16th European Signal Processing Conference (EUSIPCO 2008), Lausanne, Switzerland, August 25-29, 2008, copyright by EURASIP

TOTAL VARIATION DENOISING USING POSTERIOR EXPECTATION

Cécile Louchet and Lionel Moisan

Université Paris Descartes, MAP5 (CNRS UMR 8145)
45 rue des Saints-Pères 75006 Paris, France
email: {louchet,moisan}@math-info.univ-paris5.fr

ABSTRACT

Total Variation image denoising, generally formulated in a variational setting, can be seen as a Maximum A Posteriori (MAP) Bayesian estimate relying on a simple explicit image prior. In this formulation, the denoised image is the most likely image of the posterior distribution, which favors regularity and produces staircasing artifacts: in regions where smooth-varying intensities would be expected, constant zones appear separated by artificial boundaries. In this paper, we propose to use the Least Square Error (LSE) criterion instead of the MAP. This leads to a new denoising method called TV-LSE, that produces more realistic images by computing the expectation of the posterior distribution. We describe a Monte-Carlo Markov Chain algorithm based on Metropolis scheme, and provide an efficient convergence criterion. We also discuss the properties of TV-LSE, and show in particular that it does not suffer from the staircasing effect.

1. INTRODUCTION

Image denoising based on Total Variation (TV) was first proposed by Rudin, Osher and Fatemi (ROF) in 1992 [11]. Since then, the TV criterion was found to be very efficient in many other image processing tasks, including deblurring, interpolation, spectrum extrapolation, inpainting, decompression, etc. (see, e.g., [1, 8]). One reason for this is that the TV functional enforces a certain notion of regularity that is well suited to images: it puts a strong penalization on oscillations and random fluctuations, but allows discontinuities at the same time. This is an interesting property, because true images generally present discontinuities in the intensity map that are caused by occluding parts in the scene. However, because the ROF method is based on TV minimization, it tends to produce denoised images that present unnatural local configurations that permit to achieve a small overall TV. This is known as the *staircasing effect* [2, 10]: in ROF denoised images, one may often observe constant regions delimited by artificial discontinuities, as in Fig. 3 (middle row, left image).

As shown in [9], the staircasing effect of TV denoising is due to the fact that the total variation is not differentiable (and, as proven in [4], to the fact that noisy images are almost everywhere discontinuous). Smooth approximations and variants of the total variation functional [2, 5, 6] manage to avoid the staircasing effect, but they lose the nice geometrical properties of the total variation, in particular the co-area formula that connects the total variation measure with the image geometry via the level-set decomposition. In [3], a solution to the staircasing effect is proposed in the case of neighborhood filters, but it does not apply for variational formulations like TV denoising.

In this paper, we propose to use the TV criterion in a different framework, in order to avoid the staircasing effect while keeping the efficiency of the TV measure. In Section 2, we recall the Bayesian MAP interpretation of ROF denoising, and introduce a new denoising filter called TV-LSE, defined as the image estimate achieving the least square error Bayesian risk (like, e.g., the Wiener filter). A Monte-Carlo Markov Chain (MCMC) Metropolis sampler is then proposed in Section 3 to compute the posterior expectation required by this new filter, and a convergence criterion is given and analyzed. In Section 4, we discuss the properties of TV-LSE denoising and in particular its difference with TV-MAP (ROF) denoising. We show that unlike the latter, TV-LSE denoising does not suffer from the staircasing effect, and produce more realistic images while keeping good denoising efficiency.

2. BAYESIAN FORMULATION OF TV DENOISING

Let $u : \Omega \to \mathbb{R}$ be a discrete gray-level image defined on a rectangular domain $\Omega \subset \mathbb{Z}^2$. The discrete Total Variation of u is defined by

$$TV(u) = \sum_{(x,y)\in\Omega} |Du(x,y)|, \tag{1}$$

where $|Du(x,y)|$ is a discrete approximation of the gradient norm of u in (x,y). In this paper, we shall use the usual Euclidean norm in $\mathbb{R}^2$ and the simplest possible approximation of the gradient vector, given by

$$Du(x,y) = \begin{pmatrix} u(x+1,y)-u(x,y) \\ u(x,y+1)-u(x,y) \end{pmatrix}.$$

(with the convention that differences involving pixels outside Ω are zero). Given a (noisy) image u_0, the ROF method proposes to compute the unique image u that minimizes

$$E_\lambda(u) = \|u-u_0\|^2 + \lambda TV(u), \tag{2}$$

where $\|\cdot\|$ is the classical L^2 norm on images and λ is an hyperparameter that controls the level of denoising. This energy-minimization formulation can be translated into a Bayesian framework: let us consider, for $\beta > 0$, the prior density function

$$p_\beta(u) = \frac{1}{Z_\beta} e^{-\beta TV(u)}, \quad \text{where} \quad Z_\beta = \int_{\mathscr{E}_0} e^{-\beta TV(u)}\, du$$

$$\text{and} \quad \forall \mu \in \mathbb{R}, \quad \mathscr{E}_\mu = \left\{ u \in \mathbb{R}^\Omega,\ \sum_{\mathbf{x}\in\Omega} u(\mathbf{x}) = \mu|\Omega| \right\}.$$

The function p_β is a probability density function on each set $\mathscr{E}_\mu$, that can be used as a Bayesian prior to estimate the best

denoised image. If we assume that the noise is additive and Gaussian, i.e., that $u_0 = u + N$ where N is a Gaussian white noise with zero mean and variance σ^2, then from Bayes formula we can derive the posterior density

$$p(u|u_0) = \frac{p(u_0|u)p_\beta(u)}{p(u_0)} = \frac{1}{Z}\exp\left(-\frac{E_\lambda(u)}{2\sigma^2}\right), \quad (3)$$

where $\lambda = 2\beta\sigma^2$ and Z is a normalizing factor, depending on u_0 and λ, ensuring that $u \mapsto p(u|u_0)$ is a probability density function on $\mathbb{R}^\Omega$. Hence, the variational formulation ($\arg\min_u E_\lambda(u)$) is equivalent to the Bayesian Maximum A Posteriori (MAP) formulation

$$\hat{u}_{MAP} = \arg\max_u p(u|u_0), \quad (4)$$

which means that the ROF denoising filter simply selects the most likely image u according to the posterior distribution $p(u|u_0)$.

In a certain sense, the most complete denoising information consists in the whole posterior density function itself. However, one generally wants to build from this density a "best estimate" of the true image, according to some criterion. The MAP estimator is the one that minimizes Bayes risk when the cost function is a Dirac delta localized on the true solution. In a sense, it does not represent very well the posterior density function, because it only sees its maximum point (as we can see from (3), all posterior distributions that share the same value of λ yield the same MAP estimator, independently of their "spread" σ). Since $\hat{u}_{MAP}$ minimizes the energy $E_\lambda(u)$, it tends to present some exceptional structures that have a very small contribution to the energy, in particular "flat zones", that is, regions with uniform intensity, causing the well-known *staircasing effect*.

Instead of the hit-or-miss risk function leading to the MAP estimate, we propose to use the Least Square Error (LSE) criterion, that consists in finding the estimate $\hat{u}(u_0)$ that minimizes

$$\mathbb{E}_{u,u_0}\left(\|u - \hat{u}(u_0)\|^2\right) = \int_{\mathbb{R}^\Omega}\int_{\mathscr{E}_\mu} \|u - \hat{u}(u_0)\|^2 p(u,u_0)\,du\,du_0.$$

This minimum is attained by the posterior expectation (conditional mean), that is for

$$\hat{u}_{LSE} := \mathbb{E}(u|u_0) = \int_{u\in\mathbb{R}^\Omega} u p(u|u_0)\,du. \quad (5)$$

Thanks to (3), this can be rewritten

$$\hat{u}_{LSE} = \frac{\displaystyle\int_{\mathbb{R}^\Omega} \exp\left(-\frac{E_\lambda(u)}{2\sigma^2}\right)\cdot u\,du}{\displaystyle\int_{\mathbb{R}^\Omega} \exp\left(-\frac{E_\lambda(u)}{2\sigma^2}\right) du}. \quad (6)$$

3. MCMC ALGORITHM FOR TV-LSE DENOISING

3.1 Principle

TV-LSE denoising requires to evaluate the ratio of integrals arising in (6), each integral concerning thousands of variables (the dimension is the number of pixels). For such high-dimension integrals, only Monte-Carlo methods can be considered. Here we propose to use a Monte-Carlo Markov Chain (MCMC) following Metropolis scheme. Given a positive parameter $\alpha > 0$, let us consider a random chain of images $(\Upsilon_n)_{n\geq 0}$ consisting in an initial (random or deterministic) image Υ_0 and the transition defined by

$$\Upsilon_{n+1} = \begin{cases} \Upsilon_n + \alpha\Delta_n\delta_{\mathbf{X}_n} & \text{if} \quad R_n \geq Z_n, \\ \Upsilon_n & \text{else,} \end{cases}$$

where the random variables $(\Delta_n)_{n\geq 0}$, $(\mathbf{X}_n)_{n\geq 0}$ and $(Z_n)_{n\geq 0}$ are all independent, with $\Delta_n \sim U([-1,1])$, $\mathbf{X}_n \sim U(\Omega)$, $Z_n \sim U([0,1])$, and

$$R_n = \exp\left(-\frac{E_\lambda\left(\Upsilon_n + \alpha\Delta_n\delta_{\mathbf{X}_n}\right) - E_\lambda(\Upsilon_n)}{2\sigma^2}\right)$$

(note that if $E_\lambda\left(\Upsilon_n + \alpha\Delta_n\delta_{\mathbf{X}_n}\right) \leq E_\lambda(\Upsilon_n)$, then $R_n \geq 1$ so that $R_n \geq Z_n$ almost surely). Notice that two successive images Υ_n and Υ_{n+1} of the chain differ by one pixel at most, while Υ_n and $\Upsilon_{n+|\Omega|}$ are much less correlated. This is why we consider in the following the subsampled chain

$$U_n = \Upsilon_{|\Omega|n},$$

even if the results of this section remain true for (Υ_n). U_n is a Metropolis sampler for the posterior distribution, so it converges in law towards this distribution. This provides a way to estimate the TV-LSE denoising filter, as shown by the following Theorem.

Theorem 1 *For any $\alpha > 0$ and any distribution of U_0, we have almost surely*

$$\frac{1}{n}\sum_{k=1}^{n} U_k \quad \xrightarrow[n\to+\infty]{} \quad \hat{u}_{LSE}.$$

Proof — To simplify, the proof is given here in the case of a countable image space. Let l be a positive real number (quantization step), we assume that Δ_n follows the uniform distribution on $\{k/l,\ -l \leq k \leq l\}$, so that the discrete image space (state space) is $E = (\alpha\mathbb{Z}/l)^\Omega$. The classical ergodic Theorem on Markov chains [7] states that if the Markov chain is irreducible, has a stationary distribution π, and $h : E \to E$ satisfies $\int_E |h(u)|\,d\pi(u) < \infty$, then

$$\frac{1}{n}\sum_{k=1}^{n} h(U_k) \xrightarrow[n\to+\infty]{} \int_E h(u)d\pi(u) \text{ a.s. and in } L^1.$$

1) The chain (U_n) is irreducible because if u and u' are two images of E, then $\mathbb{P}(U_n = u'|U_0 = u) > 0$ for all $n \geq \|u' - u\|_\infty/\alpha$, so that u and u' communicate.

2) Let us write $\pi(u) = \frac{1}{Z}\exp(-\frac{E_\lambda(u)}{2\sigma^2})$ the posterior distribution. To prove that π is stationary for the subsampled chain (U_n), it is sufficient to prove that it is stationary for (Υ_n). The transition kernel P of Υ_n can be decomposed into $P_{u,u'} = q(u,u')e^{-\frac{(E_\lambda(u')-E_\lambda(u))_+}{2\sigma^2}}$, where $(x)_+ = \max(0,x)$ is the positive part of x, and

$$q(u,u') = \frac{1}{|\Omega|}\sum_{x\in\Omega}\frac{1}{2l+1}\mathbf{1}_{|u'(x)-u(x)|\leq\alpha, u'(y)=u(y)\forall y\neq x}(u,u')$$

is the instrumental distribution. If $\pi(u) \geq \pi(u')$, then $(E_\lambda(u') - E_\lambda(u))_+$ is null, and $\pi(u)P_{u,u'} = \pi(u)q(u,u')$ holds. But q is symmetric, so that $\pi(u)P_{u,u'} = \pi(u)q(u',u) = \pi(u')P_{u',u}$ since $\pi(u') \leq \pi(u)$. Consequently, π is reversible with respect to P, thus stationary for (Υ_n).

3) We conclude by applying the ergodic Theorem to the function $h = Id_E$, which is π-integrable as required. □

3.2 Convergence control

Theorem 1 is a theoretical result ensuring convergence when n tends to infinity, but in practice the real issue is: how large should n be to permit a reasonable approximation of $\hat{u}_{LSE}$? Here the speed of convergence depends on two factors: first, the number of iterations needed by the Markov Chain (U_n) to attain the stationary state; second, the number of iterations needed by the empirical average to estimate reasonably the true expectation. In MCMC simulations, it is common to introduce a "burn-in" phase, during which the random images are generated with the Markov chain but not taken into account in the expectation estimate. It amounts to consider, for $0 \leq b < n$, the partial average

$$S_n^b = \frac{1}{n-b} \sum_{k=b+1}^{n} U_k, \tag{7}$$

from which we compute

$$\mathbb{E}\|S_n^b - \hat{u}_{LSE}\|^2 = \mathbb{E}\|S_n^b - \mathbb{E}S_n^b\|^2 + \|\mathbb{E}S_n^b - \hat{u}_{LSE}\|^2.$$

The first term is the span (trace of the covariance matrix) of the estimator S_n^b (written $\mathrm{Span}(S_n^b)$), which converges to 0 like $A/(n-b)$ (for some constant A) when n tends to infinity [7]. The second term is the (squared) bias, due to the fact that the law of U_n is not exactly the posterior law, but (only) tends to it when $n \to +\infty$ by ergodic Theorem. If we assume that the state space is finite (which is always numerically the case), the convergence of U_n to the posterior distribution is geometric and

$$\forall n \geq 1, \ \ \|\mathbb{E}U_n - \hat{u}_{LSE}\| \leq B\gamma^n, \tag{8}$$

for some $B > 0$ and $0 < \gamma < 1$. From (8), we deduce that

$$\|\mathbb{E}S_n^b - \hat{u}_{LSE}\| \leq \frac{B}{n-b} \frac{\gamma^{b+1}(1-\gamma^{n-b})}{1-\gamma},$$

so that for $B' = \gamma^2 B^2/(1-\gamma)^2$ we have

$$\mathbb{E}\|S_n^b - \hat{u}_{LSE}\|^2 \leq \mathrm{Span}(S_n^b) + \frac{B'\gamma^{2b}}{(n-b)^2}. \tag{9}$$

Hence, S_n^b converges to $\hat{u}_{LSE}$ like $1/\sqrt{n}$ (recall that $\mathrm{Span}(S_n^b) \simeq A/(n-b)$), which is a rather slow convergence, that requires a good stopping criterion. Now let us consider another Markov chain $\tilde{U}_n$ defined like U_n (and independent from it). With obvious notations we have

$$\mathbb{E}\|\tilde{S}_n^b - S_n^b\|^2 = 2\ \mathrm{Span}(S_n^b),$$

and the empirical value $\|\tilde{S}_n^b - S_n^b\|^2$ is a very good approximation of its expectation since the dimension of the image space is very high. Thus, if we manage to choose b large enough to ensure that the bias term (rightmost term) is negligible in (9), then we expect to have

$$\|S_n^b - \hat{u}_{LSE}\| \simeq \frac{e_b}{\sqrt{2}} \quad \text{with} \quad e_b = \|\tilde{S}_n^b - S_n^b\|, \tag{10}$$

and we can use a test like $e_b \leq \varepsilon$ as a stopping criterion. Now how do we select the correct burn-in parameter b? Empirically, it can be observed that the function e_b decreases with b for small values of b, reaches a minimum, then increases with b (see Figure 1). This highlights a competition between the burn-in time b (that should not be too short because $\mathrm{Span}(U_n)$ decreases with n, as can be seen empirically) and the number of samples $(n-b)$ kept for the estimation. Intuitively, the value $\bar{b} = \arg\min_b e_b$ is an interesting compromise for b, since it is very close to the optimal value (see Figure 1).

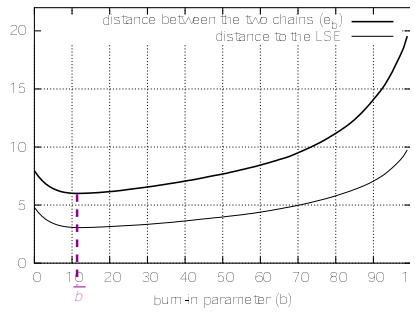

Figure 1: Selection of the burn-in parameter b. As a function of b, the distance $e_b = \|\tilde{S}_n^b - S_n^b\|$ between the two MC estimates (thick line) reaches a minimum value for $b = \bar{b}$. This value is a good choice for the burn-in parameter, because it is very close to the value of b for which the distance from $(S_n^b + \tilde{S}_n^b)/2$ to $\hat{u}_{LSE}$ is minimal (experiment made with $\lambda = 30$ and $\sigma = 10$ on a noisy image).

3.3 Algorithm

The considerations above lead to the following algorithm.

Algorithm 1 TV-LSE algorithm

draw two random images U_0 and $\tilde{U}_0$
$n \leftarrow 0$
repeat
  $n \leftarrow n+1$
  draw U_n and $\tilde{U}_n$ from U_{n-1} and $\tilde{U}_{n-1}$
  compute $\bar{b} = \arg\min_b e_b$
until $e_{\bar{b}} \leq 2\varepsilon$
return $(S_n^{\bar{b}} + \tilde{S}_n^{\bar{b}})/2$

In practice the initial images U_0 and $\tilde{U}_0$ are drawn with i.i.d uniform intensity values in $[0, 256)$. Concerning the partial sums S_n^b, since all images $(U_k)_{1 \leq k \leq n}$ cannot be kept in memory at the same time, we constrain b to belong to a discrete set of values $E = \lfloor \lambda^{\mathbb{N}} \rfloor = \{\lfloor \lambda^p \rfloor, p \in \mathbb{N}\}$ (where $\lfloor \cdot \rfloor$ denotes the lower integer part, and $\lambda = 1.2$ in practice) and to be larger than a fraction of n ($n/6$ in practice). Hence, we simply have to maintain the partial sum S_n^b for $b \in E \cap [n/6, n)$ for all n, and there are at most 10 such values of b for any n (because $-\log(\frac{1}{6})/\log 1.2 \simeq 9.8$).

At the end of the algorithm, we estimate $\hat{u}_{LSE}$ with $(S_n^b + \tilde{S}_n^b)/2$ which is better than either S_n^b or $\tilde{S}_n^b$. Indeed, since S_n^b and $\tilde{S}_n^b$ are independent, we have, with a similar computation as before

$$\mathbb{E}\left\| \frac{S_n^b + \tilde{S}_n^b}{2} - \hat{u}_{LSE} \right\|^2 \leq \frac{1}{2}\mathrm{Span}(S_n^b) + \|\mathbb{E}S_n^b - \hat{u}_{LSE}\|^2.$$

In other terms, by averaging the two chains we maintain the same bias and divide the span by two. If the bias is negligible (it is the case in general when b is chosen large enough), then

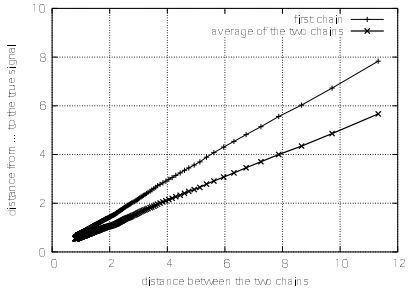

Figure 2: The curve $(e_b, \|S_n^b - \hat{u}_{LSE}\|)_n$ (thick line) and the curve $(e_b, \|(S_n^b + \tilde{S}_n^b)/2 - \hat{u}_{LSE}\|)_n$ (thin line) for a fixed value of b and $\lambda = 30$, $\sigma = 10$. Since the ratio of the two curves is approximately $\sqrt{2}$, the bias is negligible, which suggests that the MCMCs have reached the stationary regime.

we can expect to have $\|(S_n^b + \tilde{S}_n^b)/2 - \hat{u}_{LSE}\| \simeq e_b/2$, which is $\sqrt{2}$ times better than S_n^b or $\tilde{S}_n^b$ alone. This property, which can be checked numerically on Figure 1 (the ratio between the two functions is approximately 2), can be used to check the absence of bias, since averaging the two chains reduce the span but not the bias. Figure 2 corresponds to the same situation as Figure 1, that is $\lambda = 30$ and $\sigma = 10$, and also illustrates the negligibility of the bias in that case (the ratio of the two curves is approximately $\sqrt{2}$).

4. PROPERTIES OF TV-LSE DENOISING

4.1 LSE versus MAP

Whereas the classical TV denoising ($\hat{u}_{MAP}$) only depends on the parameter λ, the TV-LSE denoising depends on two parameters, λ and σ. In the Bayesian framework, σ^2 represents the variance of the noise, which indirectly controls the spread of the posterior distribution. Hence, even if it is natural to choose for σ^2 the (supposedly known) variance of the noise, we can also consider it as an abstract hyperparameter, as is λ in the variational TV denoising framework. First, we investigate the extreme values of σ from a theoretical point of view. We note $\hat{u}_{MAP}(\lambda)$ and $\hat{u}_{LSE}(\lambda,\sigma)$ the MAP and LSE results obtained from a given image u_0.

Theorem 2 *For all $\lambda > 0$, we have*

$$(i) \quad \hat{u}_{LSE}(\lambda,\sigma) \xrightarrow[\sigma\to 0]{} \hat{u}_{MAP}(\lambda),$$
$$(ii) \quad \hat{u}_{LSE}(\lambda,\sigma) \xrightarrow[\sigma\to+\infty]{} u_0.$$

Proof — When σ goes to 0, the unimodal probability distribution $\frac{1}{Z}\exp\left(-\frac{E_\lambda}{2\sigma^2}\right)$ converges to the Dirac distribution in $\hat{u}_{MAP}(\lambda) = \arg\min_u E_\lambda(u)$, whose expectation is $\hat{u}_{MAP}(\lambda)$, which proves (i). For (ii), consider the change of variable $u_0' = \frac{u_0}{\sigma}$ and $u' = \frac{u}{\sigma}$, then

$$\hat{u}_{LSE}(\lambda,\sigma) = \frac{\displaystyle\int_{\mathbb{R}^\Omega} \sigma u' e^{-\frac{1}{2}(\|u'-u_0'\|^2 + \frac{\lambda}{\sigma}TV(u'))} du'}{\displaystyle\int_{\mathbb{R}^\Omega} e^{-\frac{1}{2}(\|u'-u_0'\|^2 + \frac{\lambda}{\sigma}TV(u'))} du'}$$

so that, thanks to Lebesgue's dominated convergence theorem,

$$\hat{u}_{LSE}(\lambda,\sigma) \underset{\sigma\to\infty}{\sim} \sigma \frac{\displaystyle\int_{\mathbb{R}^\Omega} u' e^{-\frac{1}{2}\|u'-u_0'\|^2} du'}{\displaystyle\int_{\mathbb{R}^\Omega} e^{-\frac{1}{2}\|u'-u_0'\|^2} du'} \underset{\sigma\to\infty}{\sim} \sigma u_0' = u_0. \qquad \square$$

Thus, TV-MAP denoising can be seen as a special case of TV-LSE denoising, corresponding to $\sigma = 0$. When σ is very small, the posterior distribution is very concentrated around $\hat{u}_{MAP}$, so that starting the Markov chains with a random image causes a lot of bias, which makes our stopping criterion incapable of guaranteeing the precision ε on $\hat{u}_{LSE}$. We could improve a lot the previous algorithm for small values of σ by choosing to start the Markov chains with $\hat{u}_{MAP}$ (instead of random images), but this is not really worth it, since when σ is small, $\hat{u}_{LSE}$ is very close to $\hat{u}_{MAP}$, and hence not specially interesting. In practice, we never encountered convergence problems with the algorithm described in the previous section as soon as $\sigma \geq \lambda/10$.

A natural idea at this point would be to compare TV-MAP and TV-LSE denoising by keeping a fixed value of λ and making σ vary. This is not very interesting because, as one can see in the experiments, the method noise $\|\hat{u}_{LSE} - u_0\|$ decreases with σ, so that not only the denoising technique is different, but also the "amount of denoising". This is the reason why in the comparison we make, we always choose the denoising parameters (λ for the MAP, λ and σ for the LSE) so that the method noise is fixed. For TV-LSE, this leaves one degree of freedom that allows more or less departure from the TV-MAP model. A systematic exploration of this degree of freedom could be interesting, but for the sake of concision we shall only try to give an insight of TV-LSE denoising abilities by choosing one arbitrary (reasonable) value in the experiments.

Apart from computation time (typically 10 seconds for TV-MAP and 10 minutes for TV-LSE on a 512×512 image), the main difference between TV-MAP and TV-LSE denoising, illustrated on Figure 3, is the ability to TV-LSE to avoid two annoying artifacts of TV-MAP denoising: the staircasing effect and the creation of isolated pixels. These artifacts are even created by TV-MAP in a pure noise image (see Figure 4), which contradicts what could be considered as a basic requirement, that is, that a good denoising method should not create structures in noise (this requirement is very important for satellite image interpretation for example).

4.2 No staircasing effect for TV-LSE

We conclude with a theoretical result stating the absence of staircasing for TV-LSE denoised images.

Theorem 3 *Let u_0 be a random image such that the distribution of u_0 is absolutely continuous with respect to Lebesgue's measure. Let $k,k' \in \Omega$ be neighbor pixels (that is, such that $|k-k'| = 1$). Then the denoised image $\hat{u}_{LSE}$ satisfies*

$$\mathbb{P}_{u_0}\big(\hat{u}_{LSE}(k') = \hat{u}_{LSE}(k)\big) = 0.$$

Sketch of the proof — In this proof the gray value of any image u at pixel k will be denoted by u_k, and $v := u_0$ to simplify notations. Let $g : \mathbb{R} \to \mathbb{R}$ the function defined by

$$g(z) = \frac{\displaystyle\int_{\mathbb{R}^\Omega} (u_{k'} - u_k) \exp\left(-\frac{(u_k - z)^2 + \sum_{l\neq k}(u_l - v_l)^2 + \lambda TV(u)}{2\sigma^2}\right) du}{\displaystyle\int_{\mathbb{R}^\Omega} \exp\left(-\frac{(u_k - z)^2 + \sum_{l\neq k}(u_l - v_l)^2 + \lambda TV(u)}{2\sigma^2}\right) du}. \tag{11}$$

Then $\hat{u}_{LSE,k'} = \hat{u}_{LSE,k}$ is equivalent to $g(v_k) = 0$. With (11), the function g can be extended to a holomorphic $\mathbb{C} \to \mathbb{C}$ mapping, which proves that g is analytic. Now Assume that g is zero everywhere. The numerator of $g(z)$ in (11), written N, can be rewritten as the convolution product $N = G_\sigma * \varphi$, where G_σ is the cen-

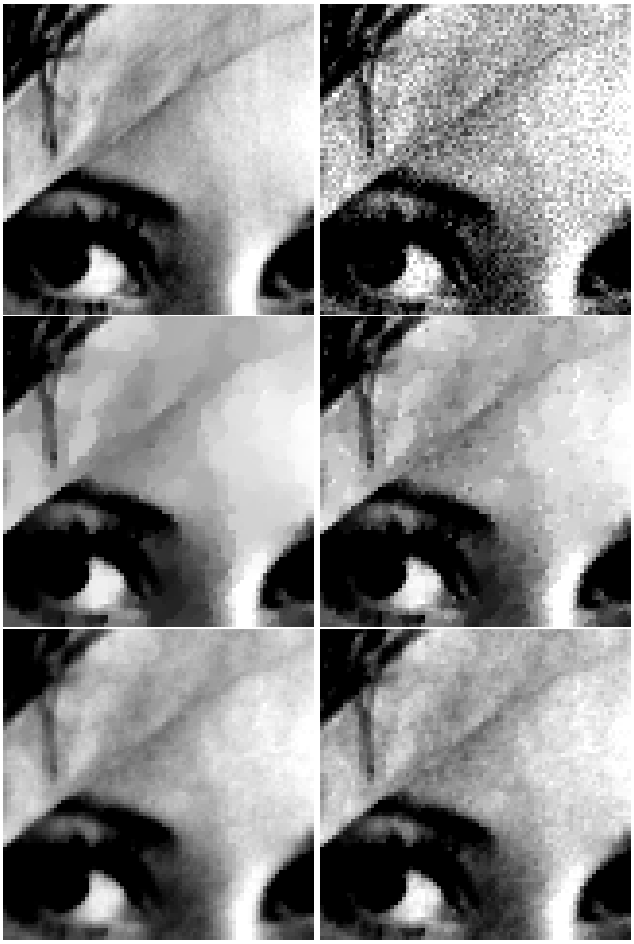

Figure 3: Comparison between classical TV-MAP denoising and the proposed TV-LSE denoising on a part of the classical Lena image (top, left) corrupted by a Gaussian white noise with standard deviation $\sigma = 10$ (top, right). On the middle row, TV-MAP denoising has been applied for two levels of denoising (the level of denoising being measured by the "method noise", that is, the L^2 distance between the noisy image and the result). Left: method noise is 9.6 (to be compared to 10, the actual noise level), obtained with $\lambda = 19.25$; right: method noise is 7.68, obtained with $\lambda = 11.1$. On the bottom row, TV-LSE denoising has been applied for the same levels of denoising (same method noise). Left: method noise is 9.6, obtained with $(\lambda, \sigma) = (50, 20)$; right: method noise is 7.68, obtained with $(\lambda, \sigma) = (25, 15)$. On both TV-MAP images (middle row), the staircasing effect is visible: artificial boundaries are created between extremely flat zones. On the right TV-MAP image, another artifact appears: isolated pixels with extreme intensity values remain. These two artifacts do not arise with the TV-LSE images (bottom row), that look much more natural.

tered Gaussian function of bandwidth σ and φ is a real function defined by $\varphi(x) = \int (u_{k'} - u_k) \exp\left(-\frac{\sum_{l\neq k}(u_l - v_l)^2 + \lambda TV(u^{k,x})}{2\sigma^2}\right) d(u_l)_{l\neq k}$, where $u^{k,x}$ is the image defined by $u_l^{k,x} = u_l$ if $l \neq k$, and $u_k^{k,x} = x$. It can be proven that the discrete formulation of TV (Equation 1) ensures that φ is in L^1, so that by considering Fourier transforms (written $\hat{\cdot}$) we get $\widehat{G_\sigma}(\xi) \cdot \hat{\varphi}(\xi) = 0$ for all $\xi \in \mathbb{R}$. Since $\widehat{G_\sigma}(\xi)$ never vanishes, we deduce that $\hat{\varphi}$ is identically null, and so is φ. But as $\varphi(z)$ can be proved to be negative for large enough z, we have a contradiction, which proves that g cannot be zero everywhere.

Last, since g is analytic and non-identically null, the isolated zero Theorem states that $g^{-1}(\{0\})$ cannot contain any accumulation point. Thus, under the marginal distribution of v_k, the event $(g(v_k) \neq 0)$ a.s. occurs. Since v has been assumed to have a density with respect to Lebesgue measure, this yields

$$\begin{aligned}\mathbb{P}_v(g(v_k) = 0) &= \int_{(v_l)_{l\neq k}} \mathbb{P}_{v_k}(g(v_k) = 0) f(v)\, d(v_l)_{l\neq k} = 0 \\ &= \int_{(v_l)_{l\neq k}} 0 \cdot f(v)\, d(v_l)_{l\neq k} = 0\end{aligned}$$

which concludes the proof. □

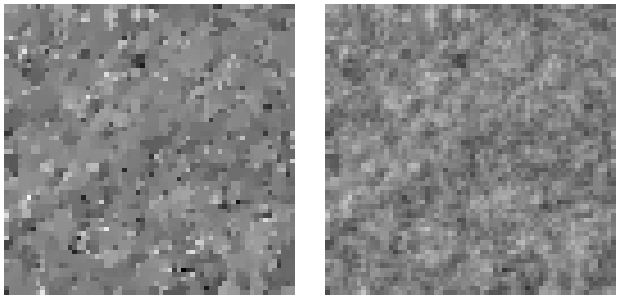

Figure 4: A pure noise image (Gaussian white noise with standard deviation 10) is denoised with the classical TV-MAP (left) and the proposed TV-LSE (right) method, with the same level of method noise (7.33), achieved with $\lambda = 9.37$ for MAP and $(\lambda, \sigma) = (40, 20)$ for LSE. As we can see, TV-MAP denoising creates severe structures in noise: artificial boundaries between artificially flat zones (the staircasing effect), and isolated pixels with extreme values. Like for the Lena image (Figure 3), these artifacts are avoided by the TV-LSE denoising method.

REFERENCES

[1] F. Alter, S. Durand, J. Froment, "Adapted total variation for artifact free decompression of JPEG images", *J. Math. Imaging and Vision* 33(2), pp. 199–211, 2005.

[2] P. Blomgren, T. Chan, P. Mulet, C.K. Wong, "Total variation image restoration: numerical methods and extensions", *Proceedings of Int. Conf. on Image Processing*, pp. 384–387, 1997.

[3] A. Buades, B. Coll, J.-M. Morel, "The staircasing effect in neighborhood filters and its solution", *IEEE Transactions on Image Processing* 15(6), pp. 1499–1505, 2006.

[4] V. Caselles, A. Chambolle, M. Novaga, "The discontinuity set of solutions of the TV denoising problem and some extensions", *Multiscale Modeling and Simulation* 6(3), pp. 879–894, 2007.

[5] T. Chan, A. Marquina, P. Mulet, "High-order total variation-based image restoration", *SIAM Journal of Scientific Computing* 22(2), pp. 503–516, 2000.

[6] T. Chan, S. Esedoglu, P. Mulet, "Image decomposition combining staircase reduction and texture extraction", *J. of Visual Comm. and Image Representation* 18(6), pp. 464–486, 2007.

[7] G. Grimmett, D. Stirzaker, *Probability and Random Processes*, Third Edition, Oxford University Press, 2001.

[8] F. Guichard, F. Malgouyres, "Total variation based interpolation", *Proceedings of Eusipco'98*, vol. 3, pp. 1741-1744, 1998.

[9] M. Nikolova, "Local strong homogeneity of a regularized estimator", *SIAM J. on Applied Math.* 61(2), pp. 633-658, 2000.

[10] M. Nikolova, "Weakly constrained minimization: application to the estimation of images and signals involving constant regions", *J. Math. Imaging and Vision* 21, pp. 155–175, 2004.

[11] L. Rudin, S. Osher, E. Fatemi, "Nonlinear total variation based noise removal algorithms," *Physica D* 60, pp. 259–268, 1992.