Posterior Expectation of the Total Variation model: Properties and Experiments

Cécile Louchet, Lionel Moisan

Posterior Expectation of the Total Variation model: Properties and Experiments

Cécile Louchet

Université d'Orléans
MAPMO, CNRS UMR 6628
Cecile.Louchet@univ-orleans.fr

Lionel Moisan

Université Paris Descartes
MAP5, CNRS UMR 8145
Lionel.Moisan@parisdescartes.fr

Abstract

The Total Variation image (or signal) denoising model is a variational approach that can be interpreted, in a Bayesian framework, as a search for the maximum point of the posterior density (Maximum A Posteriori estimator). This maximization aspect is partly responsible for a restoration bias called "staircasing effect", that is, the outbreak of quasi-constant regions separated by sharp edges in the intensity map. In this paper we study a variant of this model that considers the expectation of the posterior distribution instead of its maximum point. Apart from the least square error optimality, this variant seems to better account for the global properties of the posterior distribution. Theoretical and numerical results are presented, that demonstrate in particular that images denoised with this model do not suffer from the staircasing effect.

Keywords: Image denoising, total variation, Bayesian model, least square estimate, maximum a posteriori, estimation in high dimensional spaces, proximity operators, staircasing effect.

1 Introduction

Total Variation (TV) is probably one of the simplest analytic priors on images that favor smoothness while allowing discontinuities at the same time. Its typical use, introduced in the celebrated Rudin, Osher and Fatemi (ROF) image restoration model [58], consists in solving an inverse problem like $Au = v$ (where v is the observed image, u is the unknown ideal image, and A a given operator) by minimizing the energy

$$E(u) = \|Au - v\|^2 + \lambda TV(u). \tag{1}$$

This energy sets a trade-off between data fidelity (the first term) and data regularity (TV), the relative weight of the latter being specified by the hyperparameter λ. In a continuous formulation, the TV of a gray-level image $u : \mathbb{R}^2 \to \mathbb{R}$ is defined by

$$TV(u) = \inf \left\{ \int_{\mathbb{R}^2} u \operatorname{div} p \;;\; p \in \mathcal{C}_c^\infty(\mathbb{R}^2, \mathbb{R}^2), \|p\|_\infty \leq 1 \right\},$$

which boils down to

$$TV(u) = \int_{\mathbb{R}^2} |Du| \quad \text{with} \quad |Du| = \sqrt{\left(\frac{\partial u}{\partial x}\right)^2 + \left(\frac{\partial u}{\partial y}\right)^2} \tag{2}$$

for smooth images. Depending on the choice of A, Equation (1) can be used for image denoising (A is the identity operator), image deblurring (A is a convolution with a given blur kernel), tomography (A is a

Radon Transform), super-resolution (A is an image subsampling operator), etc. In the last two decades, the TV prior has been used in a large variety of image processing and computer vision applications (and also, of course, for applications that do not concern images): image inpainting [20], interpolation [30], segmentation [18], image quality assessment [8], scale detection [42], cartoon+texture decomposition [4, 5], motion estimation [66], and many others. For a more complete list of applications concerning image processing and the TV model, we invite the interested reader to consult [13, 16] and references therein.

Even if other prior functionals have been proposed (Besov priors, Markov random fields learned on a large bench of images, sparsity priors, fields of experts [57]), TV still frequently appears in non-linear image processing algorithms. A possible explanation for this is the simplicity of the TV operator, and its ability to penalize edges (that is, sharp transitions), but not too much: images which are smooth away from a jump set which is a finite union of smooth curves of finite length, will have a finite TV. Conversely, TV does penalize highly oscillating patterns, noise in particular. Among other reasons that make TV worth studying, we can mention the following:

- The prior model based on TV (or the *median pixel prior*, its discrete counterpart) shows a natural connection with purely discrete Markov models [7, 9].
- If u is a binary image (that is, the characteristic function of some (regular enough) subset S of $\mathbb{R}^2$), then $TV(u)$ is simply the perimeter of S. For a general real-valued image u, this correspondence is generalized thanks to the coarea formula [1]: the idea is to decompose u into nested binary images (corresponding to the level sets of u) and to sum up the infinitesimal contribution of the TV of each binary image. This geometric characterization of TV allows us to interpret the ROF model as a regularization of the level lines of v. If the data-fidelity term $\|u-v\|^2$ is replaced by its L^1-norm counterpart $\|u-v\|_1$ in (1), more suitable in the case of impulse noise, we even have a contrast-invariant transform [24], that processes the level sets of u independently. This nice analytical framework around TV and BV spaces [1] makes it particularly fitted to mathematical image analysis.
- The TV model is simple enough to produce few artifacts, which is important for applications in medical imaging for instance, be it the segmentation of an organ or the analysis of a pathology. This may not be the case for more sophisticated methods like BM3D [22] or dictionary learning methods, where the higher performance comes along with artifacts that are difficult to control and anticipate.

TV is simple and convenient, but it has its own drawbacks. First, textured image parts, that are very oscillatory in general, are highly penalized by TV and often destroyed (or at least strongly attenuated) by the ROF model. Another well-known artifact is the staircasing effect: as it was first noticed in [23], images denoised by the ROF model are piecewise constant, and present transition boundaries that may look completely artificial.

The staircasing effect. It is commonly admitted that the staircasing effect is due to the non-regularity of TV [47, 49, 51], which, in the discrete framework, comes from the singularity of TV at zero gradients. More than that, under several hypotheses [63], the ROF model is equivalent to minimizing an energy like (1), but where the TV (the ℓ^1 norm of gradient) is replaced with the ℓ^0 "norm" of the gradient, hence promoting sparsity for the gradient and favoring piecewise constant images. A way to avoid this staircasing artifact is to regularize the TV operator, as proved by [49]. Among variants that have been proposed [3, 7, 26, 50, 64], some introduce a parameter $\varepsilon > 0$ and replace the term $|Du|$ in (2) by $f_\varepsilon(|Du|)$, where

$$f_\varepsilon(t) = \sqrt{\varepsilon^2 + t^2}, \quad \text{or} \quad f_\varepsilon(t) = \begin{cases} t^2 & \text{if } |t| < \varepsilon \\ \varepsilon^2 & \text{otherwise,} \end{cases} \quad \text{or} \quad f_\varepsilon(t) = \begin{cases} \frac{t^2}{2\varepsilon} + \frac{\varepsilon}{2} & \text{if } |t| < \varepsilon \\ |t| & \text{otherwise,} \end{cases}$$

or f_ε is another even, smooth function that is non-decreasing on $\mathbb{R}^+$. More recently, different authors managed to promote sparsity for higher order derivatives [6, 10, 19, 21, 38], leading to piecewise affine or piecewise polynomial images (hence pushing the staircasing effect to higher orders). In [32], an elegant modification of the TV operator seems to avoid staircasing in denoising and deblurring experiments, but no proof is provided.

All the above-mentioned variants require modifications of TV, or the addition of higher-order terms in the variational model. One contribution of the present paper is to show that the true TV prior is compatible with the avoidance of the staircasing artifact, provided that an appropriate framework is used. Indeed, the ROF model can be reinterpreted in a statistical (Bayesian) framework, where it exactly corresponds to the Maximum A Posteriori (MAP) estimate, which means that the ROF model selects the image that maximizes the probability density function of a certain distribution (the posterior distribution, associated to the TV prior and the data-fidelity term). Several authors [51, 65] pointed out that MAP estimates tend to be very singular with regard to the prior distribution. The staircasing artifact can be considered one of these prior statistics singularities.

In the present work, we propose to keep the statistical framework associated to the ROF model, but to move away from MAP estimation and consider instead the mean of the posterior distribution (rather than its maximum). As in the preliminary work [40], we will denote this approach by TV-LSE, for it reaches the Least Square Error. This kind of approach is also often called MMSE (Minimizer of the Mean Square Error) in the literature, or sometimes CM (Conditional Mean).

LSE estimates versus MAP estimates. LSE estimates have been proposed for a long time in the context of Bayesian image restoration. As early as in 1989, Besag [7] mentioned the possibility of using the LSE estimate instead of MAP in the discrete TV framework (then called *median pixel prior*), as well as the marginal posterior mode and the median estimate. In the case of a TV prior model, LSE is presented in [25] as a favorable alternative to MAP concerning the statistics of the reconstructed image, relying on the example of binary image denoising (the TV model is then equivalent to the Ising model) where MAP provides a non-robust estimate. Lassas, Siltanen and colleagues [33, 35, 34] focus on 1-D signal restoration with a TV prior, and make a comparative study of MAP and LSE at the interface between the discrete and the continuous settings, when the quantization step goes to zero (so that the dimension goes to infinity). They show that in their asymptotic framework, the TV prior may only lead to trivial estimates (MAP equal to 0, LSE equivalent to Gaussian smoothing), and conclude by switching to a Besov prior which behaves properly when the quantization step goes to 0.

MAP estimation, seen as the minimization of an energy, is often preferred to LSE estimation because the computation is made easier and fast by a whole world of energy minimization algorithms, contrary to LSE which requires Monte-Carlo Markov Chain algorithms [28] or Gibbs samplers [27], known to be slow. This computational issue can motivate to use MAP instead of LSE, or, more interestingly, to see a LSE estimate as a MAP estimate in another Bayesian framework, as was done in [29] and [52].

The debate between MAP and LSE goes far beyond algorithmic issues, as the literature, mostly on learned prior Markov random fields, testifies. LSE estimates, regarding [52, 57, 61], seem to recover the prior statistics in a better way than MAP estimates. But in [53], it is argued that the prior learning method (maximum margin principle or maximum likelihood) has to be connected to the estimation function: maximum likelihood seem to perform better while associated to a LSE estimator, but learning with a maximum margin principle seems to perform even better while associated to a MAP estimator.

Since the preliminary work [40] in 2008, several people have taken an interest in TV-LSE. Fadili and Chambolle [32], Lefkimmiatis, Bourquard, and Unser [38], and Salmon [60] mention the TV-LSE model for its ability to naturally remove staircasing artifacts. In the conclusion of [44], Mirebeau and Cohen propose a TV-LSE-like approach to denoise images using anisotropic smoothness features, an interesting counterpart to Total Variation, arguing that LSE is able to deal with non-convex functionals. Chaari et al. propose LSE estimates for a frame-based Bayesian denoising task [14] and for a parameter estimation task [15], where a TV prior is used jointly with a prior on the frame coefficients; the abundant numerical experiments show that the proposed method compares favorably with the MAP estimate. In the Handbook chapter [13], Caselles, Chambolle and Novaga dedicate a section to TV-LSE.

Outline of the paper. The paper is organized as follows. In Section 2 we recall the Bayesian point of view on the ROF model and motivate the LSE approach using measure concentration arguments. In Section 3 we analyze the proposed TV-LSE estimator in a finite dimensional framework (finite number of pixels, but real-valued images). Simple invariance and convergence properties are first given in Section 3.1 and 3.2. Then in Section 3.3, a deeper insight is developed where the TV-LSE denoiser is viewed as the gradient of a convex function, which allows us to prove, using convex duality tools, that TV-LSE avoids the constant regions of the staircasing effect while allowing the restoration of sharp edges. We also interpret the TV-LSE denoiser as a MAP estimate, whose prior is carefully analyzed. In Section 4, we give numerical experiments on image denoising, showing that the TV-LSE offers an interesting compromise between blur and staircasing, and generally gives rise to more natural images than ROF. We then conclude in Section 5.

2 From ROF to TV-LSE: Bayes TV-based models

2.1 ROF Bayesian interpretation

Let $u : \Omega \to \mathbb{R}$ be a discrete gray-level image defined on a finite rectangular domain $\Omega \subset \mathbb{Z}^2$, that associates with each pixel $\mathbf{x} = (x, y) \in \Omega$ the gray level $u(\mathbf{x})$. The (discrete) Total Variation of the image u is defined by

$$TV(u) = \sum_{\mathbf{x}\in\Omega} |Du(\mathbf{x})|, \tag{3}$$

where $|Du(\mathbf{x})|$ is a discrete scheme used to estimate the gradient norm of u at point $\mathbf{x}$. In the sequel we shall consider either the ℓ^1 or the ℓ^2 norm on $\mathbb{R}^2$, associated with the simplest possible approximation of the gradient vector, given by

$$Du(x, y) = \begin{pmatrix} u(x+1, y) - u(x, y) \\ u(x, y+1) - u(x, y) \end{pmatrix} \tag{4}$$

(note that all the results of this paper hold for a large variety of discrete TV operators, see Appendix A.1). Concerning boundary conditions, we shall use the convention that differences involving pixels outside the domain Ω are zero. Given a (noisy) image v, the ROF method proposes to select the unique image u minimizing the energy

$$E_{v,\lambda}(u) = \|u - v\|^2 + \lambda TV(u), \tag{5}$$

where $\|\cdot\|$ is the classical L^2-norm on images, and λ is a hyperparameter which controls the denoising level. This formulation as energy minimizer can be transposed in a Bayesian framework. Indeed, for $\beta > 0$ and $\mu \in \mathbb{R}$, let us consider the probability density function (p.d.f.)

$$p_\beta(u) = \frac{1}{Z_\beta}\, e^{-\beta TV(u)}, \quad \text{where} \quad Z_\beta = \int_{\mathcal{E}_\mu} e^{-\beta TV(u)}\, du, \tag{6}$$

$$\text{and} \quad \forall \mu \in \mathbb{R}, \quad \mathcal{E}_\mu = \left\{u \in \mathbb{R}^\Omega,\ \bar{u} = \mu\right\} \qquad \text{with} \qquad \bar{u} = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega} u(\mathbf{x}).$$

Let us now suppose that instead of u, we observe the noisy image $v = u + N$, where N is a white Gaussian noise with zero mean and with variance σ^2. Applying Bayes' rule with prior distribution p_β leads to the following posterior p.d.f.

$$p(u|v) = \frac{p(v|u)p_\beta(u)}{p(v)} = \frac{1}{Z}\exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2}\right), \tag{7}$$

where $\lambda = 2\beta\sigma^2$ and Z is a normalizing constant depending on v and λ only, ensuring that $u \mapsto p(u|v)$ remains a p.d.f. on $\mathbb{R}^\Omega$. Hence, the variational formulation ($\arg\min_u E_{v,\lambda}(u)$) is equivalent to a Bayesian formulation in terms of maximum a posteriori (MAP)

$$\hat{u}_{\text{MAP}} = \arg\max_{u\in\mathcal{E}_\mu} p(u|v). \tag{8}$$

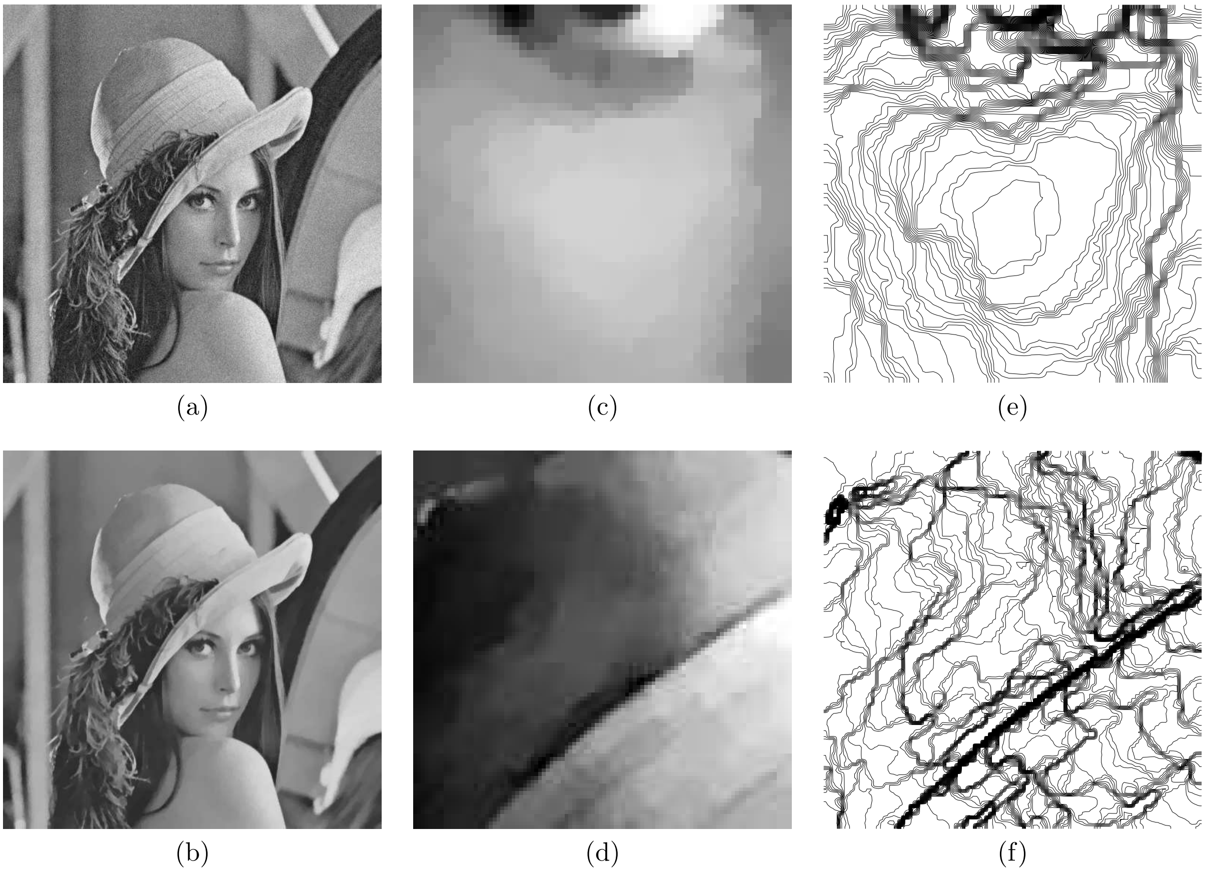


Figure 1: **The staircasing effect.** A noisy version (a) of Lena image (white Gaussian noise with standard deviation $\sigma = 10$) is denoised by TV minimization with $\lambda = 30$ (b). The details (c) and (d) of (b) reveal the so-called *staircasing effect*: TV minimization tends to create smooth regions separated by spurious edges. This effect clearly appears on the level lines (e) and (f) of images (c) and (d): most level lines (here computed using a bilinear interpolation) tend to be concentrated along spurious edges.

This means that ROF denoising amounts to select the most probable image under the posterior probability defined by $p(u|v)$. Notice that the constraint $u \in \mathcal{E}_\mu$, that was imposed to obtain a proper (that is, integrable) prior p.d.f. p_β, can be dropped out when $\mu = \bar{v}$, since this leaves the MAP estimate unchanged [2].

In a certain sense, the most complete information is given by the whole posterior distribution function. However, for obvious practical reasons, one generally seeks an “optimal” estimate of the original image built from the posterior distribution, with respect to a certain criterion. The MAP estimate is obtained by minimizing Bayes risk, when the associated cost function is a Dirac mass located on the true solution. In a certain sense, this estimator is not very representative of the posterior distribution, since it only “sees” its maximum; in particular, as shows (7), the solution does not depend on σ, which measures the “spread” of the posterior distribution. As $\hat{u}_{MAP}$ minimizes the energy $E_{v,\lambda}(u)$, it tends to concentrate certain exceptional structures which are cheap in energy, in particular regions of constant intensity, leading to the well-known *staircasing effect* (see Figure 1).

2.2 The staircasing effect

Several authors mathematically proved the existence of this staircasing effect, in the one-dimensional continuous case [55] as well as in the two-dimensional discrete [46, 47, 49, 50] and continuous [31] cases. Namely in [49], Total Variation is viewed as a particular case of regularization terms J that combine linear operators G_i (typically, finite differences) and a function φ that is non-differentiable in zero (and not necessarily convex):

Proposition 1 *[49] If $\mathcal{X}(v) = \arg\min_u \|u-v\|^2 + \lambda J(u)$ where $J(u) = \sum_{i=1}^{r} \varphi_i(G_i u)$, with G_i linear and φ a non-differentiable at 0 potential function, then there exists a neighborhood V_L of v for which*

$$\forall v' \in V_L, \qquad \{i \mid G_i(\mathcal{X}(v')) = 0\} = \{i \mid G_i(\mathcal{X}(v)) = 0\}. \tag{9}$$

In particular, the case where J is a Total Variation operator fulfills the conditions of Proposition 1, since taking the G_i's equal to gray-level differences across neighboring pixels as in (4) makes the associated function φ (the ℓ^1 or the ℓ^2 norm on $\mathbb{R}^2$) not differentiable at 0. In this case, the set $\{i \mid G_i(\mathcal{X}(v)) = 0\}$ contains the constant regions of the denoised image, and Proposition 1 tells that these constant regions have a certain stability with respect to perturbations of the observed data v. This gives a first theoretical explanation of the staircasing effect.

Let us also cite the recent work of Caselles, Chambolle and Novaga [12] where staircasing is studied not from the point of view of constant regions, but in terms of discontinuities. An interesting property concerning the jump set of the reconstructed image in the continuous framework is proved, which could suggest that staircasing is only due to a bad quantization of the total variation. The (approximate) jump set of a continuous image u is defined as the set of points $\mathbf{x} \in \mathbb{R}^2$ satisfying

$$\exists u^+(\mathbf{x}) \neq u^-(\mathbf{x}),\ \exists \nu_u(\mathbf{x}) \in \mathbb{R}^2,\ \begin{cases} |\nu_u(\mathbf{x})| = 1, \\ \lim\limits_{\rho \downarrow 0} \dfrac{\int_{B_\rho^+(x,\nu_u(x))} |u(\mathbf{y}) - u^+(\mathbf{x})|\, d\mathbf{y}}{\int_{B_\rho^+(x,\nu_u(x))} d\mathbf{y}} = 0, \\ \lim\limits_{\rho \downarrow 0} \dfrac{\int_{B_\rho^-(x,\nu_u(x))} |u(\mathbf{y}) - u^-(\mathbf{x})|\, d\mathbf{y}}{\int_{B_\rho^-(x,\nu_u(x))} d\mathbf{y}} = 0, \end{cases}$$

where

$$B_\rho^+(x, \nu_u(\mathbf{x})) = \{\mathbf{y} \mid \|\mathbf{y} - \mathbf{x}\| < \rho\,,\ \langle \mathbf{y} - \mathbf{x}, \nu_u(\mathbf{x})\rangle > 0\},$$

and $B_\rho^-(\mathbf{x}, \nu_u(\mathbf{x}))$ is the same with a negative inner product. Intuitively, the jump set of an image u is the set of points where u can be locally described as a two-dimensional Heaviside function, which corresponds to regular edges. It is shown that if the datum image v has bounded variation, then the jump set of the solution $\hat{u}$ to the continuous TV minimization problem is contained within the jump set of v. In other words, TV minimization does not create edges which did not already exist in v. This would contradict some kind of staircasing effect (the discontinuity part), if we forgot that v is generally noisy and then the jump set contains almost every point of the domain.

2.3 Concentration of the posterior distribution

Another distortion induced by the MAP approach comes from the high dimension of the problem. Indeed, the MAP estimate only depends on the location of the mode, but not on the probability mass that this mode contains [59], and this difference may become huge in high-dimensional spaces. Let us illustrate this with a simple example. If n is a positive integer and X is a random vector distributed as $\mathcal{N}(0, \sigma^2 I_n)$ (centered normal distribution with covariance matrix $\sigma^2 I_n$, I_n being the n-dimensional identity matrix), then by applying the Bienaymé-Tchebychev inequality to the random variable $|X|^2 = \sum_{i=1}^{n} X_i^2$ (which follows a $\sigma^2\chi^2(n)$ distribution), we obtain

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|\frac{1}{n}|X|^2 - \sigma^2\right| > \varepsilon\right) \leq \frac{2\sigma^4}{n\varepsilon^2}, \tag{10}$$

and the right-hand term decreases towards 0 when the dimension n grows to infinity. In this example, the mode of X is located in 0, but when n goes to $+\infty$ all the mass of the distribution "concentrates" onto the sphere centered in 0 with radius $n\sigma$, and therefore goes away from the mode. This kind of situation is quite common in high dimension. A similar example is the case of the uniform distribution on the unit ball, whose mass concentrates in an arbitrarily small neighborhood of the unit sphere when the dimension grows. Hence, the MAP estimate may be, especially in high dimension, a very special image which properties may strongly differ from those of typical samples of the posterior. This remark particularly makes sense for images, whose typical dimension can be $n = 10^6$ or more.

In our image denoising problem, we should deal with the posterior probability $\boldsymbol{\pi}$ associated to the p.d.f. $\pi(u) = p(u|v)$ (see (7)), which is a log-concave Gibbs field. For such probability distributions, we can have concentration results similar to (10), but they require more sophisticated tools [37].

A key notion to study the concentrating power of a probability distribution $\boldsymbol{\pi}$ on $\mathcal{X}$ is the concentration function $\alpha_{\boldsymbol{\pi}}(r)$ [37], defined for each $r > 0$ by

$$\alpha_{\boldsymbol{\pi}}(r) = \sup\left\{1 - \boldsymbol{\pi}(A_r),\ A \text{ Borel set of } \mathbb{R}^\Omega, \boldsymbol{\pi}(A) \geq \frac{1}{2}\right\}, \tag{11}$$

where $A_r = \{u \in \mathbb{R}^\Omega, d(u, A) < r\}$ (d is the Euclidean distance of $\mathcal{X}$). For instance, in the case of a uniform distribution on the d-dimensional unit sphere, the concentration function can be proved to be smaller than a Gaussian function of r [36], whose fast decay for large r indicates thanks to (11) that the probability is very concentrated near to any equator.

An analytical point of view for $\alpha_{\boldsymbol{\pi}}$ is useful when considering other distributions. The concentration function can be addressed in terms of concentration of Lipschitz-continuous functions around their median. Namely, a measurable function $F : \mathcal{X} \to \mathbb{R}$ is said to be 1-Lipschitz when

$$\|F\|_{\text{Lip}} := \sup_{u,v\in\mathcal{X}} \frac{|F(u) - F(v)|}{\|u - v\|} \leq 1,$$

and m_F is called a median of F if it satisfies

$$\boldsymbol{\pi}(F \leq m_F) \geq \frac{1}{2} \quad \text{and} \quad \boldsymbol{\pi}(F \geq m_F) \geq \frac{1}{2}.$$

The concentration function can be characterized for any $r > 0$ by [36]

$$\alpha_{\boldsymbol{\pi}}(r) = \sup_F \boldsymbol{\pi}(F - m_F \geq r), \tag{12}$$

where the supremum runs over all real-valued measurable 1-Lipschitz functions F, and where m_F is any median of F.

In the case of the posterior probability having p.d.f. (7) of our image denoising problem, we can have a Gaussian concentration inequality similar to the uniform distribution on the unit sphere, as shows

Proposition 2 (Measure concentration) *Let $\boldsymbol{\pi}$ denote the posterior probability with p.d.f. (7). Then*

$$\forall r > 0, \quad \alpha_{\boldsymbol{\pi}}(r) \leq 2e^{-\frac{r^2}{4\sigma^2}}. \tag{13}$$

Proof —The probability $\boldsymbol{\pi}$ has p.d.f. $\pi = \frac{1}{Z}e^{-V}$ where $V = \frac{1}{2\sigma^2}E_{v,\lambda}$. V satisfies the strong convexity inequality

$$\exists c > 0,\ \forall u, v \in \mathbb{R}^\Omega, \quad V(u) + V(v) - 2\,V\left(\frac{u+v}{2}\right) \geq \frac{c}{4}\|u - v\|^2 \tag{14}$$

with $c = 1/\sigma^2$. Then applying [37, Theorem 2.15 p. 36], we obtain (13). □

This shows that any Lipschitz-continuous function F is concentrated around its median m_F: indeed, combining (12) and (13) yields

$$\boldsymbol{\pi}(F - m_F \geq r) \leq 2e^{-\frac{\|F\|_{\text{Lip}}^2 r^2}{4\sigma^2}}, \tag{15}$$

and the same argument with $-F$ leads to

$$\boldsymbol{\pi}(F - m_F \leq -r) \leq 2e^{-\frac{\|F\|_{\text{Lip}}^2 r^2}{4\sigma^2}}. \tag{16}$$

Both inequalities put together, we deduce that for each $r > 0$,

$$\boldsymbol{\pi}(|F - m_F| \geq r) \leq 4e^{-\frac{\|F\|_{\text{Lip}}^2 r^2}{4\sigma^2}}. \tag{17}$$

Now we prove that the energy $E_{v,\lambda}$ is concentrated around a particular value. $E_{v,\lambda}$ is not Lipschitz-continuous (because the data-fidelity term is quadratic), but its square-root is Lipschitz-continuous as soon as v is not constant, which allows to conclude to a weaker form of concentration for the energy.

Proposition 3 *Let $\boldsymbol{\pi}$ denote the posterior probability with p.d.f. (7). Assume that v is not constant. Then there exist $m \in \mathbb{R}$ and $c > 0$ such that for any $r \geq 0$,*

$$\boldsymbol{\pi}(|\sqrt{E_{v,\lambda}} - m| \geq r) \leq 4e^{-\frac{c^2 r^2}{4\sigma^2}}. \tag{18}$$

Proof —For any images u and u', let us write

$$\sqrt{E_{v,\lambda}(u')} - \sqrt{E_{v,\lambda}(u)} = C_1 + C_2$$

with

$$C_1 = \sqrt{\|u' - v\|^2 + \lambda TV(u')} - \sqrt{\|u - v\|^2 + \lambda TV(u')}$$

and

$$C_2 = \sqrt{\|u - v\|^2 + \lambda TV(u')} - \sqrt{\|u - v\|^2 + \lambda TV(u)}.$$

For C_1, let us recall that when u' is fixed, $u \mapsto \sqrt{\|u - v\|^2 + \lambda TV(u')}$ is the composition of $u \mapsto \|u - v\|$ and $x \in \mathbb{R} \mapsto \sqrt{x^2 + \varepsilon}$ with $\varepsilon = \lambda TV(u') \geq 0$, which both are 1-Lipschitz. This gives the inequality

$$|C_1| \leq \|u' - u\|. \tag{19}$$

To bound C_2, we need to compute the Lipschitz constant of TV. It depends on the scheme for TV which is used, and monotonically depends on $|\Omega|$. Writing $\|TV\|_{\text{Lip}} = \kappa\sqrt{|\Omega|}$, κ can be evaluated to $\kappa = 4 + O(1/\sqrt{|\Omega|})$ for the ℓ^1-scheme, and $\kappa = 2\sqrt{2} + O(1/\sqrt{|\Omega|})$ for the ℓ^2-scheme (the approximation is due to the domain's border effect, and in both cases the Lipschitz constant is reached when computing $TV(u) - TV(0)$, where u is the chessboard image defined by $u(i,j) = (-1)^{i+j}$. We have

$$C_2 = \frac{\|u - v\|^2 + \lambda TV(u') - \|u - v\|^2 - \lambda TV(u)}{\sqrt{\|u - v\|^2 + \lambda TV(u')} + \sqrt{\|u - v\|^2 + \lambda TV(u)}}.$$

But as v is supposed to be non-constant, $E_{v,\lambda}$ is coercive and cannot equal zero, so that it is bounded from below by a positive constant. Hence, since $\sqrt{\|u - v\|^2 + \lambda TV(u')}$ is non-negative, we have

$$|C_2| \leq \frac{\lambda |TV(u') - TV(u)|}{0 + \sqrt{\min E_{v,\lambda}}} \leq \frac{\lambda \kappa \sqrt{|\Omega|} \|u' - u\|}{\sqrt{\min E_{v,\lambda}}}. \tag{20}$$

Then, combining (19) and (20), we obtain

$$\left|\sqrt{E_{v,\lambda}(u')} - \sqrt{E_{v,\lambda}(u)}\right| \leq \left(1 + \frac{\lambda\kappa\sqrt{|\Omega|}}{\sqrt{\min E_{v,\lambda}}}\right) \|u' - u\|,$$

and $\sqrt{E_{v,\lambda}}$ is Lipschitz-continuous, with constant c, with $c = \lambda\kappa\sqrt{|\Omega|/\min E_{v,\lambda}} + O(1)$ when $|\Omega|$ goes to ∞. We conclude by applying Proposition 2. □

By homogeneity of $\|\cdot - v\|^2$ and TV with respect of the dimension $|\Omega|$ of the images, it is not restrictive to assume that the median m goes to ∞ as the dimension $|\Omega|$ of images increases (by juxtaposing several versions of v for instance). This means that as the dimension increases, $\boldsymbol{\pi}(|\sqrt{E_{v,\lambda}} - m|/|m| \geq r)$ is bounded by $4e^{-\frac{c^2r^2m^2}{4\sigma^2}}$, where

$$c^2 = \frac{\lambda^2\kappa^2|\Omega|}{\min E_{v,\lambda}} + O(1)$$

is bounded because $\min E_{v,\lambda}$ is proportional to $|\Omega|$, while m goes to $+\infty$ as $|\Omega| \to +\infty$. Hence $\boldsymbol{\pi}(|\sqrt{E_{v,\lambda}} - m|/|m| \geq r)$ converges to 0, and for large domains Ω, almost any image u drawn from $\boldsymbol{\pi}$ satisfies $E_{v,\lambda}(u) \approx m^2$.

As $E_{v,\lambda}$ is strictly convex and continuous, the lower set $\{u,\ E_{v,\lambda}(u) < m^2\}$ is a bounded convex set. It is not symmetric and its boundary is not smooth as soon as v is not constant and $\lambda > 0$, but it always contains $\hat{u}_{\text{MAP}}$ (because it reaches the lowest energy). Let us define the *median energy set* as the boundary of $\{u,\ E_{v,\lambda}(u) < m^2\}$. In high dimension, (18) means that almost all the mass of $\boldsymbol{\pi}$ is supported by a thin dilation of this median energy set. Estimating the original image u by $\hat{u}_{\text{MAP}}$ does not take the geometry of this median energy set into consideration, its asymmetrical shape in particular. In high dimension, the mean of $\boldsymbol{\pi}$ approximately corresponds to the isobarycenter of the median energy set, which is likely to give interesting results in terms of image denoising performance.

2.4 Definition of the TV-LSE operator

Instead of using the risk associated to a Dirac cost (leading to a MAP estimate), we propose to use a Least Square risk, which amounts to search the image $\hat{u}(v)$ minimizing

$$\mathbb{E}_{u,v}\left(\|u - \hat{u}(v)\|^2\right) = \int_{\mathbb{R}^\Omega}\int_{\mathcal{E}_\mu} \|u - \hat{u}(v)\|^2 p(u,v)\, dv\, du. \tag{21}$$

The image reaching this minimum is the expectation of the posterior distribution (Least Square Estimate, written LSE), that is

$$\hat{u}_{\text{LSE}} := \mathbb{E}(u|v) = \int_{u\in\mathbb{R}^\Omega} p(u|v)\, u\, du, \tag{22}$$

which can be, thanks to (7), rewritten under the form below.

Definition 1 *The TV-LSE operator (denoted by S_{LSE}) maps a discrete image $v \in \mathbb{R}^\Omega$ into the discrete image $\hat{u}_{\text{LSE}}$ defined by*

$$\hat{u}_{\text{LSE}} = S_{\text{LSE}}(v) = \frac{\displaystyle\int_{\mathbb{R}^\Omega} \exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2}\right)\cdot u\, du}{\displaystyle\int_{\mathbb{R}^\Omega} \exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2}\right) du}, \tag{23}$$

where λ and σ are positive parameters and $E_{v,\lambda}$ is the energy function defined in (5).

In this paper we concentrate on TV-LSE, but we are conscious that minimizing risks other than Least Square risk in (21) can lead to other interesting estimates (a median estimate for a L^1 risk for instance), though they seem to be more difficult to analyze.

3 Properties of TV-LSE

In this section, we explore several theoretical aspects of the TV-LSE operator. We give geometric invariance properties, and study the limiting operator when one of the parameters goes either to 0 or to ∞. Finally, we use Moreau's theory of proximations (or *proximity operators*) [45, 56] to state finer properties of TV-LSE, among which the fact that the staircasing effect cannot occur in TV-LSE denoising.

3.1 Invariance properties

Here we give several geometric invariance properties of S_{LSE} (gray-level average preservation, translation and symmetry invariance, ...), all shared with MAP denoising [2], which are basic but essential requirements for image processing.

Proposition 4 (Average preservation) *For any image u, let $\bar{u} = \frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega} u(\mathbf{x})$ denote the average gray level of u. Then for every $v \in \mathbb{R}^\Omega$,*

$$\overline{S_{\mathrm{LSE}}(v)} = \bar{v}.$$

Proof —Let $\mathcal{E}_0 = \{u \in \mathbb{R}^\Omega : \bar{u} = 0\}$ denote the subspace of images with mean zero, and $\mathbb{1}$ the constant image equal to 1 everywhere. Splitting up the variable $u = u_0 + t\mathbb{1}$ into a zero-mean image $u_0 \in \mathcal{E}_0$ and a shift t (note that $\bar{u} = t = \frac{1}{|\Omega|}\langle u, \mathbb{1}\rangle$), the LSE-denoised image writes

$$\hat{u}_{\mathrm{LSE}} = \frac{1}{Z}\int_{u_0\in\mathcal{E}_0}\int_{t\in\mathbb{R}} u_0\; e^{-\frac{E_{v,\lambda}(u_0+t\mathbb{1})}{2\sigma^2}}\,dt\,du_0 + \frac{1}{Z}\int_{u_0\in\mathcal{E}_0}\int_{t\in\mathbb{R}} te^{-\frac{E_{v,\lambda}(u_0+t\mathbb{1})}{2\sigma^2}}\,dt\,du_0\;\mathbb{1}, \tag{24}$$

where Z is a generic normalizing constant. We show that $\hat{u}_{\mathrm{LSE}}$ has mean $\bar{v}$. The first integral in (24) has mean zero since it is a weighted average of zero-mean images. Let us focus on the second integral. As $TV(u) = TV(u_0)$, the energy may be written

$$E_{v,\lambda}(u_0 + t\mathbb{1}) = \|\mathbb{1}\|^2\left(t - \frac{\langle v-u_0, \mathbb{1}\rangle}{\|\mathbb{1}\|^2}\right)^2 + \|u_0 - v\|^2 - \frac{\langle \mathbb{1}, u_0 - v\rangle^2}{\|\mathbb{1}\|^2} + \lambda TV(u_0),$$

with $\|\mathbb{1}\|^2 = |\Omega|$. Then integrating along t yields

$$\int_{\mathbb{R}} te^{-\frac{|\Omega|\left(t-\frac{1}{|\Omega|}\langle v-u_0,\mathbb{1}\rangle\right)^2}{2\sigma^2}}\,dt = \frac{\sqrt{2\pi}\sigma}{|\Omega|^{3/2}}\langle v-u_0, \mathbb{1}\rangle\,.$$

Hence the second integral in (24) reads

$$\begin{aligned}\frac{1}{Z}\int_{u_0\in\mathcal{E}_0}\int_{t\in\mathbb{R}} te^{-\frac{E_{v,\lambda}(u_0+t\mathbb{1})}{2\sigma^2}}\,dt\,du_0 &= \frac{1}{Z'}\int_{u_0\in\mathcal{E}_0}\langle v-u_0, \mathbb{1}\rangle\; e^{-\frac{\|u_0-v\|^2-\frac{1}{|\Omega|}\langle\mathbb{1},u_0-v\rangle^2+\lambda TV(u_0)}{2\sigma^2}}\,du_0\\ &= \frac{1}{|\Omega|}\langle v, \mathbb{1}\rangle - \left\langle \frac{1}{Z'}\int_{u_0\in\mathcal{E}_0} u_0\; e^{-\frac{\|u_0-v\|^2-\frac{1}{|\Omega|}\langle\mathbb{1},u_0-v\rangle^2+\lambda TV(u_0)}{2\sigma^2}}\,du_0,\; \mathbb{1}\right\rangle,\end{aligned}$$

which equals $\bar{v}$, for the integral inside the inner product is again a weighted average of zero-mean images. Finally $\hat{u}_{\mathrm{LSE}}$ is a zero-mean image shifted by $\bar{v}$, therefore $\hat{u}_{\mathrm{LSE}}$ and v have the same average value. □

Proposition 5 (Invariance by composition with linear isometry) *Let $s : \mathbb{R}^\Omega \to \mathbb{R}^\Omega$ be a linear isometry such that for all $u \in \mathbb{R}^\Omega$, $TV \circ s(u) = TV(u)$ holds. Then*

$$\forall v \in \mathbb{R}^\Omega, \quad S_{\mathrm{LSE}} \circ s(v) = s \circ S_{\mathrm{LSE}}(v).$$

Proof —The change of variable $u' = s^{-1}(u)$ in the numerator and the denominator of (23) yields

$$S_{\mathrm{LSE}}(s(v)) = \frac{\int s(u') e^{-\frac{\|s(u')-s(v)\|^2+\lambda TV(s(u'))}{2\sigma^2}} du'}{\int e^{-\frac{\|s(u')-s(v)\|^2+\lambda TV(s(u'))}{2\sigma^2}} du'},$$

because s being an isometry implies $ds(u') = du'$. Furthermore s is isometric, so we have $\|s(u')-s(v)\|^2 = \|u'-v\|^2$, and $TV(s(u')) = TV(u')$ thus

$$S_{\mathrm{LSE}} \circ s(v) = \frac{\int s(u') e^{-\frac{\|u'-v\|^2+\lambda TV(u')}{2\sigma^2}} du'}{\int e^{-\frac{\|u'-v\|^2+\lambda TV(u')}{2\sigma^2}} du'} = s(S_{\mathrm{LSE}}(v)),$$

because s is linear. □

A consequence of Proposition 5 is that the TV-LSE operator inherits many properties of the discrete scheme used for TV. For the classical ℓ^1 or ℓ^2 schemes associated to Equations (3) and (4), we obtain in particular the following invariances:

1) translation invariance: $S_{\mathrm{LSE}} \circ \tau_t = \tau_t \circ S_{\mathrm{LSE}}$, where τ_t is the translation operator of vector $t \in \mathbb{Z}^2$ defined by $\tau_t \circ u(x) = u(x-t)$ (Ω is assumed to be a torus)

2) $\pi/2$-rotation invariance: if ρ is a $\pi/2$-rotation sending Ω onto itself, then $S_{\mathrm{LSE}} \circ \rho = \rho \circ S_{\mathrm{LSE}}$

3) gray-level shift invariance: $\forall u \in \mathbb{R}^\Omega, \forall c \in \mathbb{R}, \quad S_{\mathrm{LSE}}(u+c) = S_{\mathrm{LSE}}(u) + c$ (this is not a direct consequence of Proposition 5, but the proof is easily adapted to the case $s(u) = u + c$).

These properties can help find the structure of $S_{\mathrm{LSE}}(v)$ when v contains lots of redundancies and structure. For example, if v is a constant image, then $S_{\mathrm{LSE}}(v) = v$. Indeed, v is invariant under the translations of vector $(1,0)$ and $(0,1)$, and so is $S_{\mathrm{LSE}}(v)$; moreover the average gray level of $S_{\mathrm{LSE}}(v)$ is the same as v. Finally $S_{\mathrm{LSE}}(v)$ is a constant equal to v. Another example is the checkerboard, defined by

$$v_{i,j} = \begin{cases} a & \text{if } i+j \text{ is even} \\ b & \text{if } i+j \text{ is odd} \end{cases}$$

for some constants $a, b \in \mathbb{R}$. It is quite easy to see that $v' = S_{\mathrm{LSE}}(v)$ is also a checkerboard (use the invariance by translations of vectors $(1,1)$ and $(1,-1)$), even if it seems difficult to get the associated parameter values a' and b'.

3.2 Asymptotics

Unlike MAP denoising (that depends on the only parameter λ), LSE denoising depends on 2 distinct parameters λ and σ. The theorem below sums up several asymptotic behaviors of $\hat{u}_{\mathrm{LSE}}$, when one of the parameters goes to 0 or $+\infty$.

Remark 1 *By the change of variables $v' = v/\sigma$, $u' = u/\sigma$, $\lambda' = \lambda/\sigma$, $\sigma' = 1$, the transformed operator, with obvious notations, satisfies $S_{\mathrm{LSE}}^{\lambda,\sigma}(v) = \sigma S_{\mathrm{LSE}}^{\lambda/\sigma,1}(\frac{v}{\sigma})$.*

Theorem 1 *For a given image $v \in \mathbb{R}^\Omega$, let us write $\hat{u}_{\mathrm{LSE}}(\lambda,\sigma) = S_{\mathrm{LSE}}(v)$ to recall the dependency of $\hat{u}_{\mathrm{LSE}}$ with respect to λ and σ. For any fixed $\lambda > 0$, we have*

$$\begin{aligned} (i) \quad & \hat{u}_{LSE}(\lambda,\sigma) \quad \xrightarrow[\sigma\to 0]{} \quad \hat{u}_{MAP}(\lambda), \\ (ii) \quad & \hat{u}_{LSE}(\lambda,\sigma) \quad \xrightarrow[\sigma\to +\infty]{} \quad v, \end{aligned}$$

while for any $\sigma > 0$, we have

$$\begin{array}{llll} (iii) & \hat{u}_{LSE}(\lambda,\sigma) & \xrightarrow[\lambda\to 0]{} & v, \\ (iv) & \hat{u}_{LSE}(\lambda,\sigma) & \xrightarrow[\lambda\to+\infty]{} & \bar{v}\mathbb{1}, \end{array}$$

where $\bar{v}\mathbb{1}$ is the constant image equal to the average of v.

Proof —As $E_{v,\lambda}$ is strongly convex (Equation (14)), the probability distribution $\frac{1}{Z}\exp\left(-\frac{E_{v,\lambda}}{2\sigma^2}\right)$ (where Z is a normalizing constant depending on σ) weakly converges when $\sigma \to 0$ to the Dirac distribution located at $\hat{u}_{MAP}(\lambda) = \arg\min_u E_{v,\lambda}(u)$, whose expectation is $\hat{u}_{MAP}(\lambda)$, which proves (i).

For (ii), let us consider the change of variable $w = (u-v)/\sigma$. Then

$$\hat{u}_{LSE}(\lambda,\sigma) = v + \frac{\displaystyle\int_{\mathbb{R}^\Omega} \sigma w e^{-\frac{1}{2}(\|w\|^2 + \frac{\lambda}{\sigma}TV(w+\frac{v}{\sigma}))}\,dw}{\displaystyle\int_{\mathbb{R}^\Omega} e^{-\frac{1}{2}(\|w\|^2 + \frac{\lambda}{\sigma}TV(w+\frac{v}{\sigma}))}\,dw} = v + \frac{N}{D}.$$

When $\sigma \to \infty$, the function inside the denominator D converges almost everywhere (a.e.) to $e^{-\|w\|^2/2}$, ans is uniformly bounded by $e^{-\|w\|^2/2}$, thus thanks to Lebesgue's dominated convergence theorem, D converges towards $\int e^{-\|w\|^2/2}\,dw$.

For the numerator, notice that the mean value theorem applied to $x \mapsto e^{-x}$ implies the existence of a real number $c_{w,\sigma} \in [0, \frac{\lambda}{2\sigma}TV(w+\frac{v}{\sigma})]$ such that

$$e^{-\frac{\lambda}{2\sigma}TV(w+\frac{v}{\sigma})} = 1 - \frac{\lambda}{2\sigma}TV(w+\frac{v}{\sigma})e^{-c_{w,\sigma}}.$$

Hence N can be split into

$$N = \sigma\int we^{-\frac{\|w\|^2}{2}}\,dw - \frac{\lambda}{2}\int \underbrace{we^{-\frac{\|w\|^2}{2}}TV(w+\frac{v}{\sigma})e^{-c_{w,\sigma}}}_{f_\sigma(w)}\,dw.$$

The first integral is equal to zero. Concerning the second integral, when $\sigma \to \infty$, $c_{w,\sigma}$ goes to 0, and as TV is Lipschitz continuous, f_σ satisfies for every $\sigma \geq 1$,

$$f_\sigma(w) \xrightarrow[\sigma\to\infty]{} we^{-\frac{\|w\|^2}{2}}TV(w) \quad \text{a.e.}$$

$$\text{and} \qquad \|f_\sigma(w)\| \leq \|w\|e^{-\frac{\|w\|^2}{2}}(TV(w) + \alpha\|v\|), \tag{25}$$

where α is the Lipschitz-continuity coefficient of TV. As the right-hand term of (25) belongs to $L^1(\mathbb{R}^\Omega)$ (as a function of w), again Lebesgue's dominated convergence theorem applies and

$$\int f_\sigma(w)\,dw \xrightarrow[\sigma\to\infty]{} \int we^{-\frac{\|w\|^2}{2}}TV(w)\,dw = 0$$

because the function inside the integral is odd (since TV is even). Hence, N goes to 0 as σ tends to infinity, which implies the convergence of $\hat{u}_{\mathrm{LSE}}(\lambda,\sigma)$ towards v, and proves (ii).

The proof of (iii) is a simple application of Lebesgue's dominated convergence theorem.

For (iv), the dominated convergence theorem cannot be applied directly because $u \mapsto ue^{-\frac{1}{2\sigma^2}TV(u)}$ does not belong to $L^1(\mathbb{R}^\Omega)$. We need to come down to a space of constant mean images where it becomes L^1. If we assume that the data image v has zero mean (which does not reduce the generality of the proof, because of the gray-level shift invariance of Section 3.1), then thanks to the average invariance property

(Proposition 4), we simply have to show that $\hat{u}_{\mathrm{LSE}}(\lambda,\sigma)$ converges to 0 when λ goes to ∞. Let us split every $u \in \mathbb{R}^\Omega$ into $u = (\bar{u}+z)/\lambda$ with $\bar{u} \in \mathbb{R}$ and $z \in \mathcal{E}_0$ (the space of zero mean images). The fidelity term $\|u-v\|^2$ becomes

$$\|u-v\|^2 = \left\|\frac{\bar{u}+z}{\lambda} - v\right\|^2 = \frac{|\Omega|}{\lambda^2}\bar{u}^2 + \left\|\frac{z}{\lambda} - v\right\|^2,$$

since both z and v have mean zero. Hence $\hat{u}_{\mathrm{LSE}}(\lambda,\sigma)$ writes

$$\begin{aligned}
\hat{u}_{\mathrm{LSE}}(\lambda,\sigma) &= \frac{\displaystyle\int_{z\in\mathcal{E}_0}\int_{\bar{u}\in\mathbb{R}} \frac{\bar{u}+z}{\lambda} e^{-\frac{1}{2\sigma^2}(\|\frac{\bar{u}+z}{\lambda}-v\|^2+TV(z))}\,d\bar{u}\,dz}{\displaystyle\int_{z\in\mathcal{E}_0}\int_{\bar{u}\in\mathbb{R}} e^{-\frac{1}{2\sigma^2}(\|\frac{\bar{u}+z}{\lambda}-v\|^2+TV(z))}\,d\bar{u}\,dz} \\
&= \frac{1}{\lambda}\frac{\displaystyle\int_{z\in\mathcal{E}_0} e^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))}\int_{\bar{u}\in\mathbb{R}} (\bar{u}+z)e^{-\frac{|\Omega|\bar{u}^2}{2\sigma^2\lambda^2}}\,d\bar{u}\,dz}{\displaystyle\int_{z\in\mathcal{E}_0} e^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))}\int_{\bar{u}\in\mathbb{R}} e^{-\frac{|\Omega|\bar{u}^2}{2\sigma^2\lambda^2}}\,d\bar{u}\,dz} \\
&= \frac{1}{\lambda}\frac{\displaystyle\int_{\mathcal{E}_0} z e^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))}\,dz}{\displaystyle\int_{\mathcal{E}_0} e^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))}\,dz}.
\end{aligned} \tag{26}$$

For both functions $g(z)=1$ and $g(z)=z$, we have

$$\begin{cases}
g(z)e^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))} \xrightarrow[\lambda\to\infty]{} g(z)e^{-\frac{1}{2\sigma^2}(\|v\|^2+TV(z))}, \\
\left\|g(z)e^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))}\right\| \le \|g(z)\|e^{-\frac{1}{2\sigma^2}TV(z)} \le \|g(z)\|e^{-\frac{C}{2\sigma^2}\|z\|_1},
\end{cases}$$

where the last inequality comes from the fact that since TV is a norm on the finite-dimensional space $\mathcal{E}_0$, there exists $C>0$ such that for every $z \in \mathcal{E}_0$, $TV(z) \ge C\|z\|_1$ (this can be considered as a discrete version of the Poincaré inequality [1]). Thus thanks to Lebesgue's dominated convergence theorem, each integral in (26) converges to a positive value when $\lambda \to +\infty$, and dividing by λ yields the desired limit $\hat{u}_{\mathrm{LSE}}(\lambda,\sigma) \to 0$. □

3.3 TV-LSE as a proximity operator and several consequences

Proximity operators [45, 56] are mappings of a Hilbert space into itself, that extend the notion of projection onto a convex space; here we prove that the TV-LSE denoiser is a proximity operator on $\mathbb{R}^\Omega$. From that, we deduce several stability and regularity properties of TV-LSE, and prove that it cannot create staircasing artifacts.

3.3.1 S_{LSE} is a proximity operator

Let us start by setting a frame of convex analysis (in finite dimension) around TV-LSE. Let $n = |\Omega|$ denote the total size of the considered images. An image is therefore an element of $\mathbb{R}^n$. Let $\Gamma_0(\mathbb{R}^n)$ be the space of convex, lower semi-continuous functions from $\mathbb{R}^n$ to $(-\infty,+\infty]$ that are proper (that is, non identically equal to $+\infty$).

Definition 2 *[45, 56] Let f be an arbitrary function in Γ_0. The proximity operator associated to f is the mapping* $\mathrm{prox}_f : \mathbb{R}^n \to \mathbb{R}^n$ *defined by*

$$\mathrm{prox}_f(u) = \arg\min_{v\in\mathbb{R}^n} \frac{1}{2}\|v-u\|^2 + f(v).$$

Notice that if f is the characteristic function associated to a closed, convex and non-empty set C ($f = 0$ on C and $f = +\infty$ elsewhere), prox_f simply reduces to the projection on C.

Recall 1 *[45, 56] Whatever f in Γ_0, its convex conjugate f^* (Legendre-Fenchel transform), defined by*

$$f^*(v) = \sup_{u \in \mathbb{R}^n} \langle u, v\rangle - f(u),$$

is in $\Gamma_0(\mathbb{R}^n)$, and satisfies $f^{**} = f$. Moreover, the Moreau's decomposition theorem states that given $f \in \Gamma_0$, every $z \in \mathbb{R}^n$ can be decomposed into $z = u + v$, with $u = \operatorname{prox}_f(z)$ and $v = \operatorname{prox}_{f^*}(z)$.

Definition 3 *[45, 56] The primitive function associated to prox_f is the function $\Phi \in \Gamma_0(\mathbb{R}^n)$ defined by*

$$\forall z \in \mathbb{R}^n,\ \Phi(z) = \frac{1}{2}\|v\|^2 + f(u) \quad \textit{where} \quad u = \operatorname{prox}_f(z) \textit{ and } v = \operatorname{prox}_{f^*}(z).$$

Now let $f = \frac{\lambda}{2\sigma^2}TV$. f is an element of $\Gamma_0(\mathbb{R}^n)$ whose domain $\operatorname{dom}(f) = \{u \in \mathbb{R}^n \,|\, f(u) < \infty\}$ has a non-empty interior. Besides, f can be viewed as the potential of the (improper) prior distribution in our Bayesian framework whose p.d.f. is $p = \exp(-f)$.

Letting G_σ denote the Gaussian kernel $u \in \mathbb{R}^n \mapsto \frac{1}{\sigma^n (2\pi)^{n/2}} \exp(-\frac{\|u\|^2}{2\sigma^2})$, the TV-LSE operator, denoted S_{LSE}, can be written

$$\forall v \in \mathbb{R}^n, \quad S_{\mathrm{LSE}}(v) = \frac{\int u\, G_\sigma(u-v)\, p(u)\, du}{\int G_\sigma(u-v)\, p(u)\, du}. \tag{27}$$

We come to the specific study of S_{LSE}.

Lemma 1 *S_{LSE} is differentiable, and its differential dS_{LSE} is a symmetric positive-definite matrix at every point.*

The proof is left in the appendix, in Section A.2.

Lemma 2 *There exists a $\mathcal{C}^\infty$ function $\varphi \in \Gamma_0(\mathbb{R}^n)$ such that $S_{\mathrm{LSE}} = \nabla\varphi$. Furthermore, φ is strictly convex and is defined by*

$$\varphi : v \in \mathbb{R}^n \mapsto \frac{1}{2}\|v\|^2 + \sigma^2 \log(p * G_\sigma)(v). \tag{28}$$

Proof —The function φ defined by (28) is $\mathcal{C}^\infty$ since the convolution of p with a Gaussian kernel is $\mathcal{C}^\infty$. Moreover, we have

$$\nabla\varphi(v) = v + \sigma^2 \nabla_v \log \int G_\sigma(v-u)\, p(u)\, du = v + \sigma^2 \frac{\int \nabla_v G_\sigma(v-u)\, p(u)\, du}{\int G_\sigma(v-u)\, p(u)\, du}, \tag{29}$$

and since $\nabla_v G_\sigma(v-u) = -\frac{1}{\sigma^2} G_\sigma(v-u) \cdot (v-u)$, we finally get

$$\nabla\varphi(v) = \frac{\int G_\sigma(v-u)\, p(u)\, u\, du}{\int G_\sigma(v-u)\, p(u)\, du} = S_{\mathrm{LSE}}(v).$$

Now the only difficulty is to prove that φ is strictly convex (as shown in the proof of Theorem 2 below, the second term $\sigma^2 \log(p * G_\sigma)$ is actually concave). In fact, it suffices to check that the Hessian of φ is (symmetric) positive-definite. But the Hessian of φ at point v equals the differential $dS_{\mathrm{LSE}}(v)$ of S_{LSE} at point v, and by Lemma 1, $dS_{\mathrm{LSE}}(v)$ is positive-definite, which ends the proof. □

Theorem 2 *The operator S_{LSE} is a proximity operator.*

Proof —The application $p * G_\sigma$ is log-concave as the convolution of two log-concave distributions [54]. Hence $\sigma^2 \log(p * G_\sigma)$ is concave, and φ is less convex than $u \mapsto \frac{1}{2}\|u\|^2$ (that is, the mapping $u \mapsto \frac{1}{2}\|u\|^2 - \varphi$ is convex). Then, applying [45, Proposition 9.b, $(I) \Rightarrow (III)$], φ is necessarily the primitive function associated to a proximity operator, that is, there exists $g \in \Gamma_0(\mathbb{R}^n)$ such that φ is the primitive function associated to prox_g. Now, denoting $g^* \in \Gamma_0(\mathbb{R}^n)$ the Legendre-Fenchel transform of g, we have $\nabla\varphi = \text{prox}_{g*}$ [45, Proposition 7.d] which proves that S_{LSE} is a proximity operator. □

As S_{LSE} is a proximity operator, we can define the convex function to which S_{LSE} is associated.

Definition 4 (TV_σ prior) *Let us assume that $\lambda = 1$. For any $\sigma > 0$, we define TV_σ as the unique function in $\Gamma_0(\mathbb{R}^n)$ such that $TV_\sigma(0) = 0$ and $S_{\text{LSE}} = \text{prox}_{\frac{1}{2}TV_\sigma}$.*

The existence of such a function TV_σ is given by Theorem 2, while the uniqueness is a consequence of [45, Proposition 8.a]. Thus, and still for $\lambda = 1$, $S_{\text{LSE}}(v)$ corresponds to the MAP estimation of v with the prior potential TV_σ, in the same way that ROF gives a MAP estimation of v with the prior potential TV. As we shall see in Section 3.3.3, the potential TV_σ has interesting properties that significantly differ from those of TV.

Note that for other values of λ, S_{LSE} remains a proximity operator, associated to a rescaled version of TV_σ. Indeed, with obvious notations for $S_{\text{LSE}}^{\lambda,\sigma}$, since we have

$$\forall v \in \mathbb{R}^\Omega, \qquad S_{\text{LSE}}^{\lambda,\sigma}(v) = \frac{1}{\lambda} S_{\text{LSE}}^{1,\frac{\sigma}{\lambda}}(\frac{1}{\lambda}v),$$

and the scaling property of the proximity operators,

$$\forall f \in \Gamma_0(\mathbb{R}^\Omega),\ \forall \alpha > 0,\ \forall v \in \mathbb{R}^\Omega, \quad \text{prox}_{\alpha^2 f}(v) = \alpha\, \text{prox}_{f(\alpha\cdot)}(\frac{1}{\alpha}v),$$

entails

$$S_{\text{LSE}}^{\lambda,\sigma} = \text{prox}_{\frac{\lambda^2}{2}TV_{\frac{\sigma}{\lambda}}(\frac{\cdot}{\lambda})}.$$

S_{LSE} being a proximity operator is a rather strong property that implies the following stability and monotonicity properties.

Corollary 1 S_{LSE} *is non-expansive, that is,*

$$\forall v_1, v_2 \in \mathbb{R}^n,\ \|S_{\text{LSE}}(v_2) - S_{\text{LSE}}(v_1)\| \leq \|v_2 - v_1\|, \tag{30}$$

and monotone in the sense of Minty, that is,

$$\forall v_1, v_2 \in \mathbb{R}^n,\ \langle S_{\text{LSE}}(v_2) - S_{\text{LSE}}(v_1), v_2 - v_1\rangle \geq \|S_{\text{LSE}}(v_2) - S_{\text{LSE}}(v_1)\|^2. \tag{31}$$

Proof —The non-expansiveness property is a consequence of [45, Proposition 5.b], and the monotonicity a consequence of [43] or [45, 5.a] (these properties are condensed in [56, p. 340]). □

3.3.2 S_{LSE} induces no staircasing

We first show that S_{LSE} is a $\mathcal{C}^\infty$-diffeomorphism from $\mathbb{R}^n$ onto itself.

Lemma 3 S_{LSE} *is injective.*

Proof —Assume that $S_{\text{LSE}}(v_1) = S_{\text{LSE}}(v_2)$. Then considering the mapping ψ such that

$$\psi(t) = \langle S_{\text{LSE}}((1-t)v_1 + tv_2), v_2 - v_1\rangle$$

satisfying $\psi(0) = \psi(1)$, its derivative

$$\psi'(t) = \langle dS_{\text{LSE}}((1-t)v_1 + tv_2)(v_2 - v_1), v_2 - v_1\rangle$$

must vanish at a certain point $t_0 \in [0,1]$. But $dS_{\text{LSE}}((1-t_0)v_1 + t_0 v_2)$ is a positive-definite matrix (see Lemma 1), and consequently $\psi'(t) > 0$ unless $v_1 = v_2$. □

Lemma 4 *Let I denote the identity of $\mathbb{R}^n$. The operator $S_{\text{LSE}} - I$ is bounded and S_{LSE} is onto.*

The proof follows from the Lipschitz continuity of the discrete TV operator, and is detailed in Appendix A.3.

Theorem 3 *S_{LSE} is a $\mathcal{C}^\infty$-diffeomorphism from $\mathbb{R}^n$ onto $\mathbb{R}^n$.*

Proof —S_{LSE} is $\mathcal{C}^\infty$ because it satisfies $S_{\text{LSE}} = \nabla\varphi$ with φ in $\mathcal{C}^\infty$ (see Lemma 2). Now, adding the fact that dS_{LSE} is invertible at every point (Lemma 1) and that S_{LSE} is injective (Lemma 3), and we obtain by the global inversion theorem that S_{LSE} is a $\mathcal{C}^\infty$-diffeomorphism from $\mathbb{R}^n$ to $S_{\text{LSE}}(\mathbb{R}^n)$. We conclude by using the fact that $S_{\text{LSE}}(\mathbb{R}^n) = \mathbb{R}^n$ (Lemma 4). □

Beside the fact that S_{LSE} has the regularity of a $\mathcal{C}^\infty$-diffeomorphism is interesting in itself (robustness of the output with respect to the input, non-destruction of information), it allows to state the main result of this section.

Theorem 4 (S_{LSE} induces no staircasing) *If V is a random image whose p.d.f. is absolutely continuous with respect to Lebesgue's measure, then for any distinct pixels $\mathbf{x}$ and $\mathbf{y}$, one has*

$$\mathbb{P}\left\{S_{\text{LSE}}(V)(\mathbf{x}) = S_{\text{LSE}}(V)(\mathbf{y})\right\} = 0.$$

A consequence of this property is that two neighboring pixels (say, for the 4- or the 8-connectedness) have a probability zero to have the same value in $S_{\text{LSE}}(V)$. Thus, almost surely $\hat{u}_{\text{LSE}}$ contains no constant region, which means that there is no staircasing in the sense of [49], contrary to ROF.

For example, if V writes $V = u + N$ with u a fixed image and N a white Gaussian noise, that is, a realization of V is a noisy version of u, or if V is drawn from the total variation distribution (that is, $V \sim \frac{1}{Z}e^{-\lambda TV(V)}$), then the assumption on V in Theorem 4 is met, and $\hat{u}_{\text{LSE}}$ almost surely contains no staircasing. Note that it does not tell that edges should be blurred out. In Section 3.3.3 (through a theoretical argument) and 4 (through denoising experiments), we show that it is indeed not the case.

Proof of Theorem 4 — Let p_V the probability measure associated to the random image V. Let A be the event $\{V(\mathbf{x}) = V(\mathbf{y})\} \subset \mathbb{R}^n$. As A is a subspace of $\mathbb{R}^n$ with dimension strictly less than n and p_V is absolutely continuous with respect to Lebesgue's measure, the probability $p_V(A)$ is null. Now

$$\mathbb{P}\left\{S_{\text{LSE}}(V)(\mathbf{x}) = S_{\text{LSE}}(V)(\mathbf{y})\right\} = p_V(S_{\text{LSE}}^{-1}(A)),$$

and as S_{LSE} is a diffeomorphism from $\mathbb{R}^n$ onto itself and the p.d.f. of p_V is measurable, the change of variables formula can apply [62, Théorème 1.1]. In particular, S_{LSE}^{-1} changes negligible sets into negligible sets [62, Lemme 2.1], and $p_V(S_{\text{LSE}}^{-1}(A)) = 0$. □

3.3.3 Properties of TV_σ

In this section, we study the potential TV_σ introduced in Definition 4. Since we have $S_{\mathrm{LSE}} = \mathrm{prox}_{\frac{1}{2}TV_\sigma}$ for $\lambda = 1$, the S_{LSE} operator can be considered as a MAP estimator associated to the prior $p_{\mathrm{LSE}} = \frac{1}{Z}\exp\left(-\frac{1}{2\sigma^2}TV_\sigma\right)$, or, equivalently, as the minimizer of a variational formulation including the classical squared L^2 data-fidelity term and the potential TV_σ, as pointed out in [29] in a more general framework. We here specifically investigate some properties of TV_σ, that are in particular useful to compare the TV_σ and TV potentials.

Proposition 6 *TV_σ is $\mathcal{C}^\infty$.*

Proof —Let $z \in \mathbb{R}^n$. Having $u \in \frac{1}{2}\partial TV_\sigma(z)$ is equivalent to having $\|z' - (u+z)\|^2 + TV_\sigma(z')$ minimized by z among all $z' \in \mathbb{R}^n$. Hence $z = S_{\mathrm{LSE}}(u+z)$. But as S_{LSE} is invertible, the solution u is unique and satisfies $u = S_{\mathrm{LSE}}^{-1}(z) - z$. This proves the equivalence

$$u \in \frac{1}{2}\partial TV_\sigma(z) \quad \Longleftrightarrow \quad u = S_{\mathrm{LSE}}^{-1}(z) - z.$$

This means that $\partial TV_\sigma(z)$ contains a single point, so that TV_σ is differentiable at point z. Furthermore we have

$$\frac{1}{2}\nabla TV_\sigma = S_{\mathrm{LSE}}^{-1} - I, \tag{32}$$

and the right-hand term is $\mathcal{C}^\infty$ thanks to Theorem 3, which concludes the proof. □

The regularity of TV_σ distinguishes it from TV which is singular. Intuitively, this is consistent with the behavior of the denoising operator in terms of staircasing: in [49] Nikolova proves (under particular assumptions which are probably not met here) that the differentiability of the regularizing term is a necessary and sufficient condition to avoid the staircasing effect.

Corollary 2 *∇TV_σ is bounded.*

Proof —Lemma 4 states that $S_{\mathrm{LSE}} - I$ is bounded, say by $c > 0$ (that is, $\|S_{\mathrm{LSE}}(u) - u\| \leq c$ for any $u \in \mathbb{R}^n$). For any $v \in \mathbb{R}^n$, letting $u = S_{\mathrm{LSE}}^{-1}(v)$ yields

$$\|S_{\mathrm{LSE}}^{-1}(v) - v\| = \|u - S_{\mathrm{LSE}}(u)\| \leq c,$$

and hence $S_{\mathrm{LSE}}^{-1} - I$ is bounded. Writing $\nabla TV_\sigma = 2(S_{\mathrm{LSE}}^{-1} - I)$ ends the proof. □

Let us see a consequence of Corollary 2. By definition of TV_σ, $\hat{u} = S_{\mathrm{LSE}}(v)$ minimizes $\|u-v\|^2 + TV_\sigma(u)$ among all $u \in \mathbb{R}^\Omega$. As TV_σ is smooth and convex, this energy can be differentiated and $\hat{u}$ is characterized by

$$2(\hat{u} - v) + \nabla TV_\sigma(\hat{u}) = 0. \tag{33}$$

Subtracting (33) in two neighboring pixels $\mathbf{x}$ and $\mathbf{y}$ yields

$$(\hat{u}(\mathbf{x}) - v(\mathbf{x})) - (\hat{u}(\mathbf{y}) - v(\mathbf{y})) = \frac{1}{2}\Big(\nabla TV_\sigma(\hat{u})(\mathbf{y}) - \nabla TV_\sigma(\hat{u})(\mathbf{x})\Big),$$

but as $\|\nabla TV_\sigma\|$ is bounded from above by, say, a constant $c' > 0$ (depending on σ), we have

$$|\hat{u}(\mathbf{x}) - \hat{u}(\mathbf{y})| \geq |v(\mathbf{x}) - v(\mathbf{y})| - c'.$$

In particular, if the absolute gap of v between pixels $\mathbf{x}$ and $\mathbf{y}$ is greater than c', then there will be also a gap for $\hat{u}$ between these pixels. This explains why TV-LSE is able, like ROF, to restore contrasted edges.

We end up this section with an explicit (but hardly tractable) formulation connecting TV_σ to TV.

Corollary 3 *The potential TV_σ is linked to TV by the equality*

$$\left(I + \frac{1}{2}\nabla TV_\sigma\right)^{-1} = I + \sigma^2 \frac{\nabla(p * G_\sigma)}{p * G_\sigma} \tag{34}$$

or, equivalently, by

$$\frac{1}{2}\nabla TV_\sigma = \left(I + \sigma^2 \nabla \log(e^{-\frac{TV}{2\sigma^2}} * G_\sigma)\right)^{-1} - I. \tag{35}$$

Proof —Rewriting (29) gives

$$S_{\text{LSE}} = I + \sigma^2 \frac{\nabla(p * G_\sigma)}{p * G_\sigma}.$$

Now, because of (32), we can write

$$S_{\text{LSE}}^{-1} = I + \frac{1}{2}\nabla TV_\sigma.$$

Grouping these two equations yields (34), and (35) immediately follows from $p = \frac{1}{Z}e^{-\frac{TV}{2\sigma^2}}$. □

There is probably no simple closed formula for TV_σ, but (35) is a natural starting point to derive approximations of ∇TV_σ. For instance, it seems that ∇TV_σ converges to ∇TV at each point where TV is differentiable. Obtaining a higher order Taylor expansion of the right-hand side of (35) would be most helpful to get an intuition of the deviation made by TV_σ with respect to TV. Closed-form approximations of TV_σ would be very interesting too, since they could be inserted in a minimization algorithm to efficiently compute approximations of the TV-LSE operator.

4 Experiments

4.1 An algorithm for TV-LSE

As we saw in (23), the denoised image $\hat{u}_{\text{LSE}}$ can be written

$$\hat{u}_{\text{LSE}} = \int u\, \boldsymbol{\pi}(du) = \int u\, \pi(u)\, du, \quad \text{where} \quad \pi(u) = \frac{1}{Z}e^{-\frac{1}{2\sigma^2}E_{v,\lambda}(u)} \tag{36}$$

is the density of the posterior distribution $\boldsymbol{\pi}$. Hence, the computation of $\hat{u}_{\text{LSE}}$ implies an integration on the whole space of discrete images $\mathbb{R}^\Omega$. Surprisingly enough, such an integration over a very high-dimensional space can be realized in a reasonable time via a Monte-Carlo Markov Chain (MCMC) method. Here we only give a quick and intuitive explanation of the algorithm described in [40]. A more complete publication, devoted to the detailed description and the study of this algorithm, is currently in preparation; for the time being, the interested reader can find more details in [39].

The principle of the MCMC algorithm is the following: if we were able to draw i.i.d. samples from the posterior distribution $\boldsymbol{\pi}$ (36), a good approximation of the posterior mean $\hat{u}_{\text{LSE}}$ could be obtained thanks to the law of large numbers by averaging all these samples. Now, as sampling directly from $\boldsymbol{\pi}$ is computationally out of reach, we build a first-order Markov chain of images $(U_n)_{n\geq 0}$ (which means that U_{n+1} only depends on U_n and on other independent random variables) whose stationary distribution (that is, the asymptotic distribution of U_n when $n \to +\infty$) is $\boldsymbol{\pi}$. The Metropolis-Hastings algorithm provides a simple way of achieving this. Then an ergodic theorem, well adapted to our framework, states that the average of the (dependent) samples successfully approximates the mean of $\boldsymbol{\pi}$ (see [40]).

Let us detail a little bit more the construction of (U_n). The first sample U_0 is drawn at random from an initial measure μ_0 (e.g., a white noise). Then, the transition from U_k to U_{k+1} (for any $k \geq 0$) is realized in two steps. First, an intermediate image $U_{k+1/2}$ is generated by adding a uniform random perturbation to one random pixel of U_k. Second, U_{k+1} is chosen to be equal to $U_{k+1/2}$ or U_k (that is, the transition $U_k \to U_{k+1/2}$ is accepted or not) according to the following rule: if $\pi(U_{k+1/2}) > \pi(U_k)$, then $U_{k+1} = U_{k+1/2}$ (the transition is accepted); otherwise, the transition is accepted only with probability $\pi(U_{k+1})/\pi(U_k)$ (if the transition is rejected, then $U_{k+1} = U_k$, that is, nothing happens during this iteration). The chain is run until reaching a precise convergence criterion, say at iteration n. In the end, we approximate $\hat{u}_{\mathrm{LSE}}$ by $\frac{1}{n}\sum_{k=1}^{n} U_k$.

This mathematical construction can be translated into Algorithm 1, which returns an estimate of $S_{\mathrm{LSE}}(u)$. It makes use of the function $E_{v,\lambda}^{\mathbf{x}}(u,t)$, which is defined as follows: call $u_{\mathbf{x},t} \in \mathbb{R}^{\Omega}$ the image defined by

$$\forall \mathbf{y} \in \Omega, \quad u_{\mathbf{x},t}(\mathbf{y}) = \begin{cases} u(\mathbf{y}) & \text{if } \mathbf{y} \neq \mathbf{x}, \\ t & \text{if } \mathbf{y} = \mathbf{x}, \end{cases}$$

then $E_{v,\lambda}^{\mathbf{x}}(u,t)$ captures in the formula for $E_{v,\lambda}(u_{\mathbf{x},t})$ (see Equation (5)) only the terms that depend on t. It is not difficult to see that if the ℓ^2 norm is used for $|Du|$, then

$$\begin{aligned} \forall (x,y) \in \Omega, \qquad E_{v,\lambda}^{(x,y)}(u,t) &= (t - v(x,y))^2 \\ &+ \lambda\sqrt{(u(x-1,y)-t)^2 + (u(x-1,y) - u(x-1,y+1))^2} \\ &+ \lambda\sqrt{(u(x,y-1)-t)^2 + (u(x,y-1) - u(x+1,y-1))^2} \\ &+ \lambda\sqrt{(t-u(x+1,y))^2 + (t-u(x,y+1))^2}, \end{aligned} \tag{37}$$

with the boundary convention that any squared difference term that contains an undefined term ($u(\mathbf{z})$ with $\mathbf{z} \notin \Omega$) is replaced with 0.

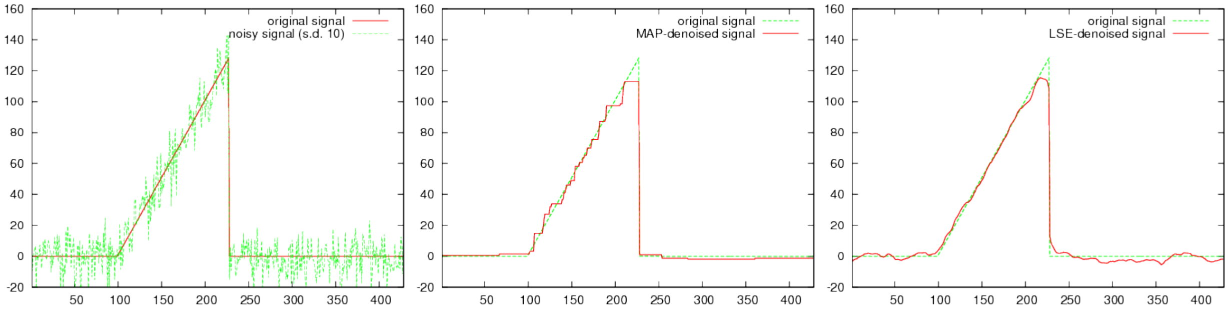
Algorithm 1 can be optimized in several ways (convergence control by means of two independent chains, use of an automatic burn-in procedure that skips the first iterations to reduce the bias, automatic setup of the α parameter, etc.), as explained in [39] and [40]. These references also contain a proof of convergence of the algorithm.

Algorithm 1 Metropolis-Hastings algorithm for $\hat{u}_{\mathrm{LSE}}$

```
$n \leftarrow 0$, $S \leftarrow 0_\Omega$
draw a white noise image $U$
repeat
    draw $\mathbf{x} \sim \mathcal{U}(\Omega)$ (uniform distribution on $\Omega$)
    $t \leftarrow U(\mathbf{x})$
    draw $t' \sim \mathcal{U}([t-\alpha, t+\alpha])$
    let $U(\mathbf{x}) \leftarrow t'$ with probability $\min\left(1, \exp\left(-\frac{E_{v,\lambda}^{\mathbf{x}}(u,t') - E_{v,\lambda}^{\mathbf{x}}(u,t)}{2\sigma^2}\right)\right)$, see Equation (37)
    $S \leftarrow S + U$
    $n \leftarrow n + 1$
until convergence is reached
return $\frac{1}{n}S$.
```

4.2 Comparison to the ROF model and the staircasing effect

In Figure 2 to 4, we show signals and images corrupted with additive Gaussian noise, and denoised using both the proposed TV-LSE method and the classical ROF method. The signal version of both denoisers consists in regarding the input signal as a one-line ($N \times 1$) image; note that in this case, both ℓ^1 and ℓ^2 schemes for $|Du|$ lead to absolute values of successive differences. On the one side, several similarities between the denoised signals or images can be noticed. Indeed, it can be seen that most of the noise is removed, and that contrasted contours (or large gaps for signals) are preserved. On the other side, the proposed TV-LSE model shows some differences with respect to the ROF model, the most striking of which being the avoidance of the staircasing effect, proved in Theorem 4, Section 3.3.2. This can be seen for instance in Figure 2, where the affine part of the signal is well restored by TV-LSE. In Figure 3, a constant image is corrupted with a Gaussian white noise ($\sigma = 20$) and then denoised by either ROF or TV-LSE for different values of the parameter λ, and we can observe that the artificial edges brought by ROF are avoided by the TV-LSE method, which manages to attenuate the noise in a much smoother way. Figure 4 again considers the images of Figure 1 and illustrates the good behavior of TV-LSE with respect to the staircasing effect, whereas the ROF denoiser transforms smooth regions into piecewise constant regions with spurious contrasted edges. Note also that TV-LSE denoised images have a more "textured" aspect than ROF denoised images. This heuristically agrees with the injectivity of the TV-LSE denoiser (Lemma 3 in Section 3.3.2), according to which two versions of the noisy image (2 different noise realizations) cannot lead to the same denoised result: there must remain some trace of the initial noise in the denoised image.



noisy ROF TV-LSE

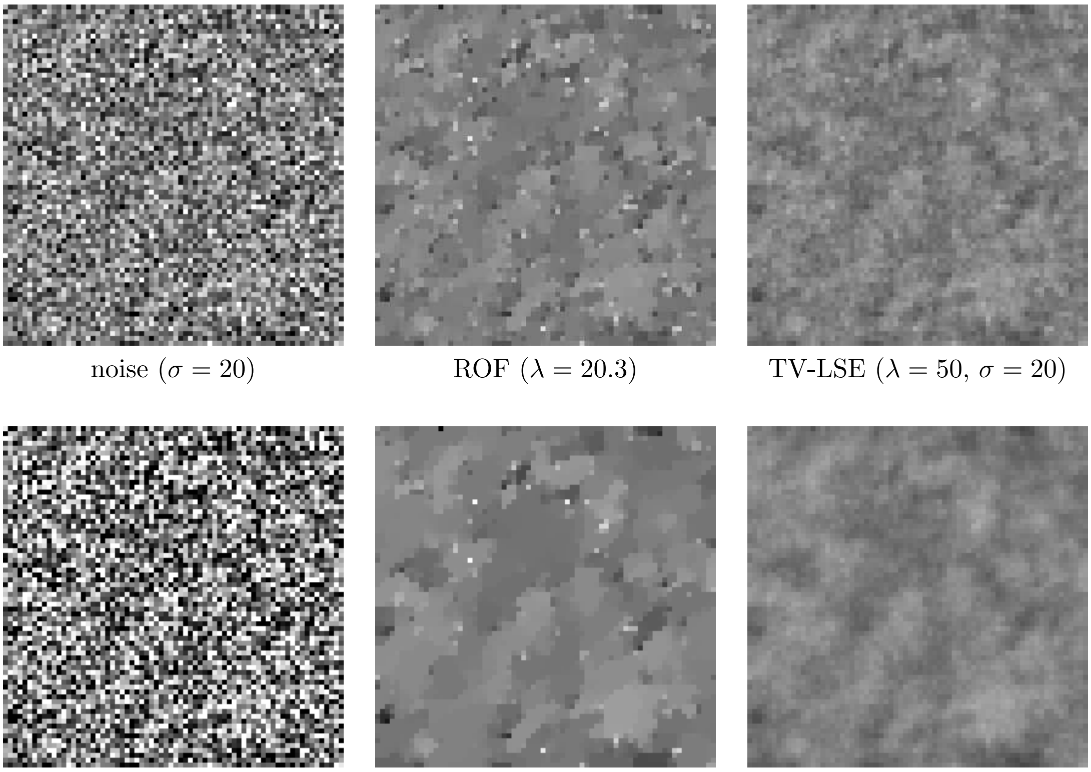
Figure 2: **Denoising of a simple synthetic signal.** A triangle-shaped signal (left figure, red curve) is corrupted by an additive white Gaussian noise, and the resulting signal (left, green curve) is then denoised using the ROF (middle) and TV-LSE (right) methods. In the ROF result, the noise has been wiped off on the initially constant parts of the signal, but a strong staircasing effect appears on the slope. The TV-LSE method behaves more smoothly: no staircasing appears on the slope and the noise is attenuated (but not completely removed) on the initially constant parts. The parameters of the ROF and TV-LSE methods have been set in order to equalize the *method noise* level (ℓ^2 distance from the noisy signal to the result).

4.3 Role of the hyperparameters

As clearly appears in Equation (23), the TV-LSE model involves two hyperparameters: the (known or estimated) noise standard deviation σ and the regularization parameter λ balancing the data-fidelity term and the regularity term. In comparison, the ROF model depends on the latter only.

Figures 5 and 6 show how the TV-LSE denoised image changes when λ is tuned while maintaining a fixed value of σ (Figure 5), or when σ is tuned with a fixed value of λ (Figure 6). One can see in Figure 5 that fixing $\sigma > 0$ and letting λ go to 0 makes the image look like the noisy initial image, and increasing λ makes the image smoother until it becomes a constant. One can also see in Figure 6 that fixing $\lambda > 0$ and letting σ go to 0 makes the image look like the ROF denoised image containing some staircasing effect, and that when σ gets larger, the image gets closer to the noisy initial image. All these observations agree with the asymptotic results of Section 3.2.

The λ parameter is useful since it permits to easily compare ROF and TV-LSE denoising methods. But a more relevant regularity parameter is $\beta = \frac{\lambda}{2\sigma^2}$, which corresponds to the inverse temperature in the prior probability (Equation 6) motivating the introduction of TV-LSE. Thus considering σ and β as the couple of hyperparameters of the model allows us to better dissociate the noise and regularization parameters. In Figure 7 a part of a noisy Lena image is denoised using TV-LSE with a constant β and increasing values of σ. The denoised image goes from the initial noisy image to a flat and smooth image:



noise ($\sigma = 20$) ROF ($\lambda = 20.3$) TV-LSE ($\lambda = 50$, $\sigma = 20$)

noise (enhanced contrast) ROF ($\lambda = 28.4$) TV-LSE ($\lambda = 100$, $\sigma = 20$)

Figure 3: **Denoising of a pure noise image**. A constant image is corrupted by a white Gaussian noise with standard deviation $\sigma = 20$ (top-left image, and bottom left image after an affine contrast change). On columns 2 and 3 we show respectively the results of ROF and TV-LSE methods on this image, the gray-level scale being the same for all images of a given row. As in Figure 2, the TV-LSE and ROF parameters are set in order to equalize (inside each row) the method noise levels of both methods. For the low denoising level (first row), isolated pixels remain in the ROF result (this can be understood by the fact that ROF is not far from being a ℓ^0 (sparse) recovery operator, and a single pixel with outstanding value has a relatively small cost for the ℓ^0 energy), which does not happen for TV-LSE. Furthermore, a staircasing effect (artificial edges) is clearly visible in the ROF result, while TV-LSE manages to maintain a smoother image. For the high denoising level (second row), ROF acts almost like a segmentation method, and breaks the domain into flat artificial regions, while the TV-LSE result gets uniformly smoother. This experiment clearly illustrates the different behavior of the ROF and TV-LSE methods on flat regions, and in particular the fact that the TV-LSE model, though being based on the TV operator, completely avoids the staircasing effect.

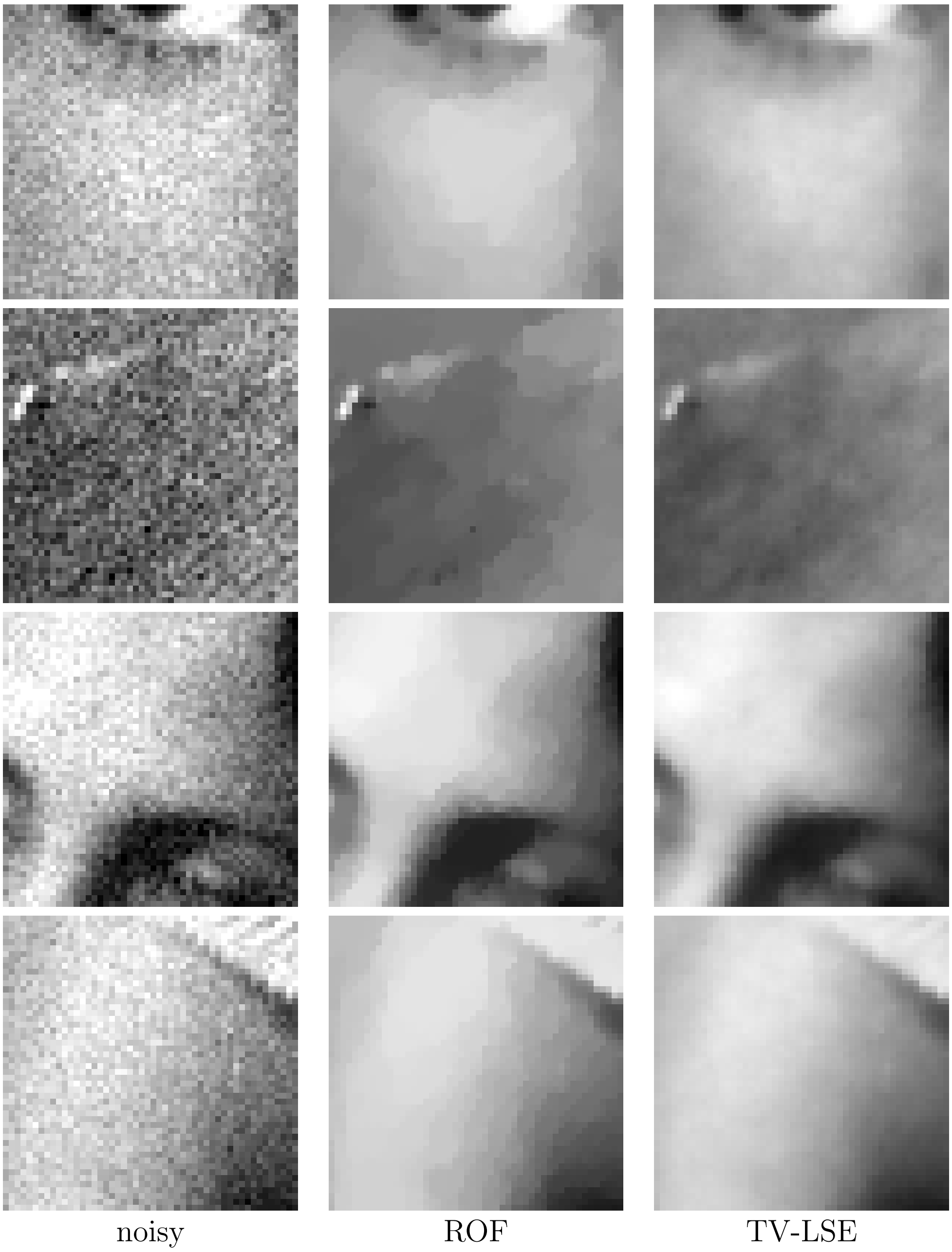


Figure 4: **No staircasing effect with TV-LSE.** We experimentally check that the TV-LSE method does not create staircasing artifacts. The left column shows parts of Lena and Barbara classical images, after they have been corrupted with an additive white Gaussian noise ($\sigma = 10$). The right column shows the corresponding TV-LSE denoised images with $(\sigma, \lambda) = (10, 40)$, while the middle column shows the ROF denoised images, with a value of λ that leads to the same method noise level in each case (from top to bottom: $\lambda_{MAP} = 25.6$, $\lambda_{MAP} = 20.3$, $\lambda_{MAP} = 29.0$, $\lambda_{MAP} = 26.9$). The main difference between the two methods is clearly the staircasing effect, which does not occur in TV-LSE images but introduces spurious edges in the ROF images.

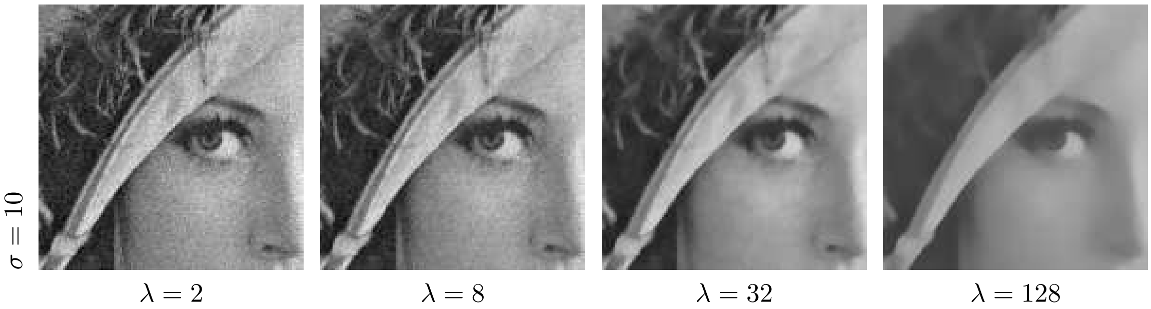


Figure 5: A noisy image is processed by TV-LSE with $\sigma = 10$ (which corresponds to the standard deviation of the noise) and increasing values of λ. When λ is small (left), the denoised image $\hat{u}_{\text{LSE}}$ is very close to the noisy image v; as λ increases, the noise gradually disappears, the homogeneous regions being smoothed out without staircasing; then, as λ increases further, the texture is erased, and the result gets close to a piecewise smooth image (right).

β really acts as the regularizing parameter. Notice that inversely, fixing σ and increasing β would be equivalent to the case of Figure 5 (fixed σ and increasing values of λ).

To compare precisely the interest of TV-LSE over ROF in terms of image denoising (see Figure 8) we fixed the level of denoising, measured by the L^2 norm of the residual image $v - \hat{u}_{\text{LSE}}$ (method noise), and considered increasing values of σ (for a given σ there exists at most one value of λ such that the desired level of denoising is reached, and this value increases with σ). For $\sigma = 0$, this corresponds to ROF denoising, but as σ increases we can observe the benefit of using TV-LSE in terms of staircasing. The fact that staircasing artifacts *gradually* disappear seems in contradiction with Theorem 4, stating that staircasing vanishes as soon as σ is positive; in fact it is not, and this simply comes from the fact that the (classical) definition of staircasing used in Theorem 4 is a qualitative (yes-no) property, while our perception is more quantitative (difference between gray level variations in flat zones and along their boundaries). By the way, it would certainly be interesting to characterize the limit TV-LSE image obtained by sending $\sigma \to +\infty$ while maintaining the method noise level as in Figure 8. Indeed, this limit image would define a filter controlled by a single parameter, the method noise level. In practice, we observed that ordinary values of σ (and in particular, choosing for σ the known or estimated noise level) lead to satisfactory results in the sense that they benefit from the good properties of the TV model (in particular edge preservation) without suffering from the staircasing effect.

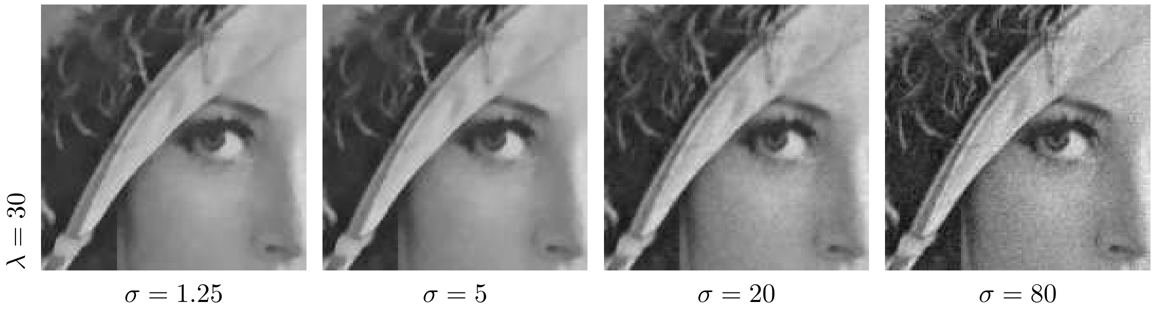


Figure 6: A noisy image is processed by TV-LSE with $\lambda = 30$ and increasing values of σ. When σ is small (left), the denoised image $\hat{u}_{\text{LSE}}$ is very close to the MAP-denoised image $\hat{u}_{\text{MAP}}(\lambda)$, with some texture erased and some staircasing visible: the cheek and hat parts contain boundaries which do not exist in the original Lena image. As σ increases, $\hat{u}_{\text{LSE}}$ looks more and more like the noisy image, which is consistent with the convergence $\hat{u}_{\text{LSE}}(\sigma, \lambda) \to v$ when $\sigma \to \infty$.

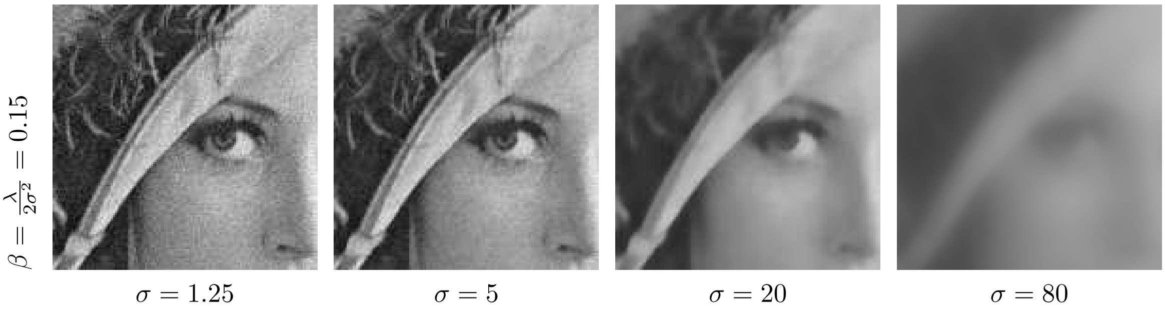


Figure 7: A noisy image is processed by TV-LSE with $\beta = \frac{\lambda}{2\sigma^2} = 0.15$ fixed and increasing values of σ. For small values of σ, the denoised image is close to the noisy image (left). As σ increases, the image is regularized, the edges are preserved but the texture is gradually erased. When σ further increases (right), the denoised image is completely blurred out.

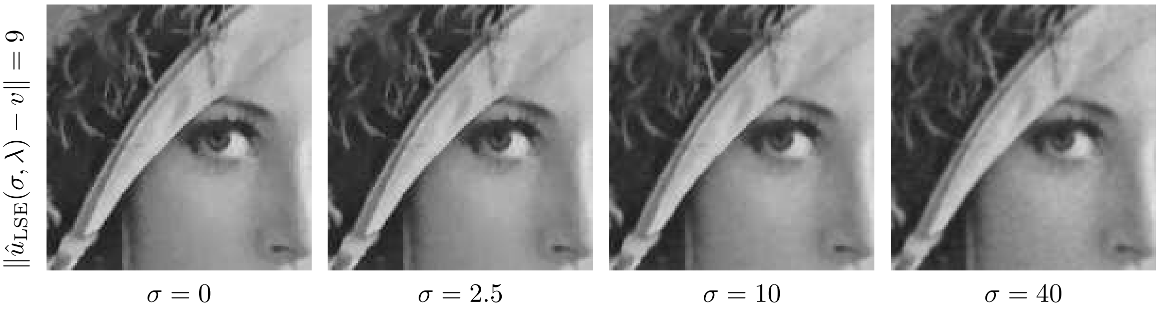


Figure 8: The level of denoising $\|\hat{u}_{\mathrm{LSE}}(\sigma, \lambda) - v\| = 9$ being fixed, TV-LSE is applied to a noisy image v for different values of σ. The value $\sigma = 0$ (left) corresponds to ROF: the image noise has been well cleaned, but some texture is erased, and staircasing is clearly visible (on the cheek for instance). As σ increases, staircasing disappears and the aspect of the denoised image becomes more natural.

Figure 9 gives a 2-dimensional view of the roles of the parameters σ and λ. The visual quality of the denoised image is good for medium values of σ and λ (typically $\sigma = 10$, corresponding to the noise level, and $\lambda = 40$), because it avoids the staircasing effect while maintaining the main structure of the image. The denoising quality is quite robust to the choice of σ, which allows for some inaccuracy in the estimation of the noise level.

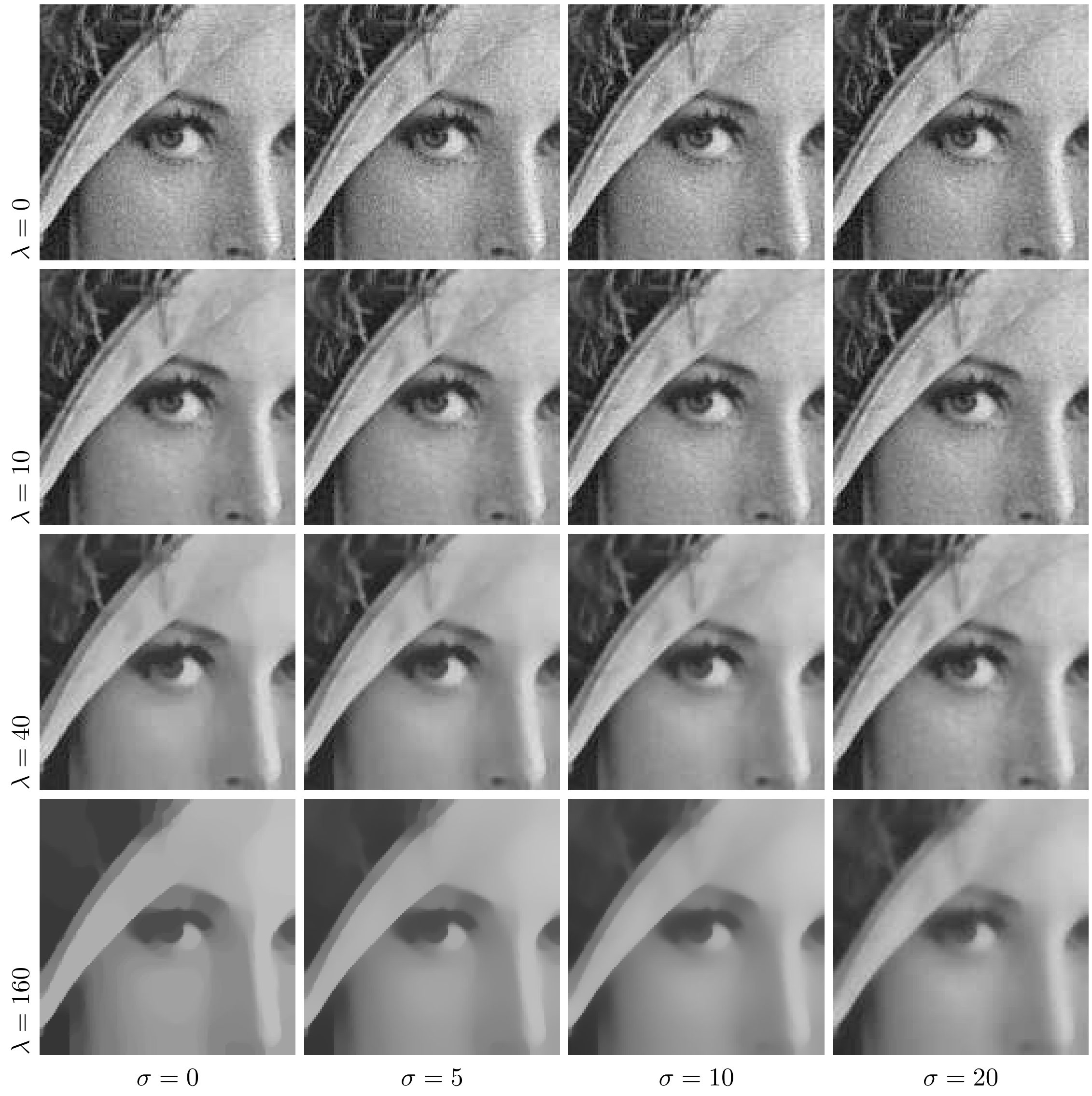


Figure 9: **Effect of the two parameters λ and σ on TV-LSE**. A noisy version of Lena image (Gaussian white noise with standard deviation equal to 10) is processed with TV-LSE for various values of λ and σ. First row: $\lambda = 0$ (the TV-LSE image is equal to the noisy image); second row: $\lambda = 10$; third row: $\lambda = 40$; last row: $\lambda = 160$. First column: $\sigma = 0$ (the TV-LSE denoised image corresponds to ROF); second column: $\sigma = 5$; third column: $\sigma = 10$; last column: $\sigma = 20$.

4.4 Comparison to other TV-based denoising methods

In this section, we propose to compare TV-LSE to other denoising methods through numerical experiments. We limit ourselves to TV-based methods, since the aim of this paper is not to bring a general and state-of-the art denoising method, but rather to explore new possibilities for Total Variation as a model for images, and in particular qualitative properties of the corresponding denoising algorithms. This is why we shall examine and discuss the visual properties of the denoised images rather than trying to blindly rank the different methods using classical metrics like PSNR or SSIM, which are poor predictors of the visual quality of the results.

Given a noisy image v, we propose to compare $\hat{u}_{\mathrm{LSE}}(\sigma, \lambda)$, the result of TV-LSE applied to v with parameters σ and λ, to:

- ROF denoising, alias TV-MAP: the denoised image is denoted by $\hat{u}_{\mathrm{MAP}}(\lambda_{MAP})$. The parameter λ_{MAP} is tuned in such a way that the denoising level $\|v - \hat{u}_{\mathrm{MAP}}(\lambda_{MAP})\|$ equals that of $\hat{u}_{\mathrm{LSE}}(\sigma, \lambda)$, $\|v - \hat{u}_{\mathrm{LSE}}(\sigma, \lambda)\|$;

- TV-barycenter: in order to be able to compare $\hat{u}_{\mathrm{LSE}}(\sigma, \lambda)$ and $\hat{u}_{\mathrm{MAP}}(\lambda)$ with the same value of λ (that is, for which both methods deal with the same energy $E_{v,\lambda}$), we propose to combine $\hat{u}_{\mathrm{MAP}}(\lambda)$ linearly with the noisy image v via
$$\hat{u}_{\mathrm{bary}} = t\, \hat{u}_{\mathrm{MAP}}(\lambda) + (1-t)\, v \quad \text{with } t = \frac{\|v - \hat{u}_{\mathrm{MAP}}(\lambda)\|}{\|v - \hat{u}_{\mathrm{LSE}}(\sigma, \lambda)\|}.$$
We obtain a barycenter of $\hat{u}_{\mathrm{MAP}}(\lambda)$ and v which has the desired denoising level. The choice of this method is also motivated by the observation that the quality of denoising often increases both visually and in PSNR when deviating the ROF estimate towards v (in other terms, visual quality is better when noise and texture are not completely removed).

- TV-ε : it is well-known that smoothing the Total Variation and embedding it in the usual variational framework leads to a staircasing-free denoising model [49]. More precisely, we can define generalizations of TV on $\mathbb{R}^\Omega$ by
$$TV_f(u) = \sum_{x \in \Omega} f(|Du|), \tag{38}$$
for specific smooth functions $f : \mathbb{R} \to \mathbb{R}$ that approximate the absolute value function, and then denoise an image v by minimizing
$$E_f(u) = \|u - v\|^2 + \lambda\, TV_f(u). \tag{39}$$
The smoothness of f in the neighborhood of 0 implies a regular processing of small gradients and avoids staircasing. A natural example of such a function f is
$$f_\varepsilon : x \mapsto \sqrt{\varepsilon^2 + x^2} \quad \text{with } \varepsilon > 0,$$
which is convex and smooth. This leads to a denoising method here called TV-ε, which is computable by a simple gradient descent. The parameter ε roughly corresponds to the minimal gradient magnitude of a discontinuity in the denoised image. We choose to set $\varepsilon = 10$ for images with gray levels lying in $[0, 255]$, while the parameter $\lambda = \lambda_\varepsilon$ is such that the denoising level of TV-LSE is reached.

- TV-Huber : another possible function f for (38) and (39), discussed in [64] for instance, is the so-called Huber norm
$$f_\alpha : x \mapsto \begin{cases} \frac{1}{2\alpha} x^2 & \text{if } |x| \le \alpha, \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha. \end{cases}$$
This leads to a denoising model here called TV-Huber model, which also has the property of avoiding the staircasing effect. A fast primal-dual algorithm can be used to compute the minimum of E_{f_α} [17]. As ε in TV-ε denoising, α corresponds to a minimal gradient for discontinuity, and is set to 10. The regularization parameter $\lambda = \lambda_{\mathrm{Huber}}$ is such that the denoising level of TV-LSE is reached.

- TV-L^1: we consider the minimizer of

  $$E(u) = \|u - v\|_1 + \lambda_{L^1} TV(u),$$

  where $\|\cdot\|_1$ is the L^1-norm. The only change of the fidelity term makes it especially adapted to remove impulse noise and makes the denoiser become contrast invariant [24, 48].

- local-TV : it has been proved in [41] that another way of avoiding staircasing in a TV framework is to "localize" it: denoising the pixel $\mathbf{x}$ of a noisy image v by the local-TV filter consists of first, extracting a patch $v(\mathcal{W}_\mathbf{x})$ centered at $\mathbf{x}$ from the image, then denoising the patch by ROF with a given regularizing parameter λ_{loc}, independent from $\mathbf{x}$, and finally assigning to the denoised image at $\mathbf{x}$ the central value of the denoised patch. The pixels of the patch can be weighted, leading to the more general scheme

  $$\hat{u}_{\text{loc}}(\mathbf{x}) = u(\mathbf{x}) \quad \text{where} \quad u \in \mathbb{R}^{\mathcal{W}_\mathbf{x}} \text{ minimizes } \sum_{\mathbf{y} \in \mathcal{W}_\mathbf{x}} \omega(\mathbf{y} - \mathbf{x})(u(\mathbf{y}) - v(\mathbf{y}))^2 + \lambda_{\text{loc}} TV(u),$$

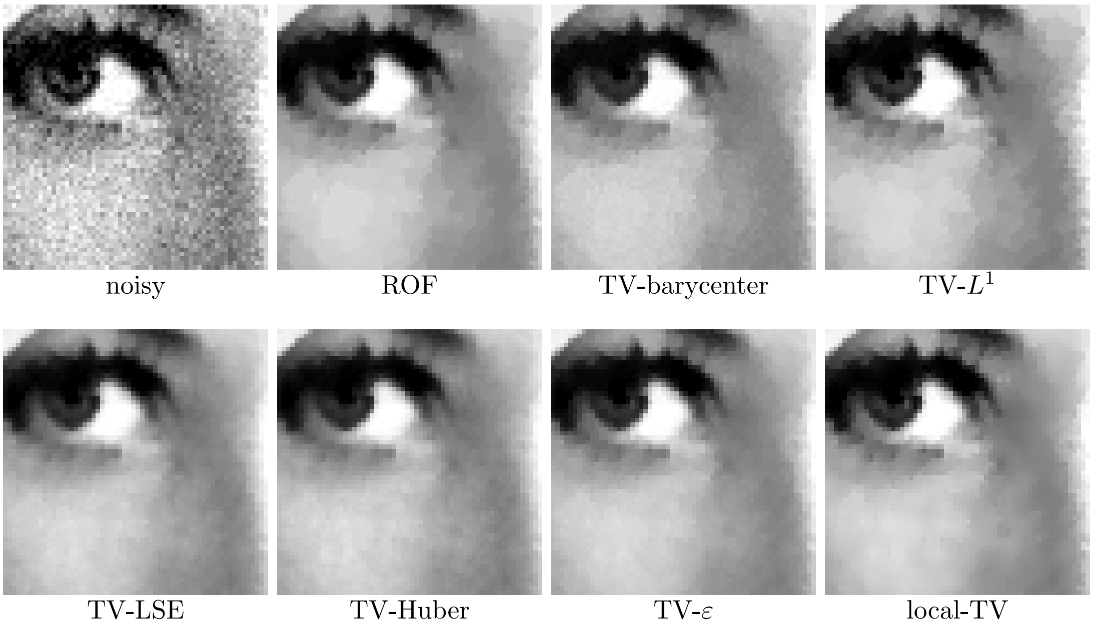
  for each pixel $\mathbf{x}$. This scheme (with Gaussian or constant weights $\omega(\mathbf{h})$ for instance) is able to avoid staircasing, in the sense that if all the patches of a given region have small enough variance, then the filter is equivalent to a blurring linear filter on this region [41]. In our present experiments, we use 5×5 patches, and Gaussian weights $\omega(\mathbf{h}) = \exp(-\|\mathbf{h}\|^2/(2a^2))$ with $a = 2$. The parameter λ_{loc} is chosen such that the denoising level is that of TV-LSE.

Figure 10, 11 and 12 zoom on different parts of the Lena image processed with all the methods listed above. As expected, ROF results present strong staircasing artifacts, and the added noise in TV-barycenter does not manage to remove them. The TV-L^1 model, due to its morphological invariance (invariance with respect to increasing contrast changes), is more suitable for granularity analysis or impulse noise removal than for piecewise smooth image retrieval, and the resulting images show even stronger staircasing artifacts. Among other methods, the similarity between the results of TV-Huber and TV-LSE is striking, both visually and qualitatively: there is no staircasing, a faithful reconstruction of contrasted edges, and a good overall quality. TV-ε also avoids staircasing and is able to reconstruct edges, but is not as good as TV-Huber and TV-LSE. Local-TV looks quite different: it is sharper than TV-LSE, but several spurious contours or spikes are still visible as in the ROF image.

We observed in our experiments that the results obtained with TV-Huber and TV-LSE could be very similar. We do not have a full explanation for this, but the results obtained in Section 3.3 shed an interesting light. Indeed, we showed that TV-LSE is a MAP estimator associated to the smooth prior potential TV_σ (see Definition 4), which seems, according to (35), to be a regularized version of TV converging to TV when σ goes to 0. Hence, it is not completely unexpected that replacing TV with a regularized prior as in TV-Huber leads to results that resemble those of TV-LSE, at least for small values of σ. It would be interesting to determine, among all regularized version of the gradient norm under the form $\varphi(\|Du\|)$, which function φ leads to the best approximation of the TV-LSE operator for a given choice of σ and λ.

5 Conclusion

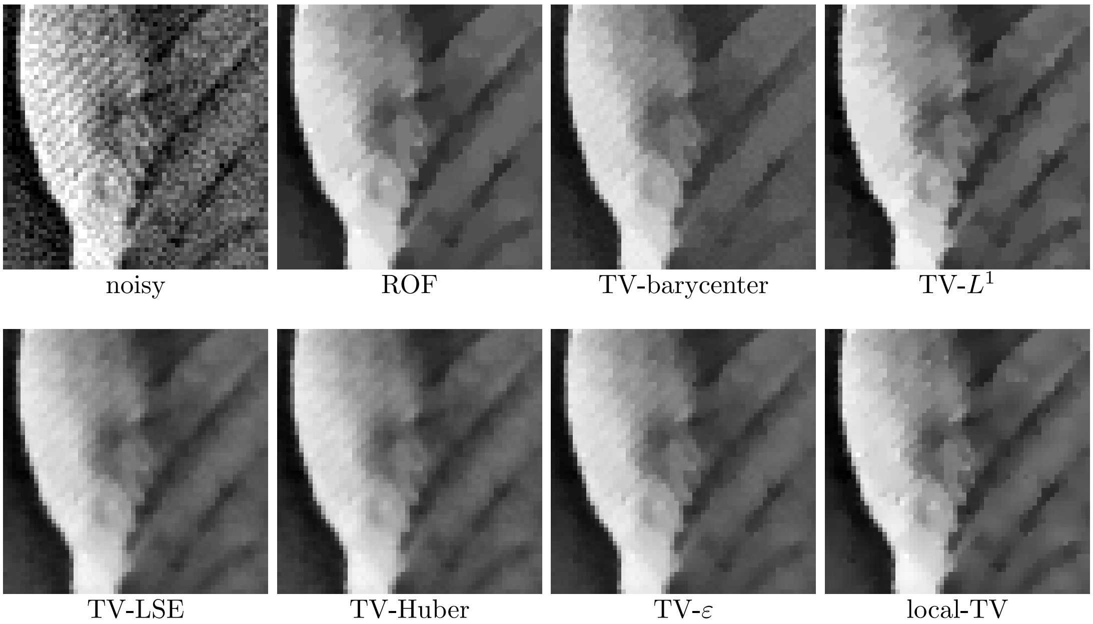
In this paper, we studied the TV-LSE variant of the Rudin-Osher-Fatemi (ROF) denoising model, which consists in estimating the expectation of the Bayesian posterior distribution rather than the image with highest posterior density (MAP). We proved, among other properties, that this denoising scheme avoids one major drawback of the classical ROF model, that is, the staircasing effect. This shows in particular that the staircasing observed with the classical ROF model is not a consequence of the TV term, but rather a model distortion due to the MAP framework, as Nikolova pointed out in [51]. As mentioned in the introduction, the posterior expectation often goes along with a better preservation of local statistics:



noisy ROF TV-barycenter TV-L^1

TV-LSE TV-Huber TV-ε local-TV

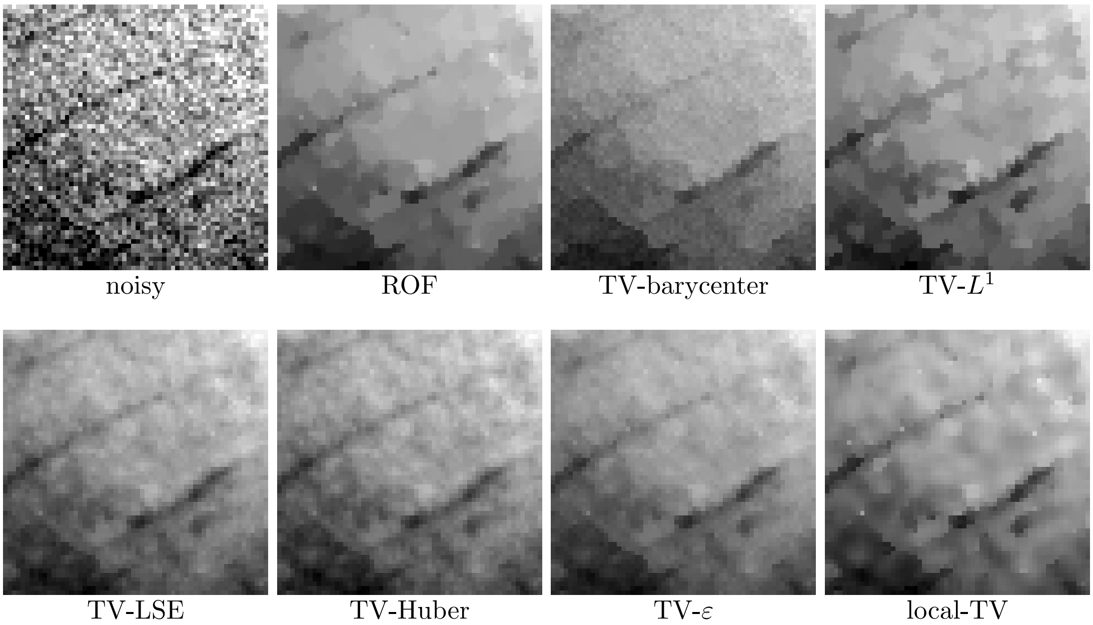
Figure 10: **Comparison of TV-LSE denoising to other TV-based denoising methods.** The Lena image is corrupted with an additive Gaussian noise with standard deviation equal to 10, and the resulting noisy image (detail on top, left) is first processed with TV-LSE using the parameters $(\sigma, \lambda) = (10, 30)$, then processed with the other above-mentioned methods. The fixed parameters for these other methods are: $\varepsilon = 10$ for TV-ε, $\alpha = 10$ for TV-Huber, while for local-TV 5×5 patches are used together with Gaussian weights with parameter $a = 2$. The remaining parameter of each method is adjusted in such a way that the resulting method noise (norm of the estimated noise image) equals the one of TV-LSE, which leads to: $\lambda_{MAP} = 17.03$ for ROF, $t = 0.87$ for TV-barycenter, $\lambda_{L^1} = 0.80$ for TV-L^1, $\lambda_{\text{Huber}} = 28.78$ for TV-Huber, and $\lambda_{\text{loc}} = 15.54$ for local-TV. The 3 results appearing in the first row (ROF, TV-barycenter and TV-L^1) all suffer from staircasing artifacts, visible in particular as spurious contrasted edges. On the second row, staircasing is avoided but TV-LSE and TV-Huber lead to better quality (and very similar) images compared to TV-ε and local-TV. Note that these pictures only show a detail of the Lena image (processed as a whole). Zooms on other details are given in Figure 11 and 12.



noisy ROF TV-barycenter TV-L^1

TV-LSE TV-Huber TV-ε local-TV

Figure 11: A second detail of Lena, denoised with various TV-based methods, as in Figure 10. The conclusions are similar: notice in particular how the stripes (top-left image part) are better restored with the TV-LSE and TV-Huber methods.



noisy ROF TV-barycenter TV-L^1

TV-LSE TV-Huber TV-ε local-TV

Figure 12: A third detail of Lena, denoised with various TV-based methods, as in Figure 10.

this is somehow the case for the gradient norm of the denoised images, that, in the TV-LSE variant, avoids the strong peak in 0 observed with the ROF model.

These theoretical properties have a direct consequence on the visual quality of the denoised images, that show a nice combination of sharp edges (the most interesting property of the TV functional) and the absence of staircase (piecewise constant) regions. In that sense, the TV-LSE model favorably compares to other TV-based denoising methods, as was shown in Section 4. Numerical experiments also revealed that the results of the TV-LSE model can be, for a certain range of parameters, very close to the images produced by the TV-Huber method, which sheds light on the latter model, and more generally on modifications of the ROF energy that would lead to good approximations of the TV-LSE method.

Beyond its use in the TV-LSE denoising variant, the theoretical and numerical framework introduced here opens interesting perspectives, not only for other restoration tasks such as deblurring, zooming, inpainting, etc. that could also be reformulated in a TV-LSE setting, but also because a very similar algorithm could be used to compute the LSE variant associated with other (non-necessarily convex) functionals, or to explore other statistics (median, maximum of marginal distribution, etc.) of the posterior distribution.

A Appendix

A.1 Mild assumptions for TV scheme

Throughout the paper, TV is assumed to be of the form

$$TV(u) = \sum_{\mathbf{x}\in\Omega} \sqrt{(Du(\mathbf{x})_1)^2 + (Du(\mathbf{x})_2)^2} \quad (\ell^2 \text{ formulation})$$

or

$$TV(u) = \sum_{\mathbf{x}\in\Omega} (|Du(\mathbf{x})_1| + |Du(\mathbf{x})_2|) \quad (\ell^1 \text{ formulation}).$$

But the only requirements we really need in the results of Section 3 are the following (and are met by both the ℓ^1 and ℓ^2 formulations):

(A1) The TV operator maps $\mathbb{R}^\Omega$ on $\mathbb{R}\cup\{+\infty\}$; it is non-negative, convex and Lipschitz continuous (so that its domain $\{u \in \mathbb{R}^\Omega, TV(u) < +\infty\}$ has a non-empty interior).

(A2) The TV operator is positively homogeneous, i.e. for every $u \in \mathbb{R}^\Omega$ and every $\alpha \in \mathbb{R}$, we have $TV(\alpha u) = |\alpha| TV(u)$.

(A3) The TV operator is shift-invariant, i.e. for every $c \in \mathbb{R}$ and every $u \in \mathbb{R}^\Omega$, we have $TV(u + c) = TV(u)$.

(A4) The TV operator satisfies the discrete form of Poincaré inequality, i.e. there exists $C > 0$ such that

$$\forall u \in \mathbb{R}^\Omega,\ \|u - \bar{u}\| \le C\, TV(u),$$

where $\bar{u}$ is the mean of u on Ω.

In particular, any norm on the space $\mathcal{E}_0$ of zero mean images, extended by shift invariance on $\mathbb{R}^\Omega$, suits to these assumptions. For example, if $(\varphi_{j,k})$ is any wavelet basis on the finite-dimensional space $\mathbb{R}^\Omega$, the function

$$F_{p,q;s}(u) = \left(\sum_j 2^{-js/2} \left(\sum_k |\langle u, \varphi_{j,k}\rangle|^p \right)^{q/p} \right)^{1/q},$$

corresponding to the discretization of a homogeneous Besov semi-norm $\|\cdot\|_{\dot{B}^s_{p,q}}$, fits the assumptions.

A.2 Proof of Lemma 1

Lemma 5 *Let $P \in \mathbb{R}[X_1, \dots, X_n]$ be a polynomial. Let p be a bounded probability density function. Let $F_P : \mathbb{R}^n \to \mathbb{R}$ be such that*

$$F_P : v \mapsto \int_{\mathbb{R}^n} P(u_1, \dots, u_n)\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du. \tag{40}$$

Then F_P is continuous and differentiable. Its derivative along the direction h is given by

$$dF_P(v)(h) = \int_{\mathbb{R}^n} \frac{\langle u-v, h\rangle}{\sigma^2} P(u_1, \dots, u_n)\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du.$$

Proof —In this proof, when $u \in \mathbb{R}^n$, we shall write $P(u)$ for $P(u_1, \dots, u_n)$ for concision. Let us start by showing that F_P is continuous, by applying the continuity theorem under the integral sign. Let g be defined by

$$g : (u, v) \mapsto P(u)\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u) \tag{41}$$

The mapping $v \mapsto g(u, v)$ is continuous. Now, note that if h is a unit vector of $\mathbb{R}^n$ then

$$|t| < \varepsilon \quad \Rightarrow \quad \|u - v - th\|^2 \geq \frac{1}{2}\|u-v\|^2 - \varepsilon^2. \tag{42}$$

Let $v \in \mathbb{R}^n$ and $\varepsilon > 0$. Let us denote $B(v, \varepsilon)$ the set of v' satisfying $\|v' - v\| \leq \varepsilon$. The mapping $g(u, \cdot)$ has an upper bound on $B(v, \varepsilon)$ thanks to (42) given by

$$\forall v' \in B(v, \varepsilon), \quad |g(u, v')| \leq |P(u)| e^{-\frac{\frac{1}{2}\|u-v\|^2 - \varepsilon^2}{2\sigma^2}} p(u),$$

which is an upper bound independent of $v' \in B(v, \varepsilon)$, and $g(u, \cdot)$ is in $L^1(\mathbb{R}^n)$ since p is bounded (i.e. $v \mapsto g(u, v)$ is locally (in v) uniformly bounded by an integrable function). Hence the continuity theorem under the integral sign applies, and F_P is continuous.

To prove the differentiability of F_P, let h be a unit vector of $\mathbb{R}^n$ and $\varepsilon > 0$. The function

$$t \in (-\varepsilon, \varepsilon) \mapsto P(u)\, e^{-\frac{\|u-v-th\|^2}{2\sigma^2}}\, p(u),$$

is $\mathcal{C}^1$, with derivative

$$t \mapsto \frac{\langle u-v, h\rangle - t}{\sigma^2} P(u)\, e^{-\frac{\|u-v-th\|^2}{2\sigma^2}}\, p(u),$$

and satisfies, thanks to (42),

$$\left| \frac{\langle u-v, h\rangle - t}{\sigma^2} P(u)\, e^{-\frac{\|u-v-th\|^2}{2\sigma^2}}\, p(u) \right| \leq \frac{\|u-v\| + \varepsilon}{\sigma^2} |P(u)|\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, e^{\frac{\varepsilon^2}{2\sigma^2}}\, p(u).$$

This bound is independent of t (provided that $|t| < \varepsilon$) and $h \in B(0, 1)$, and is integrable with respect to $u \in \mathbb{R}^n$ since the Gaussian distribution admits finite moments of order 1 and 2. Now thanks to the derivation theorem under the integral sign, the mapping $t \mapsto F_P(v + th)$ is differentiable at 0, then F_P is differentiable and its differential writes

$$dF_P(v)(h) = \frac{\partial}{\partial t} \int_{\mathbb{R}^n} P(u)\, e^{-\frac{\|u-v-th\|^2}{2\sigma^2}}\, p(u)\, du \bigg|_{t=0} = \int_{\mathbb{R}^n} \frac{\langle u-v, h\rangle}{\sigma^2} P(u)\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du,$$

which was the desired result. □

Proof of Lemma 1 — S_{LSE} is the division of two functions of the type F_P (40), with $P = X$ for the numerator and $P = 1$ for the denominator (leading to a positive value). Thanks to Lemma 5, F_P is continuous and differentiable in both cases, and finally S_{LSE} benefits from this regularity too.

Again thanks to Lemma 5,

$$\begin{aligned}\sigma^2\, dS_{\mathrm{LSE}}(v)(h) &= \frac{\int \langle h, u-v\rangle\, u\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}{\int \exp(-\frac{\|u-v\|^2}{2\sigma^2})\, p(u)\, du} - \frac{\int \langle h, u-v\rangle\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}{\int e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}\frac{\int u e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}{\int e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du} \\ &= \frac{\int \langle h, u\rangle\, u\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}{\int \exp(-\frac{\|u-v\|^2}{2\sigma^2})\, p(u)\, du} - \frac{\int \langle h, u\rangle\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}{\int e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}\frac{\int u e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}{\int e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du}.\end{aligned}$$

The differential $dS_{\mathrm{LSE}}(v)$ can be interpreted as a covariance matrix

$$\sigma^2\, dS_{\mathrm{LSE}}(v) = \mathbb{E}[Z_v Z_v^T] - \mathbb{E}Z_v\ \mathbb{E}Z_v^T = \mathrm{Cov} Z_v,$$

where Z_v follows a distribution with density $q_v(u) = \frac{1}{Z}e^{-\frac{\|u-v\|^2}{2\sigma^2}}p(u)$. Indeed, for each $h \in \mathbb{R}^n$,

$$\begin{aligned}(\mathrm{Cov}Z_v)h &= \mathbb{E}[Z_v Z_v^T h] - \mathbb{E}Z_v\mathbb{E}[Z_v^T h] \\ &= \mathbb{E}[\langle h, Z_v\rangle\, Z_v] - \mathbb{E}\,\langle h, Z_v\rangle\, \mathbb{E}Z_v,\end{aligned}$$

where we can recognize $\sigma^2\, dS_{\mathrm{LSE}}(v)(h)$. In particular, $dS_{\mathrm{LSE}}(v)$ is symmetric with non-negative eigenvalues. Let us prove now that $dS_{\mathrm{LSE}}(v)$ is positive-definite. To that end, let us assume that there exists a vector $h \neq 0$ in the kernel of $dS_{\mathrm{LSE}}(v)$, i.e. such that

$$(\mathrm{Cov}Z_v)h = 0.$$

Then multiplying on the left by h^T yields

$$h^T(\mathrm{Cov}Z_v)h = \mathrm{var}\,\langle h, Z_v\rangle = 0.$$

But the support of distribution q_v satisfies

$$\mathrm{Supp}(q_v) = \mathrm{Supp}(p) = \{v \in \mathbb{R}^n \mid f(v) < \infty\},$$

which has non-empty interior. Then $\langle h, Z_v\rangle$ cannot have a zero variance, and we obtain a contradiction. Finally $dS_{\mathrm{LSE}}(v)$ is a symmetric positive-definite matrix. □

A.3 Proof of Lemma 4

For every $v \in \mathbb{R}^n$, the triangle inequality applied to $S_{\mathrm{LSE}}(v) - v$ leads to

$$\begin{aligned}\|S_{\mathrm{LSE}}(v) - v\| &\leq \frac{\int \|u-v\| e^{-\frac{\|u-v\|^2}{2\sigma^2}} p(u)\, du}{\int e^{-\frac{\|u-v\|^2}{2\sigma^2}} p(u)\, du} \\ &\leq \frac{\int \|u\| e^{-\frac{\|u\|^2}{2\sigma^2}} p(v+u)\, du}{\int e^{-\frac{\|u\|^2}{2\sigma^2}} p(v+u)\, du}.\end{aligned}$$

Now since the potential $f = -\log p$ of the prior probability is Lipschitz continuous, we have

$$\exists k > 0, \quad \forall u, v \in \mathbb{R}^n,\ |f(v+u) - f(v)| \leq k\|u\|,$$

so that

$$p(v)e^{-k\|u\|} \leq p(v+u) \leq p(v)e^{k\|u\|},$$

each side remaining positive. This allows us to bound the expression by

$$\|S_{\mathrm{LSE}}(v) - v\| \leq \frac{\int \|u\| e^{-\frac{\|u\|^2}{2\sigma^2}} e^{k\|u\|} p(v)\, du}{\int e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-k\|u\|} p(v)\, du},$$

which simplifies into

$$\|S_{\mathrm{LSE}}(v) - v\| \leq \frac{\int \|u\| e^{-\frac{\|u\|^2}{2\sigma^2}} e^{k\|u\|}\, du}{\int e^{-\frac{\|u\|^2}{2\sigma^2}} e^{-k\|u\|}\, du}$$

which is finite and independent of v, which proves the boundedness of $S_{\mathrm{LSE}} - I$.

If the dimension $n = |\Omega|$ is equal to 1, then S_{LSE} is continuous and $S_{\mathrm{LSE}} - I$ is bounded, and thanks to the intermediate value theorem, S_{LSE} is onto. Now if $n \geq 2$, as $S_{\mathrm{LSE}} - I$ is bounded, it is straightforward that

$$\lim_{\|v\| \to \infty} \frac{| \langle S_{\mathrm{LSE}}(v), v \rangle |}{\|v\|} = +\infty, \tag{43}$$

so we can apply Corollary 16 of [11]: since S_{LSE} is continuous and satisfies (43) and

$$\forall v_1, v_2 \in \mathbb{R}^n, \quad \langle S_{\mathrm{LSE}}(v_2) - S_{\mathrm{LSE}}(v_1), v_2 - v_1 \rangle \geq 0$$

(monotony in the sense of Brezis, which is a weaker form of (31)), we conclude that S_{LSE} is onto. □

References

[1] L. Ambrosio, N. Fusco, and D. Pallara. *Functions of bounded variation and free discontinuity problems.* Oxford University Press, March 2000.

[2] G. Aubert and P. Kornprobst. *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations (second edition)*, volume 147 of *Applied Mathematical Sciences.* Springer-Verlag, 2006.

[3] G. Aubert and L. Vese. Variational methods in image restauration. *SIAM Journal of Numerical Analysis*, 34(5):1948–1979, 1997.

[4] J.-F. Aujol, G. Aubert, L. Blanc-Féraud, and A. Chambolle. Image decomposition into a bounded variation component and an oscillating component. *J. Math. Imaging Vision*, 22:71–88, 2005.

[5] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher. Structure-texture image decomposition - modeling, algorithms, and parameter selection. *International Journal of Computer Vision*, 67(1):111–136, 2006.

[6] M. Bergounioux and L. Piffet. A second-order model for image denoising. *Set Valued and Variational Analysis*, 18(3-4):483–503, 2010.

[7] J. Besag. Digital image processing : Towards Bayesian image analysis. *Journal of Applied Statistics*, 16(3):395–407, 1989.

[8] G. Blanchet and L. Moisan. An explicit sharpness index related to global phase coherence. *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2012.

[9] C. Bouman and K. Sauer. A generalized Gaussian image model for edge-preserving MAP estimation. *IEEE Trans. on Image Processing*, 2(3):296–310, 1993.

[10] K. Bredies, K. Kunisch, and T. Pock. Total generalized variation. *SIAM Journal on Imaging Sciences*, 3(3):492–526, 2010.

[11] H. R. Brezis. Les opérateurs monotones. *Séminaire Choquet, Initiation à l'analyse*, 5(2), 1965-1966.

[12] V. Caselles, A. Chambolle, and M. Novaga. The discontinuity set of solutions of the TV denoising problem and some extensions. *Multiscale Model. Simul.*, 6(3):879–894, 2007.

[13] V. Caselles, A. Chambolle, and M. Novaga. Total variation in imaging. In *Handbook of Mathematical Methods in Imaging*, pages 1016–1057. Springer, 2011.

[14] L. Chaari. *Parallel magnetic resonance imaging reconstruction problems using wavelet representations.* PhD thesis, Université Paris Est, France. Advisor J.-C. Pesquet, 2010.

[15] L. Chaari, J.-C. Pesquet, J.-Y. Tourneret, and P. Ciuciu. Parameter estimation for hybrid wavelet-total variation regularization. *IEEE Statistical Signal Processing Workshop (SSP)*, pages 461–464, 2011.

[16] A. Chambolle, V. Caselles, D. Cremers, M. Novaga, and T. Pock. An introduction to total variation for image analysis. In *Theoretical Foundations and Numerical Methods for Sparse Recovery.* De Gruyter, 2010.

[17] A. Chambolle and T. Pock. A first-order primal-dual algorithm for convex problems with applications to imaging. *Journal of Mathematical Imaging and Vision*, 40:120–145, 2011.

[18] T. Chan, S. Esedoglu, and M. Nikolova. Algorithms for finding global minimizers of image segmentation and denoising models. *SIAM Journal of Applied Mathematics*, 66(5):1632–1648, 2006.

[19] T. Chan, S. Esedoglu, and F. Park. Image decomposition combining staircase reduction and texture extraction. *Journal of Visual Communication and Image Representation*, 18(6):464–486, 2007.

[20] T. Chan and J. Shen. Mathematical models of local non-texture inpaintings. *SIAM Journal of Applied Mathematics*, 62:1019–1043, 2001.

[21] T. F. Chan, S. Esedoglu, and F. E. Park. A fourth order dual method for staircase reduction in texture extraction and image restoration problems. In *Image Processing, IEEE International Conference*, pages 4137–4140, 2010.

[22] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising with block-matching and 3D filtering. In *Electronic Imaging'06, Proc. SPIE 6064*, volume 30, 2006.

[23] D. C. Dobson and F. Santosa. Recovery of blocky images from noisy and blurred data. *SIAM J. Appl. Math.*, 56(4):1181–1198, 1996.

[24] V. Duval, J.-F. Aujol, and Y. Gousseau. The TVL^1 model: a geometric point of view. *Multiscale Model. Simul.*, 8(1):154–189, 2009.

[25] C. Fox and G. K. Nicholls. Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited. *Bayesian inference and maximum entropy methods in science and engineering: 20th International Workshop*, 2000.

[26] D. Geman. Random fields and inverse problems in imaging. In *École d'Été de Probabilités de Saint-Flour XVIII*, pages 117–193. Springer-Verlag, Lecture Notes in Mathematics 1427, 1988.

[27] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in computer vision: issues, problems, principles, and paradigms*, pages 564–584. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987.

[28] C. J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473–483, 1992.

[29] R. Gribonval. Should penalized least squares regression be interpreted as Maximum A Posteriori estimation? *IEEE Transactions on Signal Processing*, 59(5):2405–2410, May 2011.

[30] F. Guichard and F. Malgouyres. Total variation based interpolation. *Proceedings of the European Signal Processing Conference*, 3:1741–1744, 1998.

[31] K. Jalalzai. *Regularization of inverse problems in image processing.* PhD thesis, Ecole Polytechnique, France, 2012.

[32] K. Jalalzai and A. Chambolle. Enhancement of blurred and noisy images based on an original variant of the total variation. *Scale Space and Variational Methods in Computer Vision, Lecture Notes in Computer Science*, 5567:368–376, 2009.

[33] V. Kolehmainen, M. Lassas, K. Niinimäki, and S. Siltanen. Sparsity-promoting Bayesian inversion. *Inverse Problems*, 28(2):025005, 2012.

[34] M. Lassas, E. Saksman, and S. Siltanen. Discretization invariant Bayesian inversion and Besov space priors. *Inverse Problems and Imaging*, 3(1):87–122, 2009.

[35] M. Lassas and S. Siltanen. Can one use total variation prior for edge-preserving Bayesian inversion? *Inverse Problems*, 20(5):1537–1563, 2004.

[36] M. Ledoux. Measure concentration, transportation cost, and functional inequalities. *Instructional Conference on Combinatorial Aspects of Mathematical Analysis, Edinburgh, 25 March-5 April 2002 and Summer School on Singular Phenomena and Scaling in Mathematical Models, Bonn, 10-13 June 2003*, 2003.

[37] M. Ledoux. *The concentration of measure phenomenon.* Mathematical Surveys and Monographs, Second printing. American Mathematical Society, Providence, 2005.

[38] S. Lefkimmiatis, A. Bourquard, and M. Unser. Hessian-based norm regularization for image restoration with biomedical applications. *IEEE Transactions on Image Processing*, 21(3):983–995, 2012.

[39] C. Louchet. *Variational and Bayesian models for image denoising: from total variation towards non-local means.* PhD thesis, Université Paris Descartes, France, 2008.

[40] C. Louchet and L. Moisan. Total variation denoising using posterior expectation. In *Proceedings of the European Signal Processing Conference (Eusipco).* Eurasip, 2008.

[41] C. Louchet and L. Moisan. Total variation as a local filter. *SIAM Journal on Imaging Sciences*, 4(2):651–694, 2011.

[42] B. Luo, J.-F. Aujol, and Y. Gousseau. Local scale measure from the topographic map and application to remote sensing images. *SIAM Journal on Multiscale Modeling and Simulation*, 8(1):1–29, 2009.

[43] G. J. Minty. Monotone (nonlinear) operators in Hilbert space. *Duke Math. J.*, 29:341–346, 1962.

[44] J.-M. Mirebeau and A. Cohen. Anisotropic smoothness classes: From finite element approximation to image models. *Journal of Mathematical Imaging and Vision*, 38:52–69, 2010. 10.1007/s10851-010-0210-x.

[45] J.-J. Moreau. Proximité et dualité dans un espace hilbertien. *Bulletin de la Société Mathématique de France*, 93:273–299, 1965.

[46] M. Nikolova. Estimées localement fortement homogènes. *Compte-rendus de l'Académie des Sciences*, 325:665–670, 1997.

[47] M. Nikolova. Local strong homogeneity of a regularized estimator. *SIAM J. Appl. Math.*, 61(2):633–658 (electronic), 2000.

[48] M. Nikolova. A variational approach to remove outliers and impulse noise. *J. Math. Imaging Vision*, 20:99–120, 2004.

[49] M. Nikolova. Weakly constrained minimization: application to the estimation of images and signals involving constant regions. *J. Math. Imaging Vision*, 21(2):155–175, 2004.

[50] M. Nikolova. Analysis of the recovery of edges in images and signals by minimizing nonconvex regularized least-squares. *Multiscale Model. Simul.*, 4(3):960–991 (electronic), 2005.

[51] M. Nikolova. Model distortions in Bayesian MAP reconstruction. *Inverse Probl. Imaging*, 1(2):399–422, 2007.

[52] G. Papandreou and A. L. Yuille. Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models. In *International Conference on Computer Vision (ICCV)*, pages 193–200. IEEE, 2011.

[53] P. Pletscher, S. Nowozin, P. Kohli, and C. Rother. Putting MAP back on the map. In *33rd Annual Symposium of the German Association for Pattern Recognition (DAGM)*, 2011.

[54] A. Prékopa. On logarithmic concave measures and functions. *Acta Scientiarum Mathematicarum*, 34:335–343, 1973.

[55] W. Ring. Structural properties of solutions of total variation regularization problems. *ESSAIM, Math Modelling and Numerical Analysis*, 34:799–840, 2000.

[56] R. T. Rockafellar. *Convex Analysis.* Princeton University Press, 1970.

[57] S. Roth and M. J. Black. Fields of experts. *International Journal of Computer Vision (IJCV)*, 82(2):205–229, 2009.

[58] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.

[59] H. Rue and M. A. Hurn. Loss functions for Bayesian image analysis. *Scandinavian Journal of Statistics*, 24:103–114, 1997.

[60] J. Salmon. *Agrégation d'estimateurs et méthodes à patch pour le débruitage d'images numériques.* These, Université Paris-Diderot, advisors Erwan Le Pennec and Dominique Picard, December 2010.

[61] U. Schmidt, Q. Gao, and S. Roth. A generative perspective on MRFs in low-level vision. In *Proc. of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, California*, 2010.

[62] H. V. T. Si. Une remarque sur la formule du changement de variables dans R^n. *Bulletin de la Société Royale des Sciences de Liège*, 73(1):21–25, 2004.

[63] J. Tropp. Just relax: convex programming methods for identifying sparse signals in noise. *IEEE Transactions on Information Theory*, 52(3):1030–1051, 2006.

[64] P. Weiss, L. Blanc-Féraud, and G. Aubert. Efficient schemes for total variation minimization under constraints in image processing. *SIAM journal on Scientific Computing*, 31(3):2047–2080, 2009.

[65] O. J. Woodford, C. Rother, and V. Kolmogorov. A global perspective on MAP inference for low-level vision. *International Conference on Computer Vision (ICCV)*, pages 2319–2326, 2009.

[66] C. Zach, T. Pock, and H. Bischof. A duality based approach for realtime TV-L^1 optical flow. *29th DAGM Symposium on Pattern Recognition, Heidelberg, Germany*, pages 214–223, 2007.