An aggregator point of view on NL-Means

J. Salmon and E. Le Pennec

LPMA, Université Paris Diderot, France

ABSTRACT

Patch based methods give some of the best denoising results. Their theoretical performances are still unexplained mathematically. We propose a novel insight of NL-Means based on an aggregation point of view. More precisely, we describe the framework of PAC-Bayesian aggregation, show how it allows to derive some new patch based methods and to characterize their theoretical performances, and present some numerical experiments.

Keywords: Denoising, NL-Means, Aggregation, PAC-Bayesian, Patch

Some of the best denoising results are obtained by the patch based NL-means method proposed by Buades et al.[1] or by some of its variants.[2] These methods are based on a simple idea: consider the image not as a collection of pixels but as a collection of sub-images, the "patches", centered on those pixels and estimate each patch as a weighted average of patches. These weights take into account the similarities of the patches and are often chosen proportional to the exponential of the quadratic difference between the patches with a renormalization so they sum to 1. Understanding why these methods are so efficient is a challenging task.

In their seminal paper, Buades et al.[1] show the consistency of their method under a strong technical β-mixing assumption on the image. NL-Means methods can also be seen as a smoothing in a patch space with a Gaussian kernel and their performances are related to the regularity of the underlying patch manifold (see for instance Peyr[3] for a review). While intuitive and enlightening, those points of view have not yet permited to justify mathematically the performance of the NL-Means methods.

We propose to look at those methods with a different eye so as to propose a different path to their mathematical justification. We consider them as special instance of statistical aggregation. In this framework, one consider a collection of preliminary estimators and a noisy observation. We search then for a weighted average of those preliminary estimators. This "aggregate" estimate should be as close as possible to the unknown original signal. If one uses patches as preliminary estimators, a special case of a recent method inspired by PAC-Bayesian techniques[4] almost coincides with the NL-Means.

In the sequel, we describe this framework, propose some novel variants of patch based estimators and give some insights on their theoretical performances.

1. IMAGE DENOISING, KERNEL AND PATCH METHODS

We consider an image I defined on a grid (i,j), $1 \leq i \leq N$ and $1 \leq j \leq N$, of N^2 pixels and assume we observe a noisy version Y:

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

where W is a white noise, an i.i.d. standard Gaussian sequence and σ is a known standard deviation parameter. Our goal is to estimate the original image I from the noisy observation Y.

Numerous methods have been proposed to fulfill this task. Most of them share the principle that the observed value should be replaced by a suitable local average, a local smoothing. Indeed all the kernel based methods, and even the dictionary based methods (thresholding for example), can be put in this framework. They differ in the way this local average is chosen. Those methods can be represented as a locally weighted sum

$$\hat{I}(i,j) = \sum_{k,l} \lambda_{i,j,k,l} Y(k,l)$$

Corresponding author: E. Le Pennec

J. Salmon: email: salmon@math.jussieu.fr

E. Le Pennec: email: lepennec@math.jussieu.fr

where the weights $\lambda_{i,j,k,l}$ may depend in a complex way on both the indices and the values of Y. The weights $\lambda_{i,j,k,l}$ for a fixed pixel (i,j) are nothing but the weights of a local smoothing kernel. The most famous weights are probably those of the Nadaraya-Watson estimator,

$$\lambda_{i,j,k,l} = \frac{K(i-k, j-l)}{\sum_{k',l'} K(i-k', j-l')} \quad ,$$

where K is a fixed kernel (Gaussian for example). To make the estimator more efficient, the kernel and its scale can also vary depending on the local structure of the image such as in some locally adaptive method. Even if this is less explicit the representation based method can be put in this framework with a subtle dependency of the weights i, j, k, l on the values of Y.

Patch based methods can be seen as extensions of such methods in which the image f and the observation Y are lifted in a higher dimensional space of patches. More precisely, for a fixed integer odd S, we define the patch $P(I)(i,j)$ as the sub-image of I of size $S \times S$ centered on (i,j) (for the sake of simplicity we assume here a periodic extension across the boundaries):

$$P(I)(i,j)(k,l) = I\,(i+k, j+l) \quad \text{for } -\frac{S-1}{2} \leq k,l \leq \frac{S-1}{2}.$$

An image I belonging to $\mathbb{R}^{N^2}$ can thus be sent in a space of patch collection of dimension $\mathbb{R}^{N^2 \times S^2}$ through the application

$$I \mapsto P(I) = (P(I)(i,j))_{1 \leq i,j \leq N} \quad .$$

The denoising problem is reformulated as retrieving the original patch collection $\mathcal{P}(I)$ from the noisy patch collection $\mathcal{P}(Y)$. Note that an estimate $\hat{I}$ of the original image I can be obtained from any estimate $\widehat{P(I)}$ of the original patch collection through a simple projection operator for example using the central values of the patches

$$\widehat{P(I)} \to \hat{I} = \left(\hat{I}(i,j) = \widehat{P(I)(i,j)}(0,0)\right)_{1 \leq i,j \leq N} \quad .$$

This simple projection can also be replaced by a more complex one, in which the value of a given pixel is obtain by averaging the values obtained for this pixel in different patches.

Following the approach used for images, we consider in this paper patch methods based on weighted sums

$$\widehat{P(I)(i,j)} = \sum_{k,l} \lambda_{i,j,k,l} P(Y)(k,l)$$

Note that when the $\lambda_{i,j,k,l}$ are chosen as in the Nadaraya-Watson estimator, the patch based estimator and the original pixel based estimator coincide. We will thus consider some other weight choices in which the weights for a given patch depends on the values of the other patches.

The method proposed by Buades et al.[1] corresponds exactly to the use of the weights $\lambda_{i,j,k,l}$;

$$\lambda_{i,j,k,l} = \frac{e^{-\frac{1}{\beta}\|P(i,j)-P(k,l)\|^2}}{\sum_{k,l} e^{-\frac{1}{\beta}\|P(i,j)-P(k,l)\|^2}}$$

where $\|\cdot\|^2$ is the usual euclidean distance on patches. They have called this method Non Local Means (NL-Means from now on) as the weights depends only on the values of the patches and not on the distance between the patches (the distance between their centers). The influence of a patch on the reconstruction of another patch depends thus on their similarity so that the corresponding local smoothing kernels adapt themselves to the local structures in the image as illustrated in Figure 1.

They obtain the consistency of their method for stochastic process under some technical β-mixing conditions. Most of the other explanations of this method rely on the existence of a patch manifold of low dimension in which the patches live.[3] The NL-Means method appears then as a local averaging using a Gaussian kernel in

Figure 1. Adaptation of the NL-Means kernel to the local structures. The two right images show the kernel weights $(\lambda_{i,j,k,l})_{k,l}$ obtained for a patch in a uniformly regular zone and a patch centered on an edge.

this patch space. Under the strong assumptions that the patches are evenly spaced on the patch manifold and that this manifold is flat enough to be approximated by an affine space, the performance of the NL-Means can be explained. Unfortunately, there is no guarantee this is the case.

Note that the strict non locality of this construction has been contradicted by some further studies,[2] which show that using only patches in a neighborhood of the considered patch in the weights formula yields a significant improvement. The temperature parameter β is also an important issue from both the theoretical and the practical point of view. We conclude this review by stressing that adding a localization term, such as a classical spatial kernel, renders the scheme close to a bilateral filtering in which the data dependent term is computed on the patch metric.

2. AGGREGATION AND THE PAC-BAYESIAN APPROACH

In this paper, we propose a different point of view on this method: the aggregation point of view. In this setting, we consider a collection of preliminary estimates $\hat{f}_m$ of a given object f and search for the best adaptive weighted combination

$$\hat{f}_\lambda = \sum_{m=1}^{M} \lambda_m \hat{f}_m$$

of those estimates from a noisy observation $Y = f + \sigma W$. This setting has been introduced by Nemirovski[5] and Yang[6] and is the subject of a lot of studies since. This model is quite general as, for instance, both thresholding and estimator selection can be put in this framework. The key question is how to choose the aggregating weights.

We focus here on a special case in which the estimators are constructed for patches and the aggregation is based on the PAC-Bayesian approach.[4,7]

For any patch $P(I)(i,j)$, we assume we observe a noisy patch $P(Y)(i,j)$ and a collection of M preliminary estimators P_1,...,P_M. We look then for an estimate

$$\widehat{P(I)(i,j)}_\lambda = \sum_{m=1}^{M} \lambda_m P_m$$

where λ belongs to $\mathbb{R}^M$. The weights λ_m are chosen, in the PAC Bayesian approach, in a very specific way from an arbitrary prior law π on $\mathbb{R}^M$. The PAC-Bayesian aggregate $\widehat{P(I)(i,j)}^{\pi}$ is defined by the weighted "sum"

$$\widehat{P(I)(i,j)}^{\pi} = \int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\|P(Y)(i,j) - \widehat{P(I)(i,j)}_\lambda\|^2}}{\int_{\mathbb{R}^m} e^{-\frac{1}{\beta}\|P(Y)(i,j) - \widehat{P(I)(i,j)}_{\lambda'}\|^2} d\pi(\lambda')} \widehat{P(I)}_\lambda d\pi(\lambda) \quad .$$

or equivalently by its weight components

$$\lambda_\pi = \int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\|P(Y)(i,j)-\widehat{P(I)(i,j)}_\lambda\|^2}}{\int_{\mathbb{R}^m} e^{-\frac{1}{\beta}\|P(Y)(i,j)-\widehat{P(I)(i,j)}_{\lambda'}\|^2} d\pi(\lambda')} \lambda d\pi(\lambda) \quad .$$

Note that this estimator can be interpreted as a pseudo Bayesian estimator with a prior law π in which the noise of variance σ^2 is replaced by a Gaussian noise of variance $\beta/2$.

The formula defining the estimator in the PAC-Bayesian approach looks similar to the the formula defining the weights of the NL-Means, they are indeed equivalent when the preliminary estimators P_m span the set of the noisy patches $P(Y)(k,l)$ and the prior law π is chosen as the discrete law

$$\pi = \frac{1}{N^2} \sum_{(k,l)} \delta_{e_{(k,l)}}$$

where the sum runs across all the patches and $\delta_{e_{(k,l)}}$ is the Dirac measure charging only the patch $P(Y)(k,l)$. This choice leads to the estimate

$$\widehat{P(I)(i,j)}^{\pi} = \sum_{(k,l)} \frac{e^{-\frac{1}{\beta}\|P(Y)(i,j)-P(Y)(k,l)\|^2}}{\sum_{(k',l')} e^{-\frac{1}{\beta}\|P(Y)(i,j)-P(Y)(k',l')\|^2}} P(Y)(k,l) \quad ,$$

that is exactly the NL-Means estimator.

A lot of other variants of patch based method can be obtained through a suitable choice for the prior π. For example, for any kernel K,

$$\pi = \sum_{(k,l)} \frac{K(i-k, j-l)}{\sum_{(k',l')} K(i-k', j-l')} \delta_{e_{k,l}}$$

yields the localized NL-Means often used in practice.

3. STEIN UNBIASED RISK ESTIMATOR AND ERROR BOUND

The analysis of the risk of this family of estimator is based on a SURE (Stein Unbiased Risk Estimator) principle.[4,8] Indeed, assume that the preliminary estimators P_m are independent of $P(Y)(i,j)$, a simple computation shows that

$$\hat{r}_\lambda = \|P(Y)(i,j) - \widehat{P(I)(i,j)}_\lambda\|^2 - S^2\sigma^2$$

is an unbiased estimate of the risk of the estimator $\widehat{P(I)(i,j)}_\lambda$, $\|P(I)(i,j) - \widehat{P(I)(i,j)}_\lambda\|^2$. As $S^2\sigma^2$ is a term independent of λ, the PAC-Bayesian estimate of the previous section can be rewritten as

$$\widehat{P(I)(i,j)}^{\pi} = \int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^m} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}} d\pi(\lambda')} \widehat{P(I)(i,j)}_\lambda d\pi(\lambda)$$

Using Stein's formula, one is able to construct an unbiased estimate $\hat{r}$ of the risk of this estimator such that, as soon as $\beta \geq 4\sigma^2$,

$$\hat{r} \leq \int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^m} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}} d\pi(\lambda')} \hat{r}_\lambda d\pi(\lambda) \quad .$$

The key is then to notice (see for instance Catoni[7]) that this renormalized exponential weights are such that for any probability law p

$$\int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^m} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}} d\pi(\lambda')} \hat{r}_\lambda d\pi(\lambda) + \beta\mathcal{K}\left(\frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^m} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}}} \pi, \pi\right) \leq \int_{\mathbb{R}^M} \hat{r}_\lambda dp(\lambda) + \beta\mathcal{K}(p,\pi)$$

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

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

Thus, as $\mathcal{K}(p,\pi)$ is always a positive quantity,

$$\hat{r} \leq \inf_{p\in\mathcal{P}} \int_{\mathbb{R}^M} \hat{r}_\lambda dp(\lambda) + \beta\mathcal{K}(p,\pi) \quad .$$

Taking the expectation and interchanging the order of the expectation and the infimum yield

$$E(\hat{r}) \leq E\left(\inf_{p\in\mathcal{P}} \int_{\mathbb{R}^M} \hat{r}_\lambda dp(\lambda) + \beta\mathcal{K}(p,\pi)\right) \leq \inf_{p\in\mathcal{P}} \left(\int_{\mathbb{R}^M} E(\hat{r}_\lambda) dp(\lambda) + \beta\mathcal{K}(p,\pi)\right)$$

or more explicitly using the fact that the $\hat{r}$ are unbiased estimate of the risks

$$\mathbb{E}\left(\|P(I)(i,j) - \widehat{P(I)(i,j)}^{\pi}\|^2\right) \leq \inf_{p\in\mathcal{P}} \left(\int_{\mathbb{R}^M} \|P(I)(i,j) - \widehat{P(I)(i,j)}_\lambda\|^2 dp(\lambda) + \beta\mathcal{K}(p,\pi)\right) \quad .$$

The PAC-Bayesian aggregation principle is thus supported by a strong theoretical result when the preliminary estimators P_m are independent of $P(Y)$, often call the frozen preliminary estimators case, and β is larger than $4\sigma^2$. The quadratic error of the PAC-Bayesian estimate is bounded by the best trade-off between the average quadratic error of fixed λ estimators under a law p and an adaptation price corresponding to the Kullback distance between p and the prior π. The optimal p is thus one both concentrated around the best fixed λ estimator and close to the prior law π.

So far, these results have been proved only when the preliminary estimators are independent of the observation, which is obviously not the case when they are chosen as patches of the noisy images. We conjecture that the following similar inequality holds

$$\mathbb{E}\|P(I)(i,j) - \widehat{P(I)(i,j)}^{\pi}\|^2 \leq \inf_{p\in\mathcal{P}} \left(\int_{\mathbb{R}^m} \left(\|P(I)(i,j) - P(I)(i,j)_\lambda\|^2 + S^2\sigma^2\|\lambda\|^2\right) dp(\lambda) + \beta\mathcal{K}(p,\pi)\right)$$

where the $S^2\sigma^2\|\lambda\|^2$ term appears as the variance of the estimator for a fixed λ, which is nothing but a classical kernel estimator. The trade-off for p is thus between a concentration around the best linear kernel and a proximity with the prior law π. The aggregation point of view shows that this patch based procedure is close to a search for an optimal local kernel, which is one of the intuition behind the NL-Means construction.

We have obtained this result so far in three cases: when patches are computed on an other noisy image, when all patches intersecting the central patch are removed from the collection and with a small modification of the weights when the image is split into two separate images with a quincunx grid. We are still working on a proof for the general case that requires some more modifications of the weights.

4. PRIORS AND NUMERICAL ASPECTS OF AGGREGATION

The most important parameter to obtain a good control of the error is thus the prior π. A good choice is one such that for any natural patch $P(i)(i,j)$ there is a probability law p close to π and concentrated around the best kernel weights. The goal is to prove that the penalty due to the Kullback divergence term is not too big compare to the best kernel performance.

We propose here three choices for the prior:

i) the uniform discrete prior π leading to the NL-Means estimate,

$$\pi = \frac{1}{M}\sum_{m=1}^{M} \delta_m \quad ,$$

ii) a 3-Student sparsifying prior proposed by Dalalyan and Tsybakov,

$$\pi(d\lambda) \propto \prod_m (\tau^2 + \lambda_m^2)^{-2} d\lambda$$

iii) a Gaussian mixture which promotes more diversity

$$\pi(d\lambda) = \prod_m \left((1-\alpha) \frac{1}{\sqrt{2\pi}\epsilon} e^{-\lambda_m^2/(2\epsilon^2)} + \alpha \frac{1}{\sqrt{2\pi}\tau} e^{-\lambda_m^2/(2\tau^2)} \right)$$

For the two last choices, PAC-Bayesian theory relates the risk of each estimator to the one of the best kernel up to some small term due to adaptivity.

The unbiased estimated risk estimate used are

$$\hat{r}_\lambda = \|P(Y)(i,j) - \widehat{P(I)(i,j)}_\lambda\|^2 - S\sigma^2$$

when the preliminary estimators are fixed. When they are the patches themselves, a correction should be made when λ_0, the weight corresponding to the central patch, is non zero:

$$\hat{r}_\lambda = \|P(Y)(i,j) - \widehat{P(I)(i,j)}_\lambda\|^2 - S(1-2\lambda_0)^2\sigma^2 \quad .$$

Note that for standard NL-Means, when the uniform term $S\sigma^2$ is not added, this leads to a simple rule for the weight of the central patch: choose it proportional to

$$e^{-\frac{1}{\beta}2S\sigma^2} \quad .$$

Computing the proposed estimator is a non-trivial task: it requires the computation of a multi dimensional integral defining the weights λ:

$$\lambda_\pi = \int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\hat{r}_\lambda}}{\int_{\mathbb{R}^M} e^{-\frac{1}{\beta}\hat{r}_{\lambda'}} d\pi(\lambda')} \lambda d\pi(\lambda) \quad .$$

This kind of integral appears often in Bayesian approach and a huge literature already exists on the subject (see for instance[9] for a reference). Most approaches are based on a Monte-Carlo Markov Chain or a variation thereof.

Following Dalalyan and Tsybakov,[10] we propose here to use Monte-Carlo Markov Chain in which the drift is directed by a Langevin diffusion. Indeed, whenever the probability q has a density proportional to $\exp(V(\lambda))$ where V is a continuous function, there is a simple diffusion process, the Langevin diffusion,

$$d\Lambda_t = \nabla V(\Lambda_t)dt + \sqrt{2}\, dWt \quad \Lambda_0 = \lambda_0, \quad t \geq 0 \quad , \tag{1}$$

where λ_0 is a fixed vector in $\mathbb{R}^m$ and W_t is a m-dimensional Brownian motion for which q is the stationary law. Stochastic integral theory shows that under mild assumptions on V, any trajectory Λ_t solution to the equation is stationary with a stationary distribution equal to q. Any expectation again q can thus be computed along the trajectory.

We exploit this property by choosing

$$V(\lambda) = -\frac{1}{\beta}\hat{r}_\lambda - \log(\pi(\lambda)) \quad ,$$

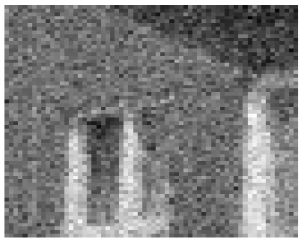
so that

$$\lambda_\pi = \lim_{T \to +\infty} \frac{1}{T} \int_0^T \Lambda_t d_t$$

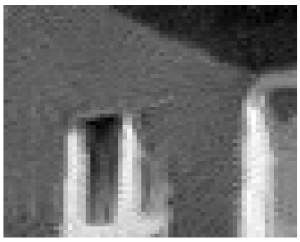
where Λ_t is any trajectory solution of the Langevin diffusion.

This integral is replaced by a Monte-Carlo Markov Chain driven by a discretized version of the diffusion. Assume a step-size h and an initial value λ_0 have been fixed, we let $\Lambda_0 = \lambda_0$ and construct recursively Λ_{k+1} from Λ_k with a usual MCMC construction:

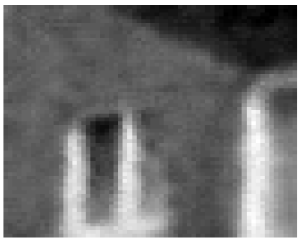
Original



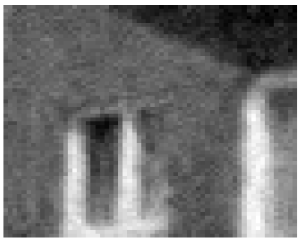
Noisy
22.0 dB



NL-Means
29.7 dB



Student
28.4 dB



Gaussian Mix.
29.5 dB

Figure 2. Numerical results on a small part of House for the 3 studied priors.

1. Compute the proposition $\Lambda^C = \Lambda_k + h\nabla V(\Lambda_k) + \sqrt{2h}W_{k+1}$ where W_k is an i.i.d. standard Gaussian sequence.
2. Compute the Metropolis-Hasting ratio
$$\alpha = \frac{e^{-V(\Lambda^C)} \times e^{-\frac{1}{4h}\|\Lambda^C + h\nabla V(\Lambda^C) - \Lambda_k\|^2}}{e^{-V(\Lambda_k)} \times e^{-\frac{1}{4h}\|\Lambda_k + h\nabla V(\Lambda_k) - \Lambda^C\|^2}}$$
3. Draw U_k uniformly on $[0,1]$ and set
$$\Lambda_{k+1} = \begin{cases} \Lambda^C & \text{if } U_k \leq \alpha \\ \Lambda_k & \text{otherwise} \end{cases}$$

Once k is large enough, we compute the approximate value of λ_π through the formula

$$\lambda_\pi \approx \frac{1}{k - k_{\min} + 1} \sum_{k'=k_{\min}}^{k} \Lambda_{k'} \quad .$$

where $k_{\min}$ is typically a small fraction of k'. General MCMC theory ensures, under mild assumptions on V, the convergence of this value to the true value.

Our numerical experiments can be summarized as follows:

- There is still a slight loss between our best method, the Gaussian mixture, and the optimized classical NL-Means. We have observed PAC-Bayesian aggregation is less sensitive to the parameters. The same parameter set yields good results for all our test images while for the NL-Means the temperature has to be tuned.

- The choice $\beta = 4\sigma^2$, recommended by the theory, does not lead to the best results: the choice $\beta = 2\sigma^2$ which corresponds to a classical Bayesian approach leads to better performances.
- The correction proposed for the central patch seems effective.
- We have observed that the central point is responsible for more than .5 dB gain in the NL-Means approach and less in the PAC-Bayesian approach.
- We are still facing some convergence issues in our Monte Carlo scheme which could explain our loss of performances. We are working on a modified scheme to overcome this issue.

The PAC-Bayesian approach provides a novel point of view on patch based method. From the theoretical point of view, we have been able to control the performance of our method. Improvements are however still required to transfer these results into numerical performances.

REFERENCES

[1] Buades, A., Coll, B., and Morel, J.-M., "A review of image denoising algorithms, with a new one," *Multiscale Model. Simul.* **4**(2), 490–530 (electronic) (2005).

[2] Kervrann, C. and Boulanger, J., "Optimal spatial adaptation for patch-based image denoising," *IEEETIP* **15**(10), '2866–2878 (2006).

[3] Peyré, G., "Manifold models for signals and images," *Computer Vision and Image Understanding* **113**, 249–260 (Feb 2009).

[4] Dalalyan, A. S. and Tsybakov, A. B., "Aggregation by exponential weighting, sharp oracle inequalities and sparsity," in [*COLT*], 97–111 (2007).

[5] Nemirovski, A., "Topics in non-parametric statistics," in [*Lectures on probability theory and statistics (Saint-Flour, 1998)*], *Lecture Notes in Math.* **1738**, 85–277, Springer, Berlin (2000).

[6] Yang, Y., "Combining different procedures for adaptive regression," *J. Multivariate Anal.* **74**(1), 135–161 (2000).

[7] Catoni, O., [*Statistical learning theory and stochastic optimization*], vol. 1851 of *Lecture Notes in Mathematics*, Springer-Verlag, Berlin (2004). Lecture notes from the 31st Summer School on Probability Theory held in Saint-Flour, July 8–25, 2001.

[8] Leung, G. and Barron, A. R., "Information theory and mixing least-squares regressions," *IEEE Trans. Inform. Theory* **52**(8), 3396–3410 (2006).

[9] Robert, C. P. and Casella, G., [*Monte Carlo Statistical Methods*], Springer (1999).

[10] Dalalyan, A. and Tsybakov, A., "Sparse regression learning by aggregation and Langevin Monte-Carlo," available at arXiv (2009).