaluated by PSNR values, which essentially measure the mean squared error (MSE) in reconstruction, averaged over a set of M test patches $\{(x_j, y_j)\}_{j=1}^{M}$

$$\text{PSNR} = 10 \log_{10}\left(\frac{1}{\text{MSE}}\right), \quad \widehat{\text{MSE}} \frac{1}{M} \sum_j (x_{j,c} - \hat{y}_{j,c})^2. \tag{2}$$

As discussed in [25], the MSE may not provide an ultimate visual quality prediction. Nonetheless, we adopt the MSE for two reasons: a) it is the classical measure being optimized by most existing denoising algorithms, and b) it is correlated, though not perfectly, with visual quality. Therefore, a limit on MSE denoising is a reasonable proxy on how much we can improve visual quality. The study of lower limits corresponding to other quality measures is an interesting future research problem.

Since each noisy patch y could actually be generated by multiple latent patches x, the problem is fundamentally ambiguous and one cannot hope for a zero reconstruction error. Our goal is to understand what is the lowest possible MSE achievable by any denoising algorithm based on $k \times k$ patches, given a-priori knowledge that the input image has been randomly sampled from the set of natural images.

To this end, we denote by $p(x)$ the density of $k \times k$ patches of natural images, and by $p(y)$ the resulting density of $k \times k$ noisy patches. When context requires, we use the notation $p_\sigma(y)$ to emphasize the explicit dependence of the density of noisy y patches on the noise level σ.

[1] The Gaussian i.i.d. noise assumption is not a hard one and the technique can be applied with many other noise models. The non parametric approach mostly requires the ability to compute $p(y|x)$.

2.2. Bayesian Minimum Mean Square Error

There are two equivalent interpretations for the MSE – the error interpretation and the variance interpretation. In the error view, we randomly sample clean patches x_i from $p(x)$, add noise to generate y_i, denoise y_i and measure the reconstruction error $(x_{i,c} - \hat{y}_{i,c})^2$. The average of this reconstruction error is

$$\text{MSE} = \int p(x) \int p(y|x)(x_c - \hat{y}_c)^2 dy dx \tag{3}$$

An equivalent interpretation of the MSE, which will serve us below, is the variance interpretation– start from a noisy patch y, and measure the variance of $p(x|y)$ around it. That is, compute the sum of weighted distances between $\hat{y}$ and all possible x explanations:

$$\text{MSE} = \int p(y) \int p(x|y)(x_c - \hat{y}_c)^2 dx dy \tag{4}$$

Eq. (4) is obtained from Eq. (3) by switching the order of integration and applying Bayes rule. The advantage of the variance view is that from it, one can easily derive the optimal estimator, namely the Bayesian minimum mean squared error (MMSE) estimator [13], which is simply given by the conditional mean

$$\begin{aligned} \mu(y) = \mathbb{E}[x_c|y] &= \int p(x|y) x_c dx \qquad (5) \\ &= \int \frac{p(y|x)}{p(y)} p(x) x_c dx \end{aligned}$$

The MMSE at any fixed y is then the conditional variance,

$$\mathcal{V}(y) = \mathbb{E}[(x_c - \mu(y))^2 | y] = \int p(x|y) \left(x_c - \mu(y)\right)^2 dx \tag{6}$$

whereas the overall MMSE is

$$\text{MMSE} \;=\; \mathbb{E}[\mathcal{V}(y)] = \int p(y) \mathcal{V}(y) dy. \tag{7}$$

In the framework of natural image denoising, the MMSE in (7) is the lowest achievable denoising error by any denoising algorithm. To compute this MMSE and the corresponding optimal estimator $\mu(y)$, we need access to the true natural image density $p(x)$. However, as discussed in the introduction, an accurate model for this density is a long standing research problem [20, 21, 26, 9]. One option would be to fit some parametric model $q(x)$, for example, a field of experts model [21] or a Gaussian mixture model. We can then compute an approximated mean $\hat{\mu}(y) = 1/q(y) \int p(y|x) q(x) x_c dx$. However, since $q(x)$ is only an approximation to the true natural image density $p(x)$, the error of the estimator $\hat{\mu}(y)$, like the error of any denoising algorithm, is only an upper bound on the MMSE. The difference in MSE depends on the quality of the approximation, which is not easy to evaluate, in particular since $p(x)$ is unknown. Our goal, in contrast, is to estimate the MMSE without committing to any specific approximate parametric model, or essentially, without even explicitly knowing $p(x)$. The key idea is that even though $p(x)$ is unknown, we are still able to sample from it. Therefore we consider a non-parametric representation of the distribution, using a large set of about $N = 10^{10}$ natural image patches. This allows us to approximate the integral of Eq. (5) by averaging over samples:

$$\hat{\mu}(y) = \frac{\frac{1}{N} \sum_i p(y|x_i) x_{i,c}}{\frac{1}{N} \sum_i p(y|x_i)}, \tag{8}$$

where for Gaussian noise

$$p(y|x) = \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{\|x-y\|^2}{2\sigma^2}} \tag{9}$$

and $d = k^2$. As $N \to \infty$, $\hat{\mu}(y)$ indeed converges to the true MMSE estimator. However, for any finite set, Eq. (8) is only an approximation, and thus, like any other denoising algorithm, its average error provides an upper bound on the MMSE. However, our goal is to bound the MMSE from below, still, without knowing $p(x)$ explicitly, but having at our disposal $N \gg 1$ samples from it. The main idea in deriving both an upper and a lower bound on the MMSE, is to use the two MSE formulations in Eqs. (3) and (4). Given a set of M clean and noisy pairs $\{(\tilde{x}_j, y_j)\}_{j=1}^{M}$ and another independent set of N clean patches $\{x_i\}_{i=1}^{N}$, both randomly sampled from natural images, we compute

$$\begin{aligned} \text{MMSE}^U &= \frac{1}{M} \sum_j (\hat{\mu}(y_j) - \tilde{x}_{j,c})^2 \qquad (10) \\ \text{MMSE}^L &= \frac{1}{M} \sum_j \hat{\mathcal{V}}(y_j) \qquad (11) \end{aligned}$$

where $\hat{\mathcal{V}}(y_j)$ is the approximated variance:

$$\hat{\mathcal{V}}(y_j) = \frac{\frac{1}{N} \sum_i p(y_j|x_i)(\hat{\mu}(y_j) - x_{i,c})^2}{\frac{1}{N} \sum_i p(y_j|x_i)} \tag{12}$$

MMSE^U and MMSE^L are both random variables which depend on the particular set of x, y samples. When the sample size approaches infinity, they converge to the exact MMSE. However, we show that for a finite sample, in expectation, MMSE^U and MMSE^L provide *upper* and *lower bounds* on the best possible MMSE. Note the key difference between these two quantities: MMSE^U uses explicit knowledge of the original noise-free patch $\tilde{x}_j$, while MMSE^L does not involve it. Since MMSE^U basically measures the error of the estimator $\hat{\mu}(y_j)$, it provides, like every denoising algorithm, an upper bound on the MMSE. The term MMSE^L is analyzed in Sec. 2.3.

When N is sufficiently large, the two computed values MMSE^U and MMSE^L are similar, and we get an accurate estimate for the actual optimal MMSE. As we show in section 3 this happens when the patch size k is small, or when the noise variance σ^2 is high, since in such cases there is a large number of valid nearest neighbors around each patch. In the harder cases, we cannot compute the MMSE exactly, but MMSE^L still provides a lower bound on the best denoising results we can expect.

2.3. The sample variance as a lower bound on the MMSE

Our goal is to show that the approximated variance estimated from a finite set of samples (Eq. (12)), provides a lower bound on the correct variance (Eq. (6)). To see this at an intuitive level, consider the numerator of Eq. (12). Since $\hat{\mu}(y)$ is the weighted mean of $\{x_i\}_{i=1}^N$, it minimizes the mean squared distance from these patches, and in particular, it achieves a smaller squared distance compared to the exact unknown mean $\mu(y)$:

$$\sum_i p(y_j|x_i)(x_{i,c} - \hat{\mu}(y_j))^2 \leq \sum_i p(y_j|x_i)(x_{i,c} - \mu(y_j))^2$$

Thus, in expectation

$$\begin{aligned} &\mathbb{E}_N\left[\frac{1}{N}\sum_i p(y_j|x_i)(x_{i,c} - \hat{\mu}(y_j))^2\right] \leq \\ &\mathbb{E}_N\left[\frac{1}{N}\sum_i p(y_j|x_i)(x_{i,c} - \mu(y_j))^2\right] = p(y_j)\mathcal{V}(y_j), \end{aligned} \tag{13}$$

where $\mathbb{E}_N[\cdot]$ denotes expectation with respect to all possible sets of N i.i.d. patches from $p(x)$. The denominator of Eq. (12) is also a random variable whose expectation is $p(y_j)$, thus, roughly, $\mathbb{E}_N[\hat{\mathcal{V}}(y)] \leq \mathcal{V}(y)$. However, since both numerator and denominator are random variables, a more careful analysis of $\hat{\mathcal{V}}(y)$ is needed.

The following claim derives a second order approximation to $\mathbb{E}_N[\hat{\mathcal{V}}(y)]$ for a fixed patch y.

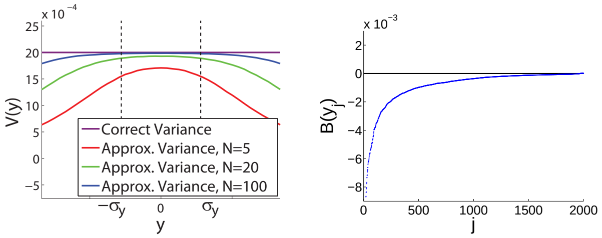
Claim 1 *Asymptotically, as the training set size $N \to \infty$, the expected value of the approximated variance defined in Eq. (12) is*

$$\mathbb{E}_N[\hat{\mathcal{V}}(y)] = \mathcal{V}(y) + C(y)\mathcal{B}(y) + o\left(\frac{1}{N}\right), \tag{14}$$

with

$$\begin{aligned} \mathcal{B}(y) &= \mathbb{E}_\sigma[x_c^2|y] - 3\mathbb{E}_\sigma[x_c|y]^2 - 2\mathbb{E}_{\sigma^*}[x_c^2|y] \\ &\quad +4\mathbb{E}_\sigma[x_c|y]\mathbb{E}_{\sigma^*}[x_c|y] \qquad (15) \\ C(y) &= \frac{1}{N}\frac{p_{\sigma^*}(y)}{(4\pi\sigma^2)^{d/2}p_\sigma(y)^2} \qquad (16) \end{aligned}$$

where p_σ, $\mathbb{E}_\sigma[\cdot]$, p_{σ^*}, $\mathbb{E}_{\sigma^*}[\cdot]$ denote probability and expectation of random variables with noise standard derivation σ and σ^* respectively. σ is the actual standard deviation and $\sigma^* = \sigma/\sqrt{2}$.



(a) (b)

Figure 1. Evaluating the bias of $\hat{\mathcal{V}}(y)$. (a) For a 1D Gaussian density $p(x)$, we compare the expected sample variance $\mathbb{E}_N[\hat{\mathcal{V}}(y)]$ to its exact value $\mathcal{V}(y)$, which is constant in y for a Gaussian distribution. As predicted theoretically, $\mathbb{E}_N[\hat{\mathcal{V}}(y)] \leq \mathcal{V}(y)$, with the bias decreasing with increasing number of samples or at high density regions. (b) Plug-in bias estimates $\hat{\mathcal{B}}(y)$ of real natural image patches, sorted in ascending order. Note that up to small numerical errors, the bias estimates of all patches are negative.

The proof of Claim 1 is provided in the appendix. It uses an asymptotic expansion of $\hat{\mathcal{V}}(y)$ in N, and neglects high order terms. It is hence valid for sufficiently large N and sufficiently common y.

Next, we consider a local Gaussian approximation for $p(x)$ around $x = y$, e.g., a Laplace approximation consisting of a second order Taylor expansion of $\ln p(x)$. For a Gaussian distribution we can analytically compute the leading order bias term $C(y)\mathcal{B}(y)$ from Eqs. (15) and (16). The following claim, proven in the appendix, shows that $\mathcal{B}(y)$ is negative for all y values. Since $C(y) > 0$, it follows that $\mathbb{E}_N[\hat{\mathcal{V}}(y)] \leq \mathcal{V}(y)$. Thus, in expectation, MMSE^L provides a lower bound on the MMSE, as we aim to show.

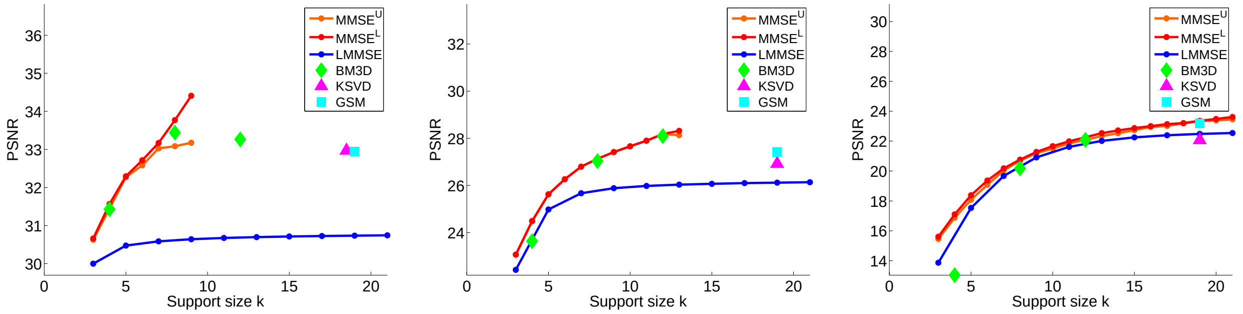
Claim 2 *For a Gaussian distribution, $\mathcal{B}(y) \leq 0$ for all y.*

Figure 1(a) considers a simple case of a 1D Gaussian distribution and compares the exact variance $\mathcal{V}(y)$ with $\mathbb{E}_N[\hat{\mathcal{V}}(y)]$, averaged over $10,000$ realizations of sets of $N = 5, 10$ or 15 samples. The bias is always negative and $\mathbb{E}_N[\hat{\mathcal{V}}(y)] \leq \mathcal{V}(y)$. As evident from Eq. (16), $C(y)$ is large when $p(y)$ is small, and so for a Gaussian distribution, the bias increases exponentially with $|y|$.

The density of natural image patches is not Gaussian, so a Laplace approximation might be inaccurate and Claim 2 may not directly apply. To evaluate this on real data, we used a plug-in estimator $\hat{\mathcal{B}}(y)$ for the term $\mathcal{B}(y)$. That is, we considered $M = 2,000$ noisy 3×3 patches $\{y_j\}$ with noise variance $\sigma = 18$. For each y_j the expectations in Eq. (15) were estimated by averaging over an independent large set of clean patches $\{x_i\}$. Figure 1(b) shows the resulting $\hat{\mathcal{B}}(y_j)$ values, sorted in ascending order for visualization purposes. As seen in the plot, the estimated values $\hat{\mathcal{B}}(y_j)$ are all negative up to small numerical errors.

3. Experiments

To evaluate the MMSE we use a set of $20,000$ images from the LabelMe dataset [22]. We thus implicitly consider this dataset an unbiased representative of noise-free



(a) $\sigma = 18$ (b) $\sigma = 55$ (c) $\sigma = 170$

Figure 2. PSNR values of several recent denoising algorithms along with our MMSE lower and upper bounds. As predicted by the theory, the performance of all algorithms are bounded by our MMSE^L estimate, although BM3D approaches the bound by fractional dB values. (Note that since PSNR $= -10 \log 10(\text{MSE})$, the MMSE lower bound turns into an upper bound on the best achievable PSNR).

natural image statistics.[2] To avoid quantization and JPEG artifacts we first low-passed all images and down-sampled them by a factor of two. These images provide a set of about $N = 10^{10}$ natural image patches $\{x_i\}$. We use another $M = 2,000$ clean and noisy pairs of patches $\{(\tilde{x_j}, y_j)\}$, denoise them by computing the weighted mean of Eq. (8), and measure the mean squared error MMSE^U (Eq. (10)), and the lower bound MMSE^L (Eq. (11)). This is an intensive computation, which took about a week of computation on a 100 CPUs cluster.

In Figure 2 we compare the PSNR values of various denoising algorithms to MMSE^U and MMSE^L with three noise levels $\sigma = 18, 55, 170$, for image intensities in the $[0, 255]$ range. We plot our bounds as a function of the window size k. We tried to match the support size of the tested algorithms as well. We used the BM3D [8] algorithm with patch sizes of $4, 8$ and 12 pixels (for $8, 12$ pixels we used the authors' parameters, and for 4 pixels, our adaptation). We associate the KSVD [10] and Portilla *et al*. GSM[20] algorithms with the largest support size k considered, since these are global denoising algorisms that do not consider finite support patches. As a baseline comparison we also present a linear minimum mean square error (LMMSE) estimator [13]. This estimator, also known as the Wiener filter, uses only the second order statistics of the data, by fitting a single k^2 dimensional Gaussian to the set of N image patches. When computing our MMSE^U and MMSE^L scores we used the bootstrapping method to evaluate the variance, by drawing multiple N patch subsets. The standard deviation of the estimation is rather small, ranging from 0.05dB for small k values to 0.2dB at the large ones.

The results in Figure 2 suggest several observations which we discuss below.

Tightness of bounds: For small window sizes or high noise, the upper and lower bounds coincide and hence we have obtained an accurate estimate of the exact optimal MMSE value. For harder cases, there is a gap between the upper and lower bounds. While we are unable to estimate the exact MMSE, we still get a valid lower bound on it. This happens when there are too few samples at a normalized distance of σ^2 from the noisy patch. To demonstrate that, we measured the number of nearest neighbors within normalized distance of one standard deviation from each noisy patch. Under the assumption of Gaussian noise the difference $\|y - x\|^2$ follows a χ^2 distribution with k degrees of freedom. Therefore, for each noisy patch y_j we consider all patches at a squared distance of $\sigma^2(1 + \sqrt{2}/k^2)$:

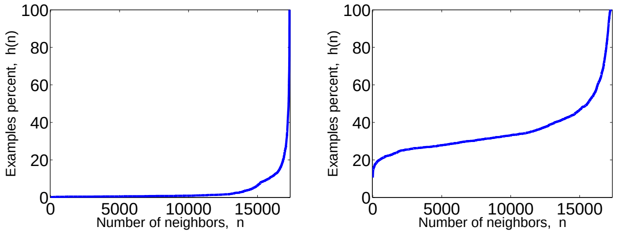
$$n_j = \# \left\{ x_i \,\middle|\, \frac{1}{k^2}\|x_i - y_j\|^2 < \sigma^2 \left(1 + \frac{\sqrt{2}}{k}\right) \right\} \tag{17}$$

In Figure 3 we plot the cumulative distribution of n_j values, for two window sizes $k = 3$ and $k = 9$, both with $\sigma = 18$. We see that for small k, most patches have a large number of nearest neighbors, whereas for $k = 9$, 13% of the patches have zero neighbors. Accordingly, at $k = 9$, $\sigma = 18$, there is a gap between MMSE^U and MMSE^L in Fig 2(a).

Near optimality of state of the art denoising results: As predicted by the theory, the empirical errors of all denoising algorithms are larger than our lower bound. Nonetheless, for most cases the PSNR values of BM3D are within 0.1dB of the optimal ones, suggesting that the BM3D results are quite close to optimality, for small patch sizes. Thus, future denoising algorithms that use small patches have very little room for improvement over BM3D[3]. However, while 0.1dB seems like a very small improvement, it may still be visually noticeable.

[2] Even if these input images do contain a small amount of noise, it is negligible w.r.t. the noise level added in our experiments. If nonetheless, these patches do contain a small amount of noise, then the goal is formulated as the reconstruction of signals with this slightly perturbed distribution, which is still a statistically well defined problem.

[3] This does not directly imply that BM3D cannot be further improved, since while BM3D is a patch based algorithm, it utilizes a wider support by looking for neighbors in the entire image.



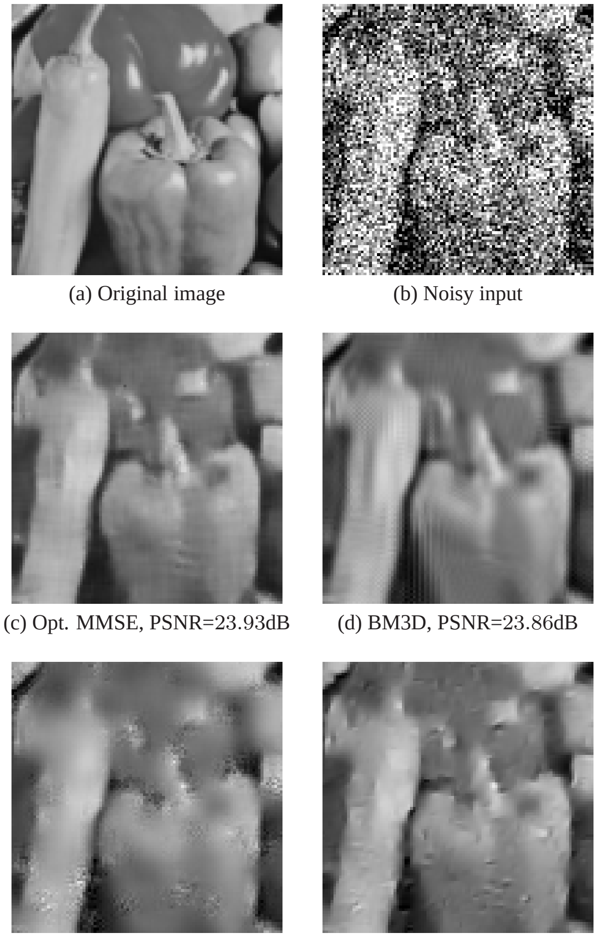
(a) $k = 3$ (b) $k = 9$

Figure 3. Cumulative distribution function of the number of neighbors within distance $\sigma^2(1 + \sqrt{2}/k)$, computed at $\sigma = 18$. For 3×3 patches, 99% of the examples had more than $2,000$ neighbors. In contrast, for a 9×9 patch size, 13% of the examples had no neighbors within this distance.

Support size: It seems that increasing the support size carries some potential for increasing the PSNR. Due to the curse of dimensionality, non-parametric techniques suffer from an unavoidable tradeoff between patch size and the number of valid nearest neighbors, and hence are unlikely to be applicable for large window sizes. Developing algorithms which utilize a larger support size thus probably requires switching from non-parametric approaches to parametric ones. Unfortunately, at the moment parametric denoising algorithms are far behind the non parametric ones.

Denoising at extreme noise levels: In Fig 2(c), the noise level is $\sigma = 170$. At this high noise level, the image prior results are very similar to the results of a simple linear minimum mean square error (LMMSE) estimator which utilizes only the second order statistics of the data. This happens since at high noise levels the signal content is lost and all the prior can do is to estimate a flat image. The square error of such an estimator is proportional to the variance of the data, therefore natural image priors or Gaussian priors utilizing second order statistics give similar results. The other extreme that we did not evaluate here is very low noise. For low noise the effect of the prior is small and the error is proportional to the noise variance. Therefore, natural image priors have an interesting effect only at medium noise levels.

Visual results: In Figure 4 we used the *Peppers* image at half resolution to visualize the denoising results of several recent denoising algorithms compared with our approach, which is the best possible denoising with a generic natural images prior. We used $\sigma = 75$ and 12×12 windows for our approach and for BM3D. Numerically, our result outperforms BM3D by 0.07dB, and this difference leads to a slightly better visual quality. The visualization of optimal denoising in Fig 4(c) is only a proof-of-concept, but not a practical denoising algorithm- denoising this 100×100 image required two weeks of computation on 100 CPUs. However, one can think of several ways to accelerate the search with approximated nearest neighbor techniques and a smarter encoding of the large set of 10^{10} image patches.



(a) Original image (b) Noisy input

(c) Opt. MMSE, PSNR=23.93dB (d) BM3D, PSNR=23.86dB

(d) KSVD, PSNR=22.41dB (e) GSM, PSNR=23.28dB

Figure 4. Visual comparison of our optimal MMSE and other algorithms, for $\sigma = 75$. The optimal MMSE and BM3D use 12×12 windows. Our approach achieves slightly higher PSNR comparing to BM3D, and a somewhat better visual quality.

Image specific bounds: Our analysis is based on the knowledge of a generic image prior. An interesting question for future research, as recently considered by [5, 6], is whether there is a significant advantage in adapting the prior to the specific statistics of the observed image, as done by some recent denoising algorithms [3, 10, 8]. In fact, such algorithms can be thought of as instances of a generic prior whose support size is the entire image, since the support size essentially means that an algorithm can see and use in the estimation all pixels within a $k \times k$ window around a noisy pixel. Unfortunately, we cannot evaluate tight bounds at very large window sizes.

4. Discussion

This paper derived a statistical measure for the best possible denoising results utilizing a generic natural image prior. Our findings suggest that for small windows, state of the art denoising algorithms cannot be further improved beyond fractional dB values. Considering a wider support carries some potential for improved results. However, increasing the support size of non-parametric algorithms might lead to a dead-end, and parametric approaches are required. Unfortunately, at the moment parametric algorithms perform well behind non-parametric ones. Developing parametric algorithms which can achieve state of the art results is also important because they can generalize for other low-level vision tasks.

The evaluation methodology used here can be easily applied to other low-level vision problems such as super-resolution, deconvolution, or inpainting. However, some of these problems, such as the deconvolution of a wide support kernel, rely on a large support of pixels. Hence we may not be able to find a sufficient number of nearest neighbors for a tight lower bound estimation. Yet, understanding the limits of image denoising is important as a step toward other low-level vision problems, and moreover, as a step toward an understanding of the inherent limits of natural image statistics.

Acknowledgments: The authors thank the ISF, BSF and ERC for providing financial support.

5. Appendix

Proof of Claim 1: For future use we denote by $A(\sigma,k) = (4\pi\sigma^2)^{-k^2/2}$ and recall that $\sigma^* = \sigma/\sqrt{2}$. The first step in the proof is to insert Eq. (8) into Eq. (12), which yields the following more convenient expression for $\hat{\mathcal{V}}(y)$,

$$\begin{aligned}\hat{\mathcal{V}}(y) &= \frac{\frac{1}{N}\sum p(y|x_i)x_{i,c}^2}{\frac{1}{N}\sum p(y|x_i)} - \frac{(\frac{1}{N}\sum p(y|x_i)x_{i,c})^2}{(\frac{1}{N}\sum p(y|x_i))^2} \\ &= A_1 - A_2. \end{aligned} \tag{18}$$

The main idea is to analyze the bias of each of these terms separately, whereby for each term we further consider the mean and fluctuations of their numerator and denominator.

For the term A_1 we shall use the following equalities

$$\begin{aligned}\mathbb{E}[p(y|x)] &= \textstyle\int p(x)p(y|x)dx = p_\sigma(y) \\ \mathbb{E}[p(y|x)x_c^2] &= p_\sigma(y)\mathbb{E}_\sigma[x_c^2|y]\end{aligned} \tag{19}$$

where $p_\sigma(y)$ is the density of y-patches at noise level σ, and $\mathbb{E}_\sigma[\cdot|y]$ denotes expectation with respect to $p_\sigma(y)$.

The two expressions in Eq.(19) are nothing but the mean of the denominator and numerator of A_1, respectively. We thus rewrite the term A_1 as

$$A_1 = \frac{p_\sigma(y)\mathbb{E}[x_c^2|y]\left(1+\frac{1}{N}\sum\frac{p(y|x_i)x_{i,c}^2-p_\sigma(y)\mathbb{E}[x_c^2|y]}{p_\sigma(y)\mathbb{E}[x_c^2|y]}\right)}{p_\sigma(y)\left(1+\frac{1}{N}\sum\frac{p(y|x_i)-p_\sigma(y)}{p_\sigma(y)}\right)} \tag{20}$$

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 the first term, (where for ease of notation we write $p(y)$ for $p_\sigma(y)$)

$$\begin{aligned} A_1 \approx \mathbb{E}[x_c^2|y]&\left(1+\frac{1}{N}\sum_i\frac{p(y|x_i)x_{i,c}^2-p(y)\mathbb{E}[x_c^2|y]}{p(y)E[x_c^2|y]}\right) \\ \cdot&\left(1-\frac{1}{N}\sum_i\frac{p(y|x_i)-p(y)}{p(y)}+\left(\frac{1}{N}\sum_i\frac{p(y|x_i)-p(y)}{p(y)}\right)^2\right)\end{aligned} \tag{21}$$

We now take the expectation of Eq. (21) over x samples. We use the fact that $\mathbb{E}[p(y|x)-p(y)] = 0$, $\mathbb{E}[p(y|x)x_c^2 - p(y)\mathbb{E}[x_c^2|y]] = 0$. We also neglect all $O(1/N^2)$ terms.

$$\begin{aligned}\mathbb{E}_N[A_1] = \mathbb{E}[x_c^2|y]&\left(1+\frac{1}{N}\frac{\mathbb{E}[p(y|x)^2]}{p(y)^2}\right. \\ &\left.-\frac{1}{N}\frac{E[p(y|x)^2x_c^2]}{p(y)^2\mathbb{E}[x_c^2|y]}+o(1/N)\right)\end{aligned} \tag{22}$$

Similarly, for the second term A_2 we use the fact that

$$\begin{aligned}\mathbb{E}[p(y|x)^2] &= \textstyle\int p(x)\frac{e^{-\|x-y\|^2/\sigma^2}}{(2\pi\sigma^2)^{k^2}}dx = A(\sigma,k)p_{\sigma*}(y) \\ \mathbb{E}[p(y|x)^2x_c] &= A(\sigma,k)p_{\sigma*}(y)\mathbb{E}_{\sigma^*}[x_c|y] \\ \mathbb{E}[p(y|x)^2x_c^2] &= A(\sigma,k)p_{\sigma*}(y)\mathbb{E}_{\sigma^*}[x_c^2|y]\end{aligned} \tag{23}$$

and hence rewrite the second term A_2 as

$$\begin{aligned} A_2 = \hat{\mu}(y)^2 &= \mathbb{E}[x_c|y]^2\frac{\left(1+\frac{1}{N}\sum\frac{p(y|x_i)x_{i,c}-p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}\right)^2}{\left(1+\frac{1}{N}\sum\frac{p(y|x_i)-p_\sigma(y)}{p_\sigma(y)}\right)^2} \\ &= E[x_c|y]^2 \\ &\cdot\frac{\left(1+\frac{2}{N}\sum\frac{p(y|x_i)x_{i,c}-p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}+\left(\frac{1}{N}\sum\frac{p(y|x_i)x_{i,c}-p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}\right)^2\right)}{\left(1+\frac{2}{N}\sum\frac{p(y|x_i)-p_\sigma(y)}{p_\sigma(y)}+\left(\frac{1}{N}\sum\frac{p(y|x_i)-p_\sigma(y)}{p_\sigma(y)}\right)^2\right)} \\ &\approx E[x_c|y]^2 \\ &\cdot\left(1+\frac{2}{N}\sum\frac{p(y|x_i)x_{i,c}-p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}+\left(\frac{1}{N}\sum\frac{p(y|x_i)x_{i,c}-p(y)\mathbb{E}[x_c|y]}{p(y)\mathbb{E}[x_c|y]}\right)^2\right) \\ &\cdot\left(1-\frac{2}{N}\sum\frac{p(y|x_i)-p_\sigma(y)}{p_\sigma(y)}+\frac{3}{N}\left(\sum\frac{p(y|x_i)-p_\sigma(y)}{p_\sigma(y)}\right)^2\right)\end{aligned} \tag{24}$$

Taking expectations and omitting $O(1/N^2)$ terms, we get:

$$\begin{aligned}\mathbb{E}_N[A_2] = \mathbb{E}[x_c|y]^2&\left(1-\frac{4}{N}\frac{\mathbb{E}[p(y|x)^2x_c]}{p(y)^2\mathbb{E}[x_c|y]}+\frac{1}{N}\frac{\mathbb{E}[p(y|x)^2x_c^2]}{p(y)^2\mathbb{E}[x_c|y]^2}\right. \\ &\left.+\frac{3}{N}\frac{\mathbb{E}[p(y|x)^2]}{p(y)^2}\ +\ o(1/N)\right)\end{aligned} \tag{25}$$

Substituting the terms from Eq. (23) in Eqs. (22) and (25) yields the desired Eq. (14). □

Proof of Claim 2: We note that $\mathcal{B}(y)$ (Eq. (15)) can be written as:

$$\mathcal{B}(y) = \mathbb{V}_\sigma[x_c|y] - 2\mathbb{V}_{\sigma^*}[x_c|y] - 2\left(\mathbb{E}_{\sigma^*}[x_c|y] - \mathbb{E}_\sigma[x_c|y]\right)^2 \tag{26}$$

where $\mathbb{V}[x_c|y] = \mathbb{E}[x_c^2|y] - \mathbb{E}[x_c|y]^2$ denotes variance. Thus, it is sufficient to show that $\mathbb{V}_\sigma[x|y] - 2\mathbb{V}_{\sigma^*}[x|y] \leq 0$. We denote by Φ the covariance matrix of $p(x)$ and by $\Psi_\sigma, \Psi_{\sigma^*}$ the covariance matrices of $p_\sigma(x|y), p_{\sigma^*}(x|y)$:

$$\Psi_\sigma = \left(\frac{1}{\sigma^2}\mathbf{I}_d + \Phi^{-1}\right)^{-1} \qquad \Psi_{\sigma^*} = \left(\frac{1}{{\sigma^*}^2}\mathbf{I}_d + \Phi^{-1}\right)^{-1}, \tag{27}$$

where $\mathbf{I}_d$ is the d dimensional identity matrix. The variance at x_c is the (c,c) entry of these matrices: $\mathbb{V}_\sigma[x|y] = \Psi_\sigma(c,c), \mathbb{V}_{\sigma^*}[x|y] = \Psi_{\sigma^*}(c,c)$.

We denote by $\{\lambda_\ell\}, \{\gamma_\ell\}, \{\gamma_\ell^*\}$ the eigenvalues of the matrices $\Phi, \Psi_\sigma, \Psi_{\sigma^*}$ respectively. We note that all these matrices are diagonal in the same basis and we can write

$$\Psi_\sigma(c,c) = \sum_\ell u_\ell^2 \gamma_\ell, \quad \Psi_{\sigma^*}(c,c) = \sum_\ell u_\ell^2 \gamma_\ell^*, \tag{28}$$

(u_ℓ is the c entry of eigenvector ℓ). We can also relate the eigenvalues of $\Psi_\sigma, \Psi_{\sigma^*}$ to the eigenvalues of the unconditional covariance Φ:

$$\gamma_\ell = (\lambda_\ell^{-1} + \sigma^{-2})^{-1}, \quad \gamma_\ell^* = (\lambda_\ell^{-1} + (\sigma^*)^{-2})^{-1}. \tag{29}$$

A simple calculation shows that for every ℓ:

$$\gamma_\ell - 2\gamma_\ell^* = \lambda_\ell \sigma^2 \left(\frac{1}{\lambda_\ell + \sigma^2} - \frac{2}{2\lambda_\ell + \sigma^2}\right) \leq 0 \tag{30}$$

Using Eqs. (30) and (28) we conclude:

$$\Psi_\sigma(c,c) - 2\Psi_\sigma(c,c) = \sum_\ell u_\ell^2(\gamma_\ell - 2\gamma_\ell^*) \leq 0. \tag{31}$$

□

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *PAMI*, 2002.

[2] A. Bell and T. Sejnowski. The ”independent components” of natural scenes are edge filters. *Vision Res.*, 1997.

[3] A. Buades, B. Coll, and J. Morel. A review of image denoising methods, with a new one. *Multiscale Model. Simul.*, 2005.

[4] D. Chandler and D. Field. Estimates of the information content and dimensionality of natural scenes from proximity distributions. *J. Opt. Soc. Am.*, 2007.

[5] P. Chatterjee and P. Milanfar. Is denoising dead? *IEEE Trans Image Processing*, 2010.

[6] P. Chatterjee and P. Milanfar. Practical bounds on image denoising: From estimation to information. *IEEE Trans Image Processing*, 2011.

[7] R. Coifman and D. Donoho. Translation-invariant denoising. In *Wavelets and Statistics*, 1995.

[8] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans Image Processing*, 2007.

[9] J. Eichhorn, F. Sinz, and M. Bethge. Natural image coding in v1: How much use is orientation selectivity? *PLoS Comput Biol*, 2009.

[10] M. Elad and M. Aharon. Image denoising via sparse and redundant representations over learned dictionaries. *IEEE Trans Image Processing*, 2006.

[11] J. Hays and A. Efros. Scene completion using millions of photographs. *SIGGRAPH*, 2007.

[12] A. Hyvarinen, J. Hurri, and P. Hoyer. *Natural Image Statistics A probabilistic approach to early computational vision*. Springer-Verlag, 2009.

[13] Steven M. Kay. *Fundamentals of statistical signal processing: estimation theory*. Prentice-Hall, Inc., 1993.

[14] A. Korostelev and A. Tsybakov. *Minimax Theory of Image Reconstruction*. Springer-Verlag, New York, 1993.

[15] A. Lee, K. Pedersen, and D. Mumford. The nonlinear statistics of high-contrast patches in natural images. *IJCV*, 2003.

[16] Z. Lin and H. Shum. Fundamental limits of reconstruction-based superresolution algorithms under local translation. *PAMI*, 2004.

[17] S. Osher, A. Sole, and L. Vese. Image decomposition and restoration using total variation minimization and the h^{-1} norm. *Multiscale Model. Simul.*, 2003.

[18] S. Osindero and G. Hinton. Modeling image patches with a directed hierarchy of markov random fields. *NIPS*, 2007.

[19] J. Polzehl and V. Spokoiny. Image denoising: Pointwise adaptive approach. *Annals of Statistics*, 31:30–57, 2003.

[20] 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.

[21] S. Roth and M.J. Black. Fields of experts: A framework for learning image priors. In *CVPR*, 2005.

[22] B. Russell, A. Torralba, K. Murphy, and W. Freeman. Labelme: a database and web-based tool for image annotation. *IJCV*, 2008.

[23] T. Treibitz and Y. Schechner. Recovery limits in pointwise degradation. In *ICCP*, 2009.

[24] M. Unser, B. Trus, and A. Steven. A new resolution criterion based on spectral signal-to-noise ratios. *Ultramicroscopy*, 1987.

[25] 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.

[26] Y. Weiss and W. T. Freeman. What makes a good model of natural images? In *CVPR*, 2007.

[27] S. Zhu and D. Mumford. Prior learning and gibbs reaction-diffusion. *PAMI*, 1997.