NL-Means and Aggregation Procedures

Joseph Salmon [1] and Erwan Le Pennec [1,2]

[1]LPMA-UMR 7599
Université Paris Diderot - Paris 7

[2]SELECT Project
INRIA - Saclay

Nov. 2009 - ICIP

Image, Noise and Estimate

Image $N \times N$

- Pixel : $i = (i_1, i_2) \in [\![1, N]\!]^2$, Image : $f(i) \in \mathbb{R}$.
- $\| \cdot \|$: Euclidean Norm

Image, Noise and Estimate

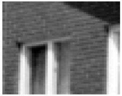


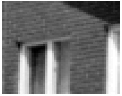
Image $N \times N$

- Pixel : $i = (i_1, i_2) \in [\![1, N]\!]^2$, Image : $f(i) \in \mathbb{R}$.
- $\| \cdot \|$: Euclidean Norm

Noisy Observation

- $Y(i) = f(i) + \sigma W(i)$
- $W(i)$ i.i.d. standard Gaussian noise, known σ
- Other noise possible

Image, Noise and Estimate



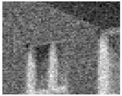


Image $N \times N$

- Pixel : $i = (i_1, i_2) \in [\![1, N]\!]^2$, Image : $f(i) \in \mathbb{R}$.
- $\| \cdot \|$: Euclidean Norm

Noisy Observation

- $Y(i) = f(i) + \sigma W(i)$
- $W(i)$ i.i.d. standard Gaussian noise, known σ
- Other noise possible

Estimate

- Estimate $f(i)$ from Y
- Non local behavior possible ...

Image, Noise and Estimate

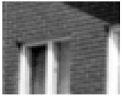
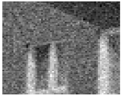
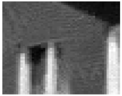
  

Image $N \times N$

- Pixel : $i = (i_1, i_2) \in [\![1, N]\!]^2$, Image : $f(i) \in \mathbb{R}$.
- $\|\cdot\|$: Euclidean Norm

Noisy Observation

- $Y(i) = f(i) + \sigma W(i)$
- $W(i)$ i.i.d. standard Gaussian noise, known σ
- Other noise possible

Estimate

- Estimate $f(i)$ from Y
- Non local behavior possible ...

Kernel Smoothing Method

General Method

- Estimate $f(i)$ through a local averaging :

$$\hat{f}(i) = \sum_{k \in [\![1,N]\!]^2} \theta_{i,k} Y(k)$$

- The weights $\theta_{i,k}$ can (will) depend on Y

Classical Kernel

- $\theta_{i,k} = \frac{K_h(i_1 - k_1, i_2 - k_2)}{\sum_{k'_1,k'_2} K_h(i_1 - k'_1, i_2 - k'_2)}$ (no dependency on Y)

- Example : Gaussian Kernel $K_h(i_1, i_2) = e^{-(i_1^2+i_2^2)/2h^2}$

Kernel Smoothing Method

General Method

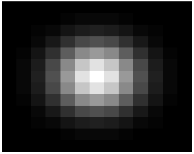
- Estimate $f(i)$ through a local averaging :

$$\hat{f}(i) = \sum_{k \in [\![1,N]\!]^2} \theta_{i,k} Y(k)$$

- The weights $\theta_{i,k}$ can (will) depend on Y

Classical Kernel

- $\theta_{i,k} = \frac{K_h(i_1 - k_1, i_2 - k_2)}{\sum_{k'_1,k'_2} K_h(i_1 - k'_1, i_2 - k'_2)}$ (no dependency on Y)



- Example : Gaussian Kernel $K_h(i_1, i_2) = e^{-(i_1^2+i_2^2)/2h^2}$

Data Dependant Kernel

Bilateral filtering

- $\theta_{i,k} = \frac{K_h(i_1 - k_1, i_2 - k_2) \times K'_{h'}(Y(i_1, i_2) - Y(k_1, k_2))}{\sum_{k'_1, k'_2} K_h(i_1 - k'_1, i_2 - k'_2) \times K'_{h'}(Y(i_1, i_2) - Y(k'_1, k'_2))}$
- Gaussian Version :

$$\theta_{i,k} = \frac{e^{-\frac{(i_1 - k_1)^2 + (i_2 - k_2)^2}{2h^2}} \times e^{-\frac{(Y(i_1, i_2) - Y(k_1, k_2))^2}{2h'^2}}}{\sum_{k'_1, k'_2} e^{-\frac{(i_1 - k'_1)^2 + (i_2 - k'_2)^2}{2h^2}} \times e^{-\frac{(Y(i_1, i_2) - Y(k'_1, k'_2))^2}{2h'^2}}}$$

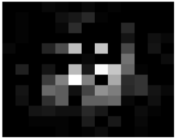
- Intuition : Average values that are close in both distance and values
- Issue : pixel value too local a feature (to be robust)

Data Dependant Kernel

Bilateral filtering

- $$\theta_{i,k} = \frac{K_h(i_1 - k_1, i_2 - k_2) \times K'_{h'}(Y(i_1, i_2) - Y(k_1, k_2))}{\sum_{k'_1, k'_2} K_h(i_1 - k'_1, i_2 - k'_2) \times K'_{h'}(Y(i_1, i_2) - Y(k'_1, k'_2))}$$
- Gaussian Version :

$$\theta_{i,k} = \frac{e^{-\frac{(i_1 - k_1)^2 + (i_2 - k_2)^2}{2h^2}} \times e^{-\frac{(Y(i_1, i_2) - Y(k_1, k_2))^2}{2h'^2}}}{\sum_{k'_1, k'_2} e^{-\frac{(i_1 - k'_1)^2 + (i_2 - k'_2)^2}{2h^2}} \times e^{-\frac{(Y(i_1, i_2) - Y(k'_1, k'_2))^2}{2h'^2}}}$$



- Intuition : Average values that are close in both distance and values
- Issue : pixel value too local a feature (to be robust)

Patches based Methods

Patch

- A patch $=$ a square sub-image of width w
- $P(f)(i)$: patch centered on i in the true image
- $P(Y)(i) = P_i$: patch centered on i in the noisy image
- A less localized version of pixel values : more robust
- Easy reprojection from patch collection $P(f)$ to an image f

Intuition

- Use weights that take into account the patch similarity :
  - Patch P to denoise
  - Similar Patches, useful : large weights
  - Less Similar Patches, less useful : small weights
  - Very Different Patches , useless : very small weights

Patches based Methods

Patch

- A patch $=$ a square sub-image of width w
- $P(f)(i)$: patch centered on i in the true image
- $P(Y)(i) = P_i$: patch centered on i in the noisy image
- A less localized version of pixel values : more robust
- Easy reprojection from patch collection $P(f)$ to an image f

Intuition

- Use weights that take into account the patch similarity :

  - Patch P to denoise
  - Similar Patches, useful : large weights
  - Less Similar Patches, less useful : small weights
  - Very Different Patches , useless : very small weights

Patches based Methods

Patch

- A patch $=$ a square sub-image of width w
- $P(f)(i)$: patch centered on i in the true image
- $P(Y)(i) = P_i$: patch centered on i in the noisy image
- A less localized version of pixel values : more robust
- Easy reprojection from patch collection $P(f)$ to an image f

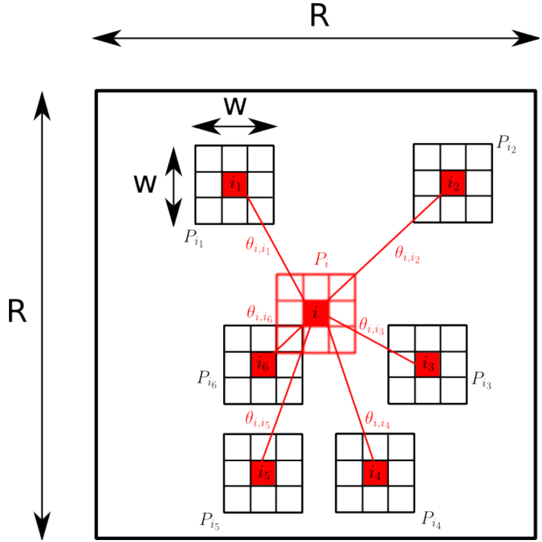
Intuition

- Use weights that take into account the patch similarity :

- Patch P to denoise
- Similar Patches, useful : large weights
- Less Similar Patches, less useful : small weights
- Very Different Patches , useless : very small weights

Searching Zone, Weights and Patches

Patch width : w=3, Searching zone width : R=15



NL-Means I

NL-Means [BCM05]

- Choose a dissimilarity measure D between patches.
- Use weights $\theta_{i,k} = \frac{K'(D(P_i, P_k))}{\sum_k K'(D(P_i, P_k))}$ with $D(P_i, P_k) = \|P_i - P_k\|$ to measure the dissimilarity, a Gaussian kernel $K'(x) = \exp(-x^2/\beta)$ and a temperature β.

Variations

- Adapt automatically the search zone (Kervrann et al. [KB06])
- Use higher order local approximations (Takeda et al. [TFM07])
- Use different dissimilarity measures (Azzabou et al. [APG07])

NL-Means I

NL-Means [BCM05]

- Choose a dissimilarity measure D between patches.
- Use weights $\theta_{i,k} = \dfrac{K'(D(P_i, P_k))}{\sum_k K'(D(P_i, P_k))}$ with $D(P_i, P_k) = \|P_i - P_k\|$ to measure the dissimilarity, a Gaussian kernel $K'(x) = \exp(-x^2/\beta)$ and a temperature β.

Variations

- Adapt automatically the search zone (Kervrann et al. [KB06])
- Use higher order local approximations (Takeda et al. [TFM07])
- Use different dissimilarity measures (Azzabou et al. [APG07])

NL-Means II

Advantages

- Performance close to “state-of-the-art” methods (in 2005)
- Easy to implement

Limits :

- Consistency requires strong hypotheses : stationary and $\beta-$mixing process (true for textures . . .)
- Searching zone = entire image : too slow in practice and no benefit if $R \geq 21$ for common images
- $\beta \longrightarrow 0$ (temperature) : [BCM05] $\beta = 12\sigma^2$ choice ?

NL-Means II

Advantages

- Performance close to “state-of-the-art” methods (in 2005)
- Easy to implement

Limits :

- Consistency requires strong hypotheses : stationary and $\beta-$ mixing process (true for textures . . .)
- Searching zone = entire image : too slow in practice and no benefit if $R \geq 21$ for common images
- $\beta \longrightarrow 0$ (temperature) : [BCM05] $\beta = 12\sigma^2$ choice ?

NL-Means Interpretation

Intuitive explanation

- Smoothing on the patch manifold

Optimized local kernel

- NL-Means induces a local kernel adapted to the geometry

NL-Means Interpretation

Intuitive explanation

- Smoothing on the patch manifold

Optimized local kernel

- NL-Means induces a local kernel adapted to the geometry

A best local kernel ?

- Can we compare the NL-Means to the best local kernel :

$$\mathbb{E}(\|f-\hat{f}\|^2) \leq C \arg\min_{\theta} \sum_i |f(i) - \sum_k \theta_{i,k} f(k)|^2 + N^2\sigma^2\|\theta\|^2 \ ?$$

NL-Means Interpretation

Intuitive explanation

- Smoothing on the patch manifold

Optimized local kernel

- NL-Means induces a local kernel adapted to the geometry

A best local kernel ?

- Can we compare the NL-Means to the best local kernel :

$$\mathbb{E}(\|f-\hat{f}\|^2) \leq C \arg\min_{\theta} \sum_i |f(i) - \sum_k \theta_{i,k} f(k)|^2 + N^2\sigma^2\|\theta\|^2 \ ?$$

Statistical Aggregation

Model and preliminary estimators

- $Y = f + \sigma W$ of size $N \times N$.
- $\{P_k\}$ set of M preliminary estimators of f (obtained independently).

Aggregation

- Estimate f as a weighted average $\hat{f} = P_\theta = \sum_k \theta_k P_k$
- Aggregation procedure : way to choose θ_k from Y.

Statistical Aggregation

Model and preliminary estimators

- $Y = f + \sigma W$ of size $N \times N$.
- $\{P_k\}$ set of M preliminary estimators of f (obtained independently).

Aggregation

- Estimate f as a weighted average $\hat{f} = P_\theta = \sum_k \theta_k P_k$
- Aggregation procedure : way to choose θ_k from Y.

Oracle Inequality

- Typical result : "Best" aggregation amongst a class $\Theta \subset \mathbb{R}^M$,

$$\mathbb{E}(\|f - \hat{f}\|^2) \leq C \inf_{\theta \in \Theta} \|f - P_\theta\|^2 + \mathsf{V}(\theta, \sigma)$$

- C, Θ and V depend on the procedure.

Statistical Aggregation

Model and preliminary estimators

- $Y = f + \sigma W$ of size $N \times N$.
- $\{P_k\}$ set of M preliminary estimators of f (obtained independently).

Aggregation

- Estimate f as a weighted average $\hat{f} = P_\theta = \sum_k \theta_k P_k$
- Aggregation procedure : way to choose θ_k from Y.

Oracle Inequality

- Typical result : "Best" aggregation amongst a class $\Theta \subset \mathbb{R}^M$,

$$\mathbb{E}(\|f - \hat{f}\|^2) \leq C \inf_{\theta \in \Theta} \|f - P_\theta\|^2 + \mathsf{V}(\theta, \sigma)$$

- C, Θ and V depend on the procedure.

Aggregation PAC-Bayesian

Aggregation PAC-Bayesian

- Specific aggregation procedure based on exponential weights.
- Defined from a prior π on $\mathbb{R}^M$ by $\hat{f} = P_{\theta_\pi}$, with

$$\theta_\pi = \int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\|Y-P_\theta\|^2}}{\int_{\mathbb{R}^M} e^{-\frac{1}{\beta}\|Y-P_{\theta'}\|^2} d\pi(\theta')} \theta \, d\pi(\theta) \, .$$

- $\pi = \frac{1}{M}\sum_k \delta_k \quad \Longrightarrow \quad \hat{f} = \sum_k \frac{e^{-\frac{1}{\beta}\|Y-P_k\|^2}}{\sum_{k'} e^{-\frac{1}{\beta}\|Y-P_{k'}\|^2}} P_k \, .$

Oracle Inequality

- Sharp oracle inequality : if the temperature $\beta \geq 4\sigma^2$,

$$\mathbb{E}(\|f-\hat{f}\|^2) \leq \inf_p \left[\int_{\theta \in \mathbb{R}^M} \|f - P_\theta\|^2 dp(\theta) + \beta \mathcal{K}(p,\pi) \right]$$

$\mathcal{K}(p,\pi)$: Kullback-Leibler divergence, p : measure on $\mathbb{R}^M$

Aggregation PAC-Bayesian

Aggregation PAC-Bayesian

- Specific aggregation procedure based on exponential weights.
- Defined from a prior π on $\mathbb{R}^M$ by $\hat{f} = P_{\theta_\pi}$, with

$$\theta_\pi = \int_{\mathbb{R}^M} \frac{e^{-\frac{1}{\beta}\|Y - P_\theta\|^2}}{\int_{\mathbb{R}^M} e^{-\frac{1}{\beta}\|Y - P_{\theta'}\|^2} d\pi(\theta')} \theta \, d\pi(\theta) \, .$$

- $\pi = \frac{1}{M}\sum_k \delta_k \quad \Longrightarrow \quad \hat{f} = \sum_k \frac{e^{-\frac{1}{\beta}\|Y - P_k\|^2}}{\sum_{k'} e^{-\frac{1}{\beta}\|Y - P_{k'}\|^2}} P_k \, .$

Oracle Inequality

- Sharp oracle inequality : if the temperature $\beta \geq 4\sigma^2$,

$$\mathbb{E}(\|f - \hat{f}\|^2) \leq \inf_p \left[\int_{\theta \in \mathbb{R}^M} \|f - P_\theta\|^2 dp(\theta) + \beta \mathcal{K}(p, \pi) \right]$$

$\mathcal{K}(p, \pi)$: Kullback-Leibler divergence, p : measure on $\mathbb{R}^M$

Prior Choice

Error bound and prior

- $\mathbb{E}(\|f - \hat{f}\|^2) \leq \inf_p \left[\int_{\theta \in \mathbb{R}^M} \|f - P_\theta\|^2 dp(\theta) + \beta \mathcal{K}(p, \pi) \right].$
- Compromise between a localization of p close to the best "oracle" aggregation P_θ and a proximity with the prior π.
- Choose π so that this quantity is small "uniformly"...

Discrete Prior case

- $\pi = \frac{1}{M} \sum_k \delta_k$ gives $\mathbb{E}(\|f - \hat{f}\|^2) \leq \inf_k \|f - P_k\|^2 + \beta \log M$.
- As good as the best preliminary estimator...

Prior Choice

Error bound and prior

- $\mathbb{E}(\|f - \hat{f}\|^2) \leq \inf_p \left[\int_{\theta \in \mathbb{R}^M} \|f - P_\theta\|^2 dp(\theta) + \beta \mathcal{K}(p, \pi) \right].$
- Compromise between a localization of p close to the best "oracle" aggregation P_θ and a proximity with the prior π.
- Choose π so that this quantity is small "uniformly"...

Discrete Prior case

- $\pi = \frac{1}{M} \sum_k \delta_k$ gives $\mathbb{E}(\|f - \hat{f}\|^2) \leq \inf_k \|f - P_k\|^2 + \beta \log M.$
- As good as the best preliminary estimator...

Sparsifying Prior

- π : i.i.d. Student (Dalalyan et al. [DT09]) or Gaussian mixture
- Bound : $\mathbb{E}(\|f - \hat{f}\|^2) \leq \inf_{\theta \in \mathbb{R}^M} \|f - P_\theta\|^2 + C\beta \|\theta\|_0 \log M.$
- As good as the best "sparse aggregation'...

Prior Choice

Error bound and prior

- $\mathbb{E}(\|f-\hat{f}\|^2) \leq \inf_p \left[\int_{\theta \in \mathbb{R}^M} \|f - P_\theta\|^2 dp(\theta) + \beta \mathcal{K}(p, \pi) \right]$.
- Compromise between a localization of p close to the best "oracle" aggregation P_θ and a proximity with the prior π.
- Choose π so that this quantity is small "uniformly"...

Discrete Prior case

- $\pi = \frac{1}{M} \sum_k \delta_k$ gives $\mathbb{E}(\|f-\hat{f}\|^2) \leq \inf_k \|f - P_k\|^2 + \beta \log M$.
- As good as the best preliminary estimator...

Sparsifying Prior

- π : i.i.d. Student (Dalalyan et al. [DT09]) or Gaussian mixture
- Bound : $\mathbb{E}(\|f-\hat{f}\|^2) \leq \inf_{\theta \in \mathbb{R}^M} \|f - P_\theta\|^2 + C\beta \|\theta\|_0 \log M$.
- As good as the best "sparse aggregation'...

Patch based aggregation

Patches as preliminary estimators

- Use the patches as preliminary estimators $P_i = P(Y)(i)$
- Only issue : not independent with the observation $P(Y)(i_0)$.

Theorem

- Same flavor than for regular aggregation :

$$\mathbb{E}(\|P(f)(i) - P(\hat{f})(i)\|^2)$$
$$\leq \inf_p \int_{\theta \in \mathbb{R}^M} \left(\|P(f)(i) - P_\theta\|^2 + N^2\sigma^2\|\theta\|^2 \right) dp(\theta) + \beta\mathcal{K}(p, \pi)$$

Patch based aggregation

Patches as preliminary estimators

- Use the patches as preliminary estimators $P_i = P(Y)(i)$
- Only issue : not independent with the observation $P(Y)(i_0)$.

Theorem

- Same flavor than for regular aggregation :

$$\mathbb{E}(\|P(f)(i) - P(\hat{f})(i)\|^2) \leq \inf_p \int_{\theta \in \mathbb{R}^M} \left(\|P(f)(i) - P_\theta\|^2 + N^2\sigma^2\|\theta\|^2 \right) dp(\theta) + \beta\mathcal{K}(p,\pi)$$

Patch based aggregation

Patches as preliminary estimators

- Use the patches as preliminary estimators $P_i = P(Y)(i)$
- Only issue : not independent with the observation $P(Y)(i_0)$.

Theorem ?

- Same flavor than for regular aggregation :

$$\mathbb{E}(\|P(f)(i) - P(\hat{f})(i)\|^2)$$
$$\leq \inf_p \int_{\theta \in \mathbb{R}^M} \left(\|P(f)(i) - P_\theta\|^2 + N^2\sigma^2\|\theta\|^2 \right) dp(\theta) + \beta\mathcal{K}(p, \pi)$$

- Proof requires either some splitting or some more homework...

Patch based priors

- Discrete Uniform (NL-Means) : selection ...
- Sparsifying (Student, Gaussian mixture) : sparse kernel optimization !

Patch based aggregation

Patches as preliminary estimators

- Use the patches as preliminary estimators $P_i = P(Y)(i)$
- Only issue : not independent with the observation $P(Y)(i_0)$.

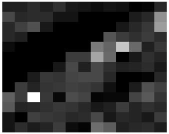
Theorem ?

- Same flavor than for regular aggregation :

$$\mathbb{E}(\|P(f)(i) - P(\hat{f})(i)\|^2) \leq \inf_p \int_{\theta \in \mathbb{R}^M} \left(\|P(f)(i) - P_\theta\|^2 + N^2\sigma^2\|\theta\|^2 \right) dp(\theta) + \beta\mathcal{K}(p, \pi)$$

- Proof requires either some splitting or some more homework...

Patch based priors



- Discrete Uniform (NL-Means) : selection ...
- Sparsifying (Student, Gaussian mixture) : sparse kernel optimization !

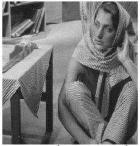
PAC-Bayesian estimate and Monte Carlo method

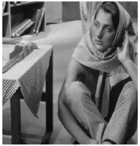
The PAC-Bayesian estimate

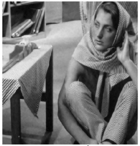
- High dimensional integral similar to some integrals appearing in the Bayesian framework...
- Important Issue !
- Monte Carlo method based on a Langevin diffusion equation
- Approximate values only... but sufficient precision
- Some convergence issues still under investigation
- Patch preselection seems to help...

Numerical Results (PSNR)


Original


Noisy (28.13 dB)


NL Means (31.19 dB)


PAC-Bayesien (32.80 dB)

Experimental setting

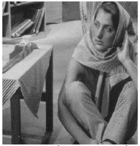
- Comparison with classic NL-Means with $\beta = 12\sigma^2$
- PAC-Bayesian aggregation with Student prior

Results

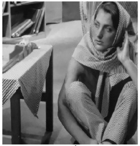
- Results on par with NL-Means
- Room for improvement.

Numerical Results (PSNR)

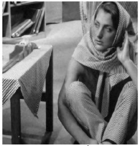
Original



Noisy (28.13 dB)



NL Means (31.19 dB)



PAC-Bayesien (32.80 dB)

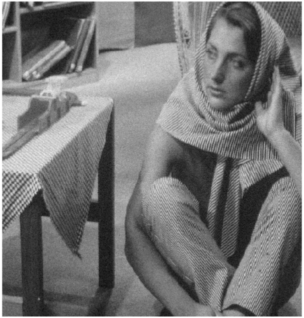
Experimental setting

- Comparison with classic NL-Means with $\beta = 12\sigma^2$
- PAC-Bayesian aggregation with Student prior

Results

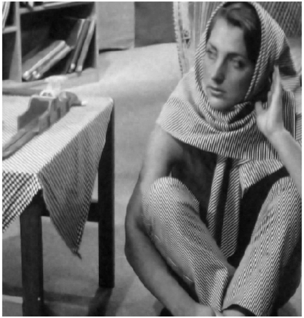
- Results on par with NL-Means
- Room for improvement.

Original

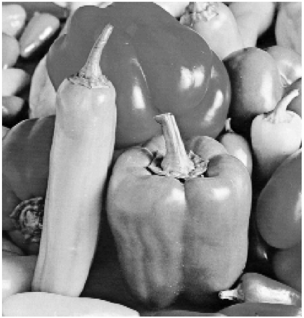


Noisy (28.13 dB)

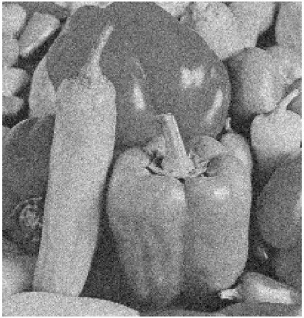
NL Means (31.19 dB)



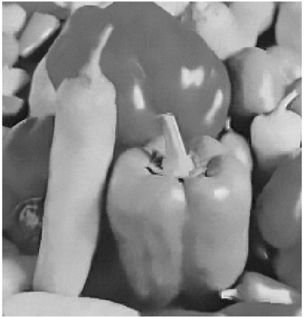
PAC-Bayesien (32.20 dB)



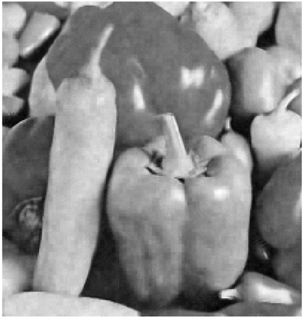
Original



Noisy (22.12 dB)

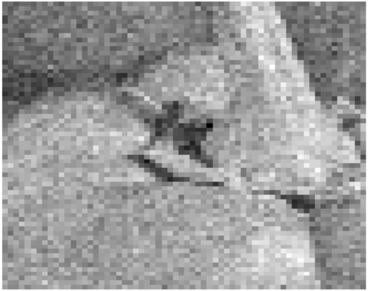


NL Means (29.59 dB)



PAC-Bayesien (29.46 dB)

Original



Noisy (22.21 dB)

NL Means (24.23dB)

PAC-Bayesien (26.96 dB)

Conclusion

A novel aggregation point of view on the NL-Means

- New look on the exponential weights and the L_2 patch dissimilarity measure
- Stein Unbiased Risk Estimate : a tool in proofs leading to a new approach for the central patch weight
- Proposition of a new aggregation procedure which is on par with NL-Means but with (some) theoretical control
- Framework adaptable for other dictionaries

A huge to-do list

- Extend the theorem to the fully dependent case
- Choice of the best prior
- Accelerated convergence of the Monte Carlo chain
- ...

Conclusion

A novel aggregation point of view on the NL-Means

- New look on the exponential weights and the L_2 patch dissimilarity measure
- Stein Unbiased Risk Estimate : a tool in proofs leading to a new approach for the central patch weight
- Proposition of a new aggregation procedure which is on par with NL-Means but with (some) theoretical control
- Framework adaptable for other dictionaries

A huge to-do list

- Extend the theorem to the fully dependent case
- Choice of the best prior
- Accelerated convergence of the Monte Carlo chain
- ...

References

- Noura Azzabou, Nikos Paragios, and Frederic Guichard.
  Image denoising based on adapted dictionary computation.
  In *ICIP (3)*, pages 109–112. IEEE, 2007.
- Antoni Buades, Bartomeu Coll, and Jean-Michel Morel.
  A review of image denoising algorithms, with a new one.
  Multiscale Model. Simul., 4(2) :490–530 (electronic), 2005.
- 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.
- Charles Kervrann and Jérôme Boulanger.
  Optimal spatial adaptation for patch-based image denoising.
  IEEETIP, 15(10) :'2866–2878, 2006.
- H. Takeda, S. Farsiu, and P. Milanfar.
  Kernel regression for image processing and reconstruction.
  IEEE Transactions on Image Processing, 16(2) :349–366, 2007.