NL-MEANS AND AGGREGATION PROCEDURES

J. Salmon and E. Le Pennec

Laboratoire de Probabilité et Modèles Aléatoires, CNRS-UMR 7599,
Université Paris 7- Diderot
175 rue du Chevaleret 75013 Paris

ABSTRACT

Patch based denoising methods, such as the NL-Means, have emerged recently as simple and efficient denoising methods. This paper provides a new insight on those methods by showing their connection with recent statistical aggregation techniques. Within this aggregation framework, we propose some novel patch based denoising methods. We provide some theoretical justification and then explain how to implement them with a Monte Carlo based algorithm.

Index Terms— Diffusion processes, Statistics, Monte Carlo methods, Gaussian noise, Image processing

1. INTRODUCTION

Non local patch based denoising methods, such as the NL-Means of Buades, Coll and Morel [1] and its variants [2], give some of the best results to date. Those efficient methods are based on a simple idea: look at the image as a collection of small subimages, the ”patches”, and estimate each patch as a weighted average of patches. The weights should take into account their similarities and are often chosen proportional to the exponential of the quadratic error between the patches. Those methods, and their graph variations, are often seen as a kind of smoothing on a patch manifold. However, this manifold point of view does not permit to explain mathematically their performance.

We propose to tackle this challenge through a very different angle, the angle of statistical aggregation. In statistical aggregation, one consider a collection of pre-estimators and a noisy observation, and look for an “aggregation” procedure that build a weighted average of the pre-estimators that should be as close as possible to the unknown original signal. Juditsky et al. [3] and Dalalyan and Tsybakov [4, 5] propose some aggregation rules for which they control the theoretical performance. Interestingly, choosing the patches as pre-estimators leads to a a special case of recent PAC-Bayesian methods which coincides almost exactly with the NL-Means.

In this paper, we describe how this aggregation framework applies to patch based estimators and how one can obtain novel patch based method.

2. IMAGE, PATCHES AND KERNELS

Let $I = I(i, j)$, with $1 \leq i \leq N$ and $1 \leq j \leq N$, be an image with N^2 pixels. Assume we observe only a noisy version Y obtained with an additive random error W:

$$Y(i, j) = I(i, j) + \sigma W(i, j)\,.$$

We assume that $W(i, j)$ are i.i.d standard normal distributions and the variance σ^2 is supposed to be known.

Let S be an odd integer, we define the patch $P(I)(i, j)$ as the subimage of I of size $S \times S$ centered on (i, j):

$$P(I)(i,j)(k,l) = I\left(i + k - \frac{S-1}{2}, j + l - \frac{S-1}{2}\right)\,.$$

In patch based methods, one is interested by estimates of the patches $P(I)(i_0, j_0)$ obtained from the collection of noisy patches $P(Y)(i, j)$. More precisely, we consider weighted estimators

$$P_\lambda(Y)(i_0, j_0) = \sum_{1\leq i\leq N, 1\leq j\leq N} \lambda_{(i_0,j_0)(i,j)}(Y) P(Y)(i,j)\,,$$

where the weights $\lambda_{(i_0,j_0)(i,j)}(Y)$ should not depend on the unobserved I.

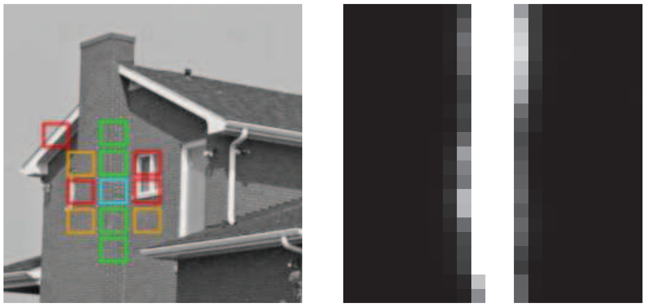

(a) Patches to consider (b) Associated kernel

Fig. 1. (a) House image with some patches used to denoise the blue central patch (the lighter the closer). (b) Associated local kernel showing geometrical structure.

Those weights should be chosen in such a way that this average is close to the true patch $P(I)$, as illustrated in Fig.1. The most classical choice is the one where those weights depend only on the relative position of (i_0, j_0) and (i, j). This case is nothing but a kernel method in which a fixed kernel is chosen and used to uniformly smooth the noisy image. Often, a single kernel is not adapted to the whole image and several methods propose to change locally this kernel according to the observation itself.

A simple and very efficient way has been proposed by Buades, Coll and Morel [1]: just replace the dependency on the position by a dependency on the distance between the corresponding patches and renormalize the weights so that they sum to 1. Indeed, the weights they proposed are defined as

$$\lambda_{(i_0,j_0)(i,j)}(Y) = \frac{e^{-\frac{1}{\beta}\|P(Y)(i_0,j_0)-P(Y)(i,j)\|^2}}{\sum_{i',j'} e^{-\frac{1}{\beta}\|P(Y)(i_0,j_0)-P(Y)(i',j')\|^2}} ,$$

where $\|\cdot\|$ denotes the usual l^2 norm and β is an important tuning parameter. This method can be seen as an extension of the bilateral filter [6] where the pixels difference based kernel is replaced by patches difference based kernel. The weights can accordingly vary, for example to take into account the relative spatial position of the patches. Note that, as observed by Boulanger and Kervrann [2], it is important to restrict the patch averaging to a neighborhood of (i, j) for both speed and performance reasons. In the following, we will assume that the weights are zero as soon as $\|(i,j)-(i_0,j_0)\|_\infty > \frac{\sqrt{m}-1}{2}$, so that we only combine m patches.

3. AGGREGATION AND PAC-BAYESIAN APPROACH

The aggregation framework is very similar. One also has a noisy observation Y of size n of a signal I:

$$Y(\mathbf{i}) = I(\mathbf{i}) + \sigma W(\mathbf{i}) ,$$

where $W(\mathbf{i})$ are standard i.i.d. gaussian variables. σ is known and one has a collection of m pre-estimators $P_1, \ldots, P_m$. In this context, we look for an estimate of I with the following form

$$P_\lambda = \sum_{k=1}^{m} \lambda_k P_k ,$$

where λ belongs to $\mathbb{R}^m$. The aggregation theory deals with the choice of a data dependent λ that leads to a theoretical control on the estimation error.

In the PAC-Bayesian approach, for any probability measure π, called the prior, on the parameter λ, one defined the exponential weight aggregate $\hat{I}_\pi$ by

$$\hat{I}_\pi = \frac{\int_{\mathbb{R}^m} I_\lambda \exp^{-\frac{1}{\beta}\|Y-P_\lambda\|^2} d\pi(\lambda)}{\int_{\mathbb{R}^m} \exp^{-\frac{1}{\beta}\|Y-P_\lambda\|^2} d\pi(\lambda)} ,$$

where $\|\cdot\|$ is the standard ℓ^2 norm. This estimator can be interpreted in the Bayesian framework as the posterior mean in a phantom model $Z = I + \sqrt{\frac{\beta}{2}} W_g$ where W_g is a standard gaussian white noise. This estimate can also be recast through its coordinates: straightforward computation show that $\hat{I}_\pi = I_{\hat{\lambda}_\pi}$ with

$$\hat{\lambda}_\pi = \frac{\int_{\mathbb{R}^m} \lambda \exp^{-\frac{1}{\beta}\|Y-P_\lambda\|^2} d\pi(\lambda)}{\int_{\mathbb{R}^m} \exp^{-\frac{1}{\beta}\|Y-P_\lambda\|^2} d\pi(\lambda)} .$$

The key result is that this estimator satisfies an "oracle" inequality whenever the pre-estimators P_k are independent of Y and β large enough. Let $\mathcal{P}$ be the set of all probability distributions on $\mathbb{R}^m$. For any $\pi \in \mathcal{P}$ and any $\beta \geq 4\sigma^2$, Dalalyan and Tsybakov [4] prove that

$$\mathbb{E}\left[\|I - \hat{I}_\pi\|^2\right] \leq \inf_{p \in \mathcal{P}} \left(\int_{\mathbb{R}^m} \|I - P_\lambda\|^2 dp(\lambda) + \beta \mathcal{K}(p, \pi) \right) ,$$

where $\mathcal{K}(p,\pi)$ is the Kullback divergence between p and π:

$$\mathcal{K}(p,\pi) = \begin{cases} \int_{\mathbb{R}^m} \log\left(\frac{dp}{d\pi}(\lambda)\right) p(d\lambda) & \text{if } p \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

The first term in the above inequality is an approximation term. It relates the risk to a deterministic quantity, close to the risk of the best possible approximation in the family $(P_\lambda)_{\lambda \in \mathbb{R}^m}$. The second term is the price to pay for adaptivity, the fact that you do not know in advance the best probability distribution p, the one that makes the risk small.

4. PAC-BAYESIAN AND NL-MEANS

NL-Means and its variants with exponential weights for the patches can be seen as a special case of PAC-Bayesian estimation. Indeed, it suffices to use $P(Y)$ as the observation and to choose the pre-estimators P_k as the noisy patches $P(Y)(i,j)$ to obtain, for any prior π, an estimator of the following form

$$P_{\hat{\lambda}_\pi}(Y) = \sum_{1 \leq i \leq N, 1 \leq j \leq N} \hat{\lambda}_{\pi,(i,j)}(Y) P(Y)(i,j) ,$$

where $$\hat{\lambda}_\pi = \frac{\int_{\mathbb{R}^m} \lambda \exp^{-\frac{1}{\beta}\|P(Y)-P_\lambda(Y)\|^2} d\pi(\lambda)}{\int_{\mathbb{R}^m} \exp^{-\frac{1}{\beta}\|P(Y)-P_\lambda(Y)\|^2} d\pi(\lambda)} .$$

When π is chosen as the uniform discrete probability on the noisy patches, one obtains the simple formula

$$\hat{I}_\pi = \frac{\sum_{i,j} P(Y)(i,j) \exp^{-\frac{1}{\beta}\|P(Y)-P(Y)(i,j)\|^2}}{\sum_{i,j} \exp^{-\frac{1}{\beta}\|P(Y)-P(Y)(i,j)\|^2}} ,$$

that is the classical NL-Means estimator.

The theoretical results of the previous section have been proved only when the pre-estimators are independent of the observations, which is obviously not the case when they are chosen as patches of the noisy image. However, numerical experiments suggest that a similar result is valid. More precisely, we expect

$$\mathbb{E}\left[\|P(I) - P_{\hat{\lambda}_\pi}(Y)\|^2\right] \leq$$
$$\inf_{p\in\mathcal{P}}\biggl(\int_{\mathbb{R}^m}\bigl(\|P(I) - P_\lambda(I)\|^2 + S^2\sigma^2\|\lambda\|^2\bigr)dp(\lambda) + \beta\mathcal{K}(p,\pi)\biggr)\,.$$

The supplementary term, $S^2\sigma^2\|\lambda\|^2$, is exactly the variance of the kernel estimator corresponding to the choice of weights λ: the value in the integral is thus the expected error of a fixed kernel smoothing estimator. This oracle inequality means that the risk of the PAC-Bayesian estimator is controlled by the average of the risk of any kernel estimator up to a penalty which measures the distance between the averaging probability p and the prior used. The point is that this is valid for every p so one can mimic the best local kernel up to this penalty term.

The choice of the prior π is crucial to control the error. A good choice will be one such that for any patch $P(I)$ there is a probability p which makes the left-hand side of the error bound small. An ideal one is one for which the attained minimum is close to the minimum without the complexity penalty. It is impossible to reach in general but possible if one restricts the choice of p and π within a certain class, so that we can control the sum of the two terms.

The simple uniform prior corresponding to the NL-Means is not yet handled efficiently by these analysis. Indeed, the previous inequality gives

$$\mathbb{E}\|P(I) - P_{\hat{\lambda}_\pi}(Y)\|^2 \leq S\sigma^2 + \beta\log(m)\,,$$

meaning that one does better than nothing up to a logarithmic factor.

A much more interesting results has been obtained by Dalalyan and Tsybakov [4] with a 3-Student law as a prior π, ie. $\pi(d\lambda) \propto (\tau^2 + \lambda_j^2)^{-2}d\lambda$. They obtain a sparse oracle inequality showing that if the best kernel has only few non zero elements then the PAC-Bayesian estimate behaves almost as well as this best kernel.

The question of a better prior choice remains open as we want to have simultaneously a theoretical control, an efficient estimator and an efficient algorithm.

5. PAC BAYESIAN ESTIMATORS AND MONTE CARLO METHOD

Computing efficiently the proposed estimator is indeed a challenging task. Recall that our estimate has the form $P_{\hat{\lambda}_\pi}(Y)$ with

$$\hat{\lambda}_\pi = \frac{\int_{\mathbb{R}^m}\lambda\exp^{-\frac{1}{\beta}\|P(Y)-P_\lambda(Y)\|^2}\,d\pi(\lambda)}{\int_{\mathbb{R}^m}\exp^{-\frac{1}{\beta}\|P(Y)-P_\lambda(Y)\|^2}\,d\pi(\lambda)}\,.$$

Thus, the challenge is in computing such a multi dimensional integral. This "posterior" computation appears quite often in Bayesian approach and a huge literature already exists on the subject (see [7] for instance).

Most approaches are based on the Monte-Carlo Markov Chain (MCMC) machinery which yields efficient approximation scheme for this type of integrals. We focus here on a method based on diffusion techniques called Langevin Walk Monte Carlo. Note that as in most MCMC method, we only need to know the distribution up to a multiplicative constant, so we do not need to compute the normalization constant appearing in the previous formulae.

The key observation on the Langevin Walk Monte Carlo is that whenever the probability q has a density proportional to $\exp(V(\lambda))$ where V is a continuous function, there is a simple diffusion process for which q is the stationary law. Indeed, stochastic integral theory shows that the solution of the Langevin diffusion equation

$$d\Lambda_t = \nabla V(\Lambda_t)dt + \sqrt{2}\,dWt \quad \Lambda_0 = \lambda_0, \quad t \geq 0\,,$$

where λ_0 is a fixed vector in $\mathbb{R}^m$ and W_t is a m-dimensional Brownian motion, is such that, under mild assumptions on V, any trajectory Λ_t is stationary with a stationary distribution equal to q. The expectation of any function of λ with respect to q can thus be obtained as the average of this function along the trajectory Λ_t.

With the choice $V(\lambda) = -\beta^{-1}\|P(Y) - P_\lambda(Y)\|^2 - \log(\pi(\lambda))$, we obtain

$$\hat{\lambda}_\pi = \lim_{T\to+\infty}\frac{1}{T}\int_0^T \Lambda_t d_t,$$

where Λ_t is any trajectory solution of the Langevin diffusion. We approximate this integral by discretizing it with step h and computing an approximate diffusion Λ^E_{kh} with an Euler scheme. We let $\Lambda^E_0 = 0$ and compute recursively the approximate solution Λ^E_{hk} for $k = 1, \ldots, [T/h] - 1$:

$$\Lambda^E_{(k+1)h} = \Lambda^E_{kh} + h\nabla V(\Lambda^E_{kh}) + \sqrt{2h}W_k,$$

with $W_1, W_2, \ldots$ i.i.d standard gaussian random vectors in $\mathbb{R}^m$. The estimated coefficients $\hat{\lambda}_\pi$ are then replaced by the approximation $\tilde{\lambda}_\pi$,

$$\tilde{\lambda}_\pi = \frac{h}{T}\sum_{k=0}^{[T/h]-1}\Lambda^E_{hk}\,.$$

For a small enough h, both theory [8] and numerical experiments ensure the closeness with the true value.

Note that the we can also use the Langevin diffusion to define the drift of a Metropolis Walk Monte Carlo chain. This corresponds to a small correction in the recursion of the previous discrete approximation which ensures the convergence but slows down the algorithm.

6. NUMERICAL RESULTS

We implemented the proposed approach in Matlab, focusing on two priors: the uniform discrete prior corresponding to the NL-Means (with $\beta = 12\sigma^2$ as recommanded in [1]) and the 3-Student prior. We used square patches with length $S = 5$ and a search region of size $m = 13 \times 13$ in all our experiments. The choice $\beta = 4\sigma^2$, recommended by the theory, does not lead to the best results: the choice $\beta = 2\sigma^2$ corresponding to a classical Bayesian approach leads to better performances.

To better control the convergence of our procedure, we used several chains instead of only one. We speed up the process by eleminating from pre-estimators patches that are too much different from the central patch.

The results are on par with the classical NL-Means procedure as illustrated in Fig.2 with two classical images. One can still see some kind of grain in images treated by our method, what seems more comfortable for our eyes than too flat region obtained with NL-Means. Moreover, we have observed that in PAC-Bayesian aggregation, the same parameter set yields good results for all our test images while the optimization is important for the NL-Means.

We have describe a new denoising algorithm inspired by the NL-Means procedure and we have proposed a new framework to transfer statistical aggregation results into image processing theorems. In the future, we plan to improve the algorithm by exploring several other heavy tail priors such as Gaussian mixture or Cauchy.

7. REFERENCES

[1] A. Buades, B. Coll, and J-M. Morel, "A review of image denoising algorithms, with a new one," *Multiscale Model. Simul.*, vol. 4, no. 2, pp. 490–530 (electronic), 2005.

[2] C. Kervrann and J. Boulanger, "Optimal spatial adaptation for patch-based image denoising," *IEEETIP*, vol. 15, no. 10, pp. ’2866–2878, 2006.

[3] A. B. Juditsky, A. V. Nazin, A. B. Tsybakov, and N. Vayatis, "Recursive aggregation of estimators by the mirror descent method with averaging," *Problemy Peredachi Informatsii*, vol. 41, no. 4, pp. 78–96, 2005.

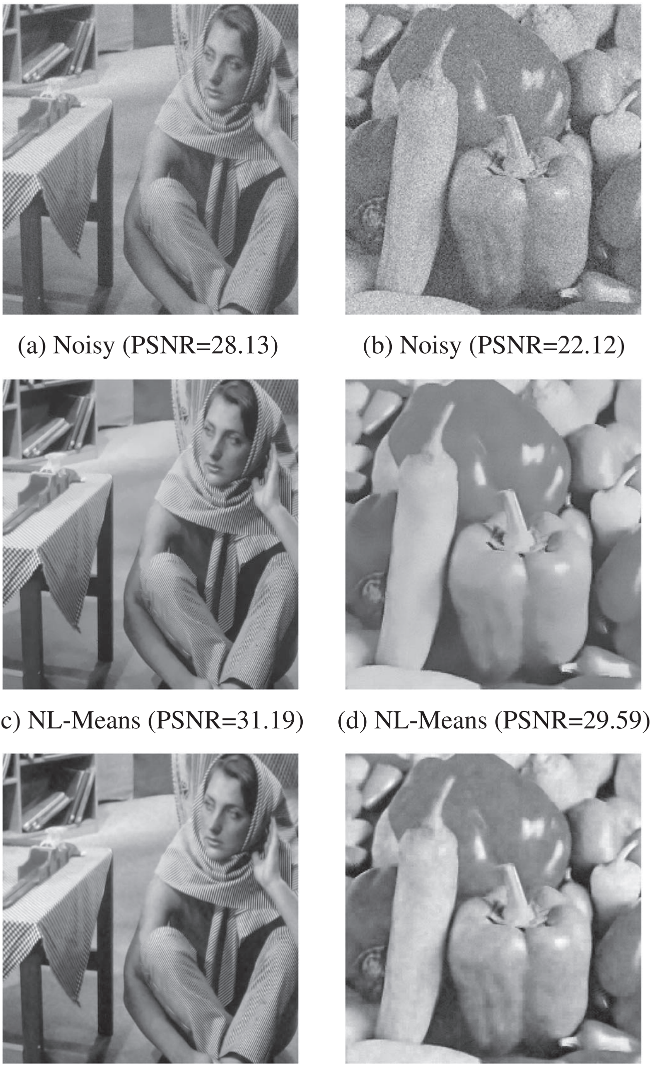
[4] A. S. Dalalyan and A. B. Tsybakov, "Aggregation by exponential weighting, sharp oracle inequalities and sparsity," in *20th Annual Conference on Learning Theory, COLT*, 2007, pp. 97–111.

[5] A. S. Dalalyan and A. B. Tsybakov, "Sparse regression learning by aggregation and langevin monte-carlo," in *22th Annual Conference on Learning Theory, COLT*, 2009.

[6] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," *Computer Vision, IEEE International Conference on*, vol. 0, pp. 839, 1998.

[7] C.P. Robert and G. Casella, *Monte Carlo statistical methods*, Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

[8] G.O. Roberts and O. Stramer, "Langevin diffusions and metropolis-hastings algorithms," *Methodol. Comput. Appl. Probab.*, vol. 4, no. 4, pp. 337–357 (2003), 2002.



(a) Noisy (PSNR=28.13)

(b) Noisy (PSNR=22.12)

(c) NL-Means (PSNR=31.19)

(d) NL-Means (PSNR=29.59)

(e) Our method (PSNR=32.79)

(f) Our method (PSNR=29.46)

Fig. 2. (a) Barbara noisy $\sigma = 10$. (b) Peppers noisy $\sigma = 20$. (c) and (d) Images denoised with the NL-Means. (e) and (f) Images denoised with our method.