SIAM J. IMAGING SCIENCES
Vol. 6, No. 4, pp. 2640–2684
© 2013 Society for Industrial and Applied Mathematics

Posterior Expectation of the Total Variation Model: Properties and Experiments*

Cécile Louchet† and Lionel Moisan‡

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 the "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. We present theoretical and numerical results that demonstrate in particular that images denoised with this model do not suffer from the staircasing effect.

Key words. image denoising, total variation, Bayesian model, least square estimate, maximum a posteriori, estimation in high-dimensional spaces, proximity operators, staircasing effect

AMS subject classifications. 68U10, 62H35

DOI. 10.1137/120902276

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 [64], consists in solving an inverse problem like $Au = v$ (where v is the observed image, u is the unknown ideal image, and A is a given operator) by minimizing the energy

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

This energy establishes 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

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

which boils down to

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

*Received by the editors December 12, 2012; accepted for publication (in revised form) August 26, 2013; published electronically December 17, 2013.
http://www.siam.org/journals/siims/6-4/90227.html
†MAPMO, Université d'Orléans, CNRS, UMR 6628, 45067 Orléans, France (Cecile.Louchet@univ-orleans.fr).
‡MAP5, Université Paris Descartes, CNRS UMR 8145, 75006 Paris, France (Lionel.Moisan@parisdescartes.fr).

for smooth images. Depending on the choice of A, (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), superresolution (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: image inpainting [23], interpolation [35], segmentation [20], image quality assessment [8], scale detection [49], cartoon+texture decomposition [4, 5], motion estimation [76], and many others (and also, of course, in applications that do not concern images). For a more complete list of applications concerning image processing and the TV model, we invite the interested reader to consult [15, 18] 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 [63]), TV still frequently appears in nonlinear 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 that 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), which is more suitable in the case of impulse noise, we even have a contrast-invariant transform [26], which processes the level sets of u independently. This nice analytical framework around TV and bounded variation (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 [24] 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, which are very oscillatory in general, are highly penalized by TV and are often destroyed (or at least strongly attenuated) by the ROF model. Another well-known artifact is the staircasing effect.

The staircasing effect. A simple example of the staircasing effect is obtained when a noisy affine one-dimensional signal is processed with the ROF model: the denoised signal is not smooth but piecewise constant (see, e.g., the ROF-denoised signal that is displayed in the middle plot of Figure 2). The staircase shape obtained in this case is quite general: as first noticed in [25] and then analyzed by [11, 15, 33, 57, 61] in different frameworks, the

application of ROF denoising leads to blocky structures on most signals and images; this phenomenon is called the staircasing effect. As far as we know, all image processing papers discussing this effect on natural images consider it an artifact rather than a desirable property [6, 10, 13, 15, 21, 22, 45]. Indeed, in [36], histograms of local features (such as directional gradients) computed on a large bunch of natural images are faithfully modeled by generalized Laplace distributions with a probability density function (p.d.f.) proportional to $\exp(-s|x|^\alpha)$ with $\alpha \approx 0.55$. An image restored with the ROF model will have a potentially large proportion of pixels with a strictly zero gradient, thus adding a Dirac mass to the generalized Laplace distribution and contradicting the observed statistics of natural images. Furthermore, the staircasing effect is incompatible with the Shannon sampling theory [68] in the sense that a nonconstant bandlimited image (in the continuous domain) cannot be locally constant. Even from a physical image formation viewpoint, truly constant regions are rare: because of little variations in illumination, orientation, or simply because of the perspective, a minimum drift of gray levels is generally observed. Last, the staircasing effect comes with the creation of spurious edges in what should be smooth areas. Using an ROF filter as a preprocessing step before a higher-level processing, such as segmentation, may be a source of systematic failures.

It is commonly admitted that the staircasing effect is due to the nonregularity of TV [53, 55, 57], which, in the discrete framework, comes from the singularity of TV at zero gradients. More than that, under several hypotheses [71], the ROF model is equivalent to minimizing an energy like (1), but where the TV (the ℓ^1-norm of the 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 in [55]. Among variants that have been proposed [3, 7, 28, 56, 74], 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 nondecreasing on $\mathbb{R}^+$. More recently, different authors managed to promote sparsity for higher-order derivatives [6, 10, 21, 22, 45], leading to piecewise affine or piecewise polynomial images (hence pushing the staircasing effect to higher orders). In [38], 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 p.d.f. of a certain distribution (the posterior distribution, associated to the TV prior and the data-fidelity term). Several authors [57, 75] 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.

Let us also mention the nonlocal extension of ROF (NL-ROF) proposed in [32], where the neighborhood between pixels is based on a comparison of patches as in the NL-means method [13]. As the authors notice in their experiments, it clearly improves the basic ROF model, but produces staircasing effects. Another NL extension of ROF (NLBV) proposed in [39] avoids staircasing artifacts by considering differences of gradients instead of differences of gray values, in a spirit similar to that of the second-order TV models considered in [6, 10].

In the present work, we propose keeping the statistical framework associated to the ROF model, but moving away from MAP estimation and considering instead the mean of the posterior distribution (rather than its maximum). As in the preliminary work [47], 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 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 [27] 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 nonrobust estimate. Lassas, Siltanen, and colleagues [40, 42, 41] focus on one-dimensional 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 lead only to trivial estimates (MAP equal to 0, LSE equivalent to Gaussian smoothing), and they 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 faster by a whole world of energy minimization algorithms, contrary to LSE which requires Monte-Carlo Markov chain algorithms [31] or Gibbs samplers [29], which are known to be slow. This computational issue can motivate one to use MAP instead of LSE or, more interestingly, to see an LSE estimate as a MAP estimate in another Bayesian framework, as was done in [34] and [58].

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 [58, 63, 67], seem to recover the prior statistics in a better way than MAP estimates. But in [59], it is argued that the prior learning method (maximum margin principle or maximum likelihood) has to be connected to the estimation function: maximum likelihood seems to perform better while associated to an LSE estimator, but learning with a maximum margin principle seems to perform even better while associated to a MAP estimator.

Since the preliminary work [47] in 2008, several researchers have taken an interest in TV-LSE. Jalalzai and Chambolle [38], Lefkimmiatis, Bourquard, and Unser [45], and Salmon [66] mention the TV-LSE model for its ability to naturally remove staircasing artifacts. In the conclusion of [51], Mirebeau and Cohen propose a TV-LSE–like approach to denoising images using anisotropic smoothness features, an interesting counterpart to TV, arguing that LSE

is able to deal with nonconvex functionals. Chaari and colleagues propose LSE estimates for a frame-based Bayesian denoising task [16] and for a parameter estimation task [17], 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 [15], 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 justify 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 sections 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 $\Omega \subset \mathbb{Z}^2$, which maps each pixel $\mathbf{x} = (x, y) \in \Omega$ to 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 what follows 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). 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 selecting 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 p.d.f.

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

$$\text{and} \quad \forall \mu \in \mathbb{R}, \quad \mathcal{E}_\mu = \left\{ u \in \mathbb{R}^\Omega,\ \bar{u} = \mu \right\} \quad \text{with} \quad \bar{u} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} u(\mathbf{x}). \tag{7}$$

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 variance σ^2. Applying Bayes' rule with prior distribution p_β leads to the following posterior p.d.f.:

$$p(u|v) = \frac{p(v|u)p_\beta(u)}{p(v)} = \frac{1}{Z} \exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2} \right), \tag{8}$$

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 MAP

$$\hat{u}_{\mathrm{ROF}} = \arg\max_{u \in \mathcal{E}_\mu} p(u|v). \tag{9}$$

This means that ROF denoising amounts to selecting the most probable image under the posterior probability defined by $p(u|v)$. Notice that the constraint $u \in \mathcal{E}_\mu$, which 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 the 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 (8) shows, the solution does not depend on σ, which measures the "spread" of the posterior distribution. As $\hat{u}_{\mathrm{ROF}}$ 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. In order to establish the existence of this staircasing effect for ROF denoising in the discrete setting, Nikolova [55] remarks that the discrete TV operator (3) can be written under the more general form

$$TV(u) = \sum_{i=1}^{r} \varphi_i(G_i u),$$

where $G_i : \mathbb{R}^\Omega \to \mathbb{R}^m$ $(1 \le i \le r)$ are linear operators (here, differences between neighboring pixels), and $\varphi_i : \mathbb{R}^m \to \mathbb{R}$ are piecewise smooth functions that are not differentiable in zero (here, all φ_i correspond to the L^1-norm on $\mathbb{R}^2$).

Proposition 2.1 (see [55]). *If $S(v) = \arg\min_u \|u - v\|^2 + \lambda J(u)$, where $J(u) = \sum_{i=1}^r \varphi_i(G_i u)$ and, for all i, G_i is linear and φ_i is piecewise smooth and not differentiable in 0, then, under some technical assumptions, there exists an open neighborhood V of v for which*

$$\forall v' \in V, \qquad \left\{ i \mid G_i(S(v')) = 0 \right\} = \left\{ i \mid G_i(S(v)) = 0 \right\}. \tag{10}$$

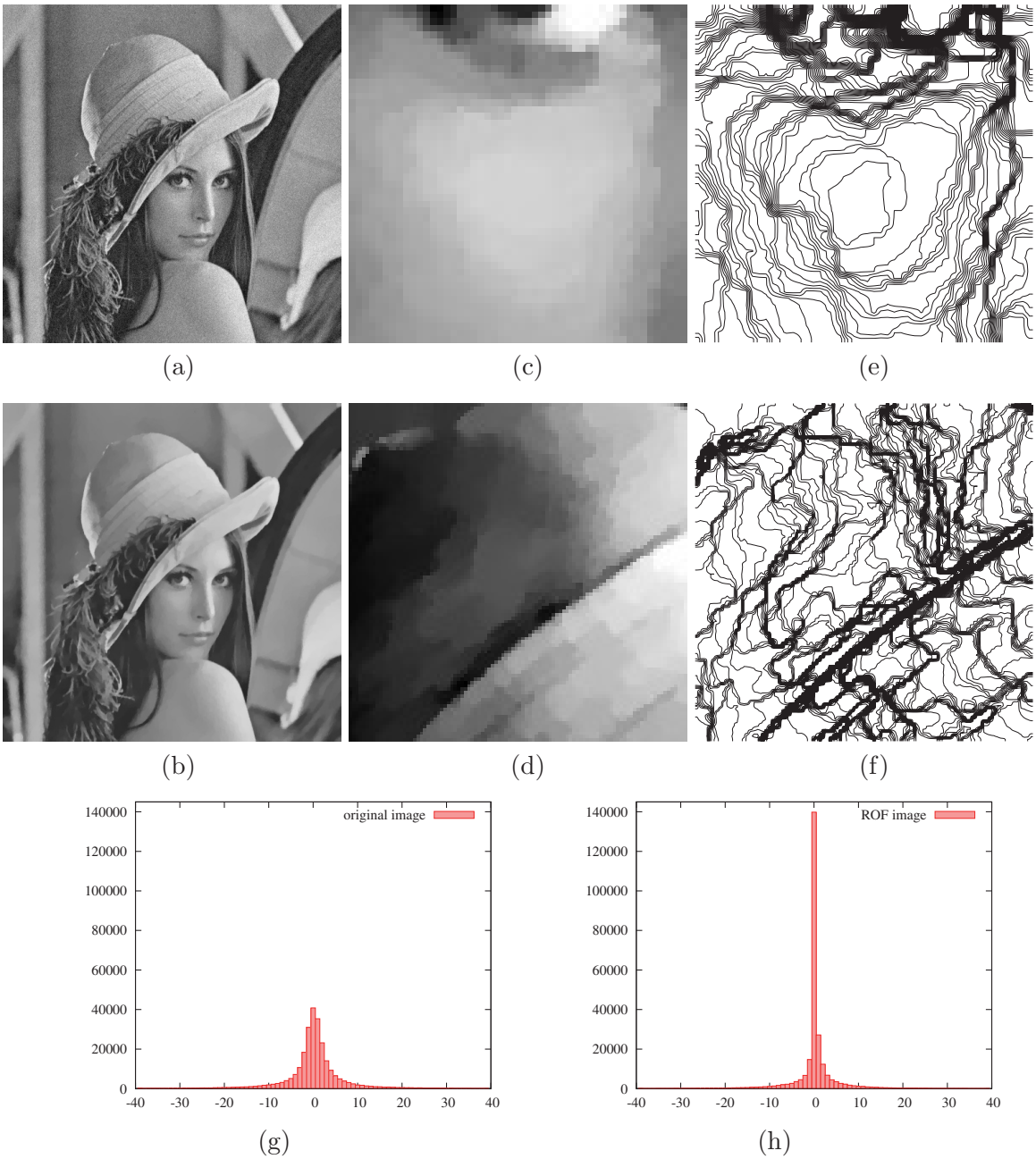

Figure 1. The staircasing effect. *A noisy version* (a) *of the Lena image (additive white Gaussian noise with standard deviation* $\sigma = 10$*) is denoised with the ROF model with* $\lambda = 40$ (b)*. The details* (c) *and* (d) *of* (b) *reveal the so-called staircasing effect: ROF denoising tends 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. The histograms of the horizontal derivative of the original Lena image* (g) *and the ROF-denoised image* (h) *also reveal this staircasing effect: whereas such a histogram is generally well modeled by a generalized Laplace distribution for natural images, the ROF version presents a large peak in* 0 *that is a direct consequence of the staircasing effect. Similar plots would be obtained with the vertical derivative.*

In the case of TV, the G_i are differences between neighboring pixels, so the set $\{i \in \{1,\dots,r\} \mid G_i(S(v)) = 0\}$ corresponds to the constant sets of $S(v) = \hat{u}_{\mathrm{ROF}}$. The existence of an open neighborhood $V = \{v'\}$ of v for which any $S(v')$ has the same constant set as $S(v)$ indicates that the regions of constant gray level have a certain stability with respect to perturbations of the observed data v. This gives a first theoretical explanation of the staircasing effect. In other words, if the space of noisy images $\mathbb{R}^\Omega$ is endowed with a probability distribution that is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^\Omega$, then for any $\mathbf{x} \in \Omega$, the probability of having a zero gradient at pixel $\mathbf{x}$ in the *denoised* image is positive. Hence, there is a bias toward constant regions, which can be measured by a Dirac mass at zero on the histograms of gradients (see Figure 1, plots (g) and (h)).

In the continuous domain, Jalalzai [37] assesses the presence of staircasing by testing the positivity of $|\{\mathbf{x} \in \Omega \mid S(v)(\mathbf{x}) = c\}|$ for some c. He proves that this property occurs for $c = \max S(v)$ and $c = \min S(v)$ when the datum v is in $L^2(\Omega) \cap L^\infty(\Omega)$ and $\Omega = \mathbb{R}^2$. In particular the gradient of $S(v)$ is zero in the interior of $\{\mathbf{x} \in \mathbb{R}^2 \mid S(v)(\mathbf{x}) = c\}$. As this definition is specific to the continuous setting, in what follows we shall rather focus on Nikolova's viewpoint [55].

Let us also cite the recent work of Caselles, Chambolle, and Novaga [14], 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 due only to a bad quantization of the TV. The (approximate) jump set of a continuous image u is defined as the set of points $\mathbf{x} \in \mathbb{R}^2$ satisfying

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

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 ROF denoising problem is contained within the jump set of v. In other words, ROF denoising 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 so that 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 depends only on the location of the mode (that is, the point with maximum density), not

on the probability mass that this mode contains [65], 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–Chebyshev 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

$$\forall \varepsilon > 0, \quad \mathbb{P}\left(\left|\frac{1}{n}\|X\|^2 - \sigma^2\right| > \varepsilon\right) \leq \frac{2\sigma^4}{n\varepsilon^2}, \tag{11}$$

and the right-hand term decreases toward 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" around the sphere centered in 0 with radius $\sigma\sqrt{n}$ 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 whose 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 (8)), which is a log-concave Gibbs field. For such probability distributions, we can have concentration results similar to (11), but they require more sophisticated tools [44].

A key notion in studying the concentrating power of a probability distribution $\boldsymbol{\pi}$ on a Euclidean space $\mathcal{X}$ is the concentration function $\alpha_{\boldsymbol{\pi}}(r)$ [44], defined for each $r > 0$ by

$$\alpha_{\boldsymbol{\pi}}(r) = \sup\left\{1 - \boldsymbol{\pi}(A_r);\ A \text{ Borel set of } \mathbb{R}^\Omega \text{ and } \boldsymbol{\pi}(A) \geq \frac{1}{2}\right\}, \tag{12}$$

where $A_r = \{u \in \mathbb{R}^\Omega,\ d(u, A) < r\}$ (d is the Euclidean distance on $\mathcal{X}$). 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 [43], whose fast decay for large r indicates, thanks to (12), that the probability is very concentrated near 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 the concentration of Lipschitz-continuous functions around their medians. 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 [43]

$$\alpha_{\boldsymbol{\pi}}(r) = \sup_F \boldsymbol{\pi}(F - m_F \geq r), \tag{13}$$

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. (8) of our image denoising problem, we can have a Gaussian concentration inequality similar to the uniform distribution on the unit sphere, as shown in the following proposition.

Proposition 2.2 (concentration property for the ROF posterior distribution). *Let $\boldsymbol{\pi}$ denote the posterior probability with p.d.f. (8). Then*

$$\forall r > 0, \quad \alpha_{\boldsymbol{\pi}}(r) \leq 2e^{-\frac{r^2}{4\sigma^2}}. \tag{14}$$

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

$$\exists c > 0 \quad \forall u, v \in \mathbb{R}^{\Omega}, \quad V(u) + V(v) - 2\,V\left(\frac{u+v}{2}\right) \geq \frac{c}{4}\|u - v\|^2 \tag{15}$$

with $c = 1/\sigma^2$. Then applying [44, Theorem 2.15, p. 36], we obtain (14). ■

This shows that any Lipschitz-continuous function F is concentrated around its median m_F: indeed, combining (13) and (14) yields

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

and the same argument with $-F$ leads to

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

Putting both inequalities together, we deduce that for each $r > 0$,

$$\boldsymbol{\pi}(|F - m_F| \geq r) \leq 4e^{-\frac{\|F\|_{\mathrm{Lip}}^2 r^2}{4\sigma^2}}. \tag{18}$$

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 leads to a weaker form of concentration for the energy.

Proposition 2.3. *Let $\boldsymbol{\pi}$ denote the posterior probability with p.d.f. (8). Assume that v is not constant. Then there exist $m \in \mathbb{R}$ and $c > 0$ such that for any $r \geq 0$,*

$$\boldsymbol{\pi}(|\sqrt{E_{v,\lambda}} - m| \geq r) \leq 4e^{-\frac{c^2 r^2}{4\sigma^2}}. \tag{19}$$

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 are both 1-Lipschitz. This gives the inequality

$$|C_1| \leq \|u'-u\|. \tag{20}$$

To bound C_2, we need to compute the Lipschitz constant of the discrete TV. It depends on the scheme for TV which is used and depends monotonically 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 nonconstant, $E_{v,\lambda}$ 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 nonnegative, we have

$$|C_2| \leq \frac{\lambda|TV(u') - TV(u)|}{0 + \sqrt{\min E_{v,\lambda}}} \leq \frac{\lambda\kappa\sqrt{|\Omega|}\|u'-u\|}{\sqrt{\min E_{v,\lambda}}}. \tag{21}$$

Then, combining (20) and (21), 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 (18). ■

By homogeneity of $\|\cdot - v\|^2$ and TV with respect to 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 $4\exp\left(-\frac{c^2r^2m^2}{4\sigma^2}\right)$, 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 $\hat{u}_{\mathrm{ROF}}$ (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, (19) 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 $\hat{u}_{\mathrm{ROF}}$ 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 using a least square risk, which amounts to searching the image $\hat{u}(v)$ minimizing

$$\mathbb{E}_{u,v}\big(\|u - \hat{u}(v)\|^2\big) = \int_{\mathbb{R}^\Omega} \int_{\mathcal{E}_\mu} \|u - \hat{u}(v)\|^2 p(u,v)\, dv\, du. \tag{22}$$

The image reaching this minimum is the expectation of the posterior distribution (least square estimate (LSE)), that is,

$$\hat{u}_{\mathrm{LSE}} := \mathbb{E}(u|v) = \int_{\mathbb{R}^\Omega} p(u|v)\, u\, du, \tag{23}$$

which, thanks to (8), can be rewritten in the form below.

Definition 2.4. *The TV-LSE operator (denoted by S_{LSE}) maps a discrete image $v \in \mathbb{R}^\Omega$ into the discrete image $\hat{u}_{\mathrm{LSE}}$ defined by*

$$\hat{u}_{\mathrm{LSE}} = S_{\mathrm{LSE}}(v) = \frac{\displaystyle\int_{\mathbb{R}^\Omega} \exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2}\right) \cdot u\, du}{\displaystyle\int_{\mathbb{R}^\Omega} \exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2}\right) du}, \tag{24}$$

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 the least square risk in (22) can lead to other interesting estimates (a median estimate for an 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 $+\infty$. Finally, we use Moreau's theory of *proximations* (or proximity operators) [52, 62] 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} (such as gray-level average preservation, translation, and symmetry invariance), all shared with ROF denoising [2], which are basic but essential requirements for image processing.

First we establish a facilitating formulation of the TV-LSE operator. It makes use of integrals on the smaller space $\mathcal{E}_{\bar{v}}$, on which the prior p_β is proper and compatible with (6).

Lemma 3.1. *Let $\bar{v}$ be the average of v, and let $\mathcal{E}_{\bar{v}}$ be the space of images having average $\bar{v}$ (see (7)). Then (24) can be rewritten as*

$$\forall v \in \mathbb{R}^{\Omega}, \quad S_{\mathrm{LSE}}(v) = \frac{\displaystyle\int_{\mathcal{E}_{\bar{v}}} \exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2}\right) u \, du}{\displaystyle\int_{\mathcal{E}_{\bar{v}}} \exp\left(-\frac{E_{v,\lambda}(u)}{2\sigma^2}\right) du}. \tag{25}$$

Proof. For a given $v \in \mathbb{R}^{\Omega}$, let us make the change of variable $u = \bar{u} + z$ in (24), where $\bar{u} \in \mathbb{R}$ is the mean of u and z is in $\mathcal{E}_{\bar{v}}$. We have

$$S_{\mathrm{LSE}}(v) = \frac{\displaystyle\int_{z \in \mathcal{E}_{\bar{v}}} \int_{\bar{u} \in \mathbb{R}} (\bar{u} + z) e^{-\frac{1}{2\sigma^2}(\|\bar{u}+z-v\|^2 + \lambda TV(z))} \, d\bar{u} \, dz}{\displaystyle\int_{z \in \mathcal{E}_{\bar{v}}} \int_{\bar{u} \in \mathbb{R}} e^{-\frac{1}{2\sigma^2}(\|\bar{u}+z-v\|^2 + \lambda TV(z))} \, d\bar{u} \, dz}.$$

As both z and v have mean $\bar{v}$, the quadratic term $\|\bar{u} + z - v\|^2$ equals $|\Omega|\bar{u}^2 + \|z - v\|^2$. Hence $\hat{u}_{\mathrm{LSE}}$ becomes

$$S_{\mathrm{LSE}}(v) = \frac{\displaystyle\int_{z \in \mathcal{E}_{\bar{v}}} e^{-\frac{1}{2\sigma^2}(\|z-v\|^2 + \lambda TV(z))} \int_{\bar{u} \in \mathbb{R}} (\bar{u} + z) e^{-\frac{|\Omega|\bar{u}^2}{2\sigma^2}} \, d\bar{u} \, dz}{\displaystyle\int_{z \in \mathcal{E}_{\bar{v}}} e^{-\frac{1}{2\sigma^2}(\|z-v\|^2 + \lambda TV(z))} \int_{\bar{u} \in \mathbb{R}} e^{-\frac{|\Omega|\bar{u}^2}{2\sigma^2}} \, d\bar{u} \, dz},$$

which, thanks to the properties of the normal distribution $\mathcal{N}(0, \frac{\sigma^2}{|\Omega|})$, simplifies into the desired expression (25). ■

Proposition 3.2 (average preservation). *For any image u, let $\bar{u} = \frac{1}{|\Omega|} \sum_{\mathbf{x} \in \Omega} u(\mathbf{x})$ denote the mean gray level of u. Then for every $v \in \mathbb{R}^{\Omega}$,*

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

Proof. Thanks to Lemma 3.1, $S_{\mathrm{LSE}}(v)$ is written as a weighted average of images all having mean $\bar{v}$. Hence the result has mean $\bar{v}$. ■

Proposition 3.3 (invariance by composition with a 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 (24) yields

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

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

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

because s is linear. ■

A consequence of Proposition 3.3 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 used in (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: for all $u \in \mathbb{R}^\Omega$, for all $c \in \mathbb{R}$, $S_{\mathrm{LSE}}(u + c) = S_{\mathrm{LSE}}(u) + c$ (this is not a direct consequence of Proposition 3.3, 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 many redundancies and much structure. For example, if v is a constant image, then $S_{\mathrm{LSE}}(v) = v$. Indeed, v is invariant under the translations of vectors $(1, 0)$ 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 gray levels a' and b'.

3.2. Asymptotics. Unlike ROF denoising (which depends on the single parameter λ), TV-LSE denoising depends on two distinct parameters λ and σ. The strict Bayesian point of view (see section 2.1) would rather encourage the focus on the single parameter $\beta = \lambda/(2\sigma^2)$ associated to the TV prior, while σ^2 is set as the actual noise variance. In practice, it is more interesting to relax this point of view and to consider σ as a parameter, because, as we shall see later, in general the best denoising results are not obtained when σ^2 equals the actual noise variance. A second reason is that when an image is corrupted with a noise of variance σ^2, the parameter λ in ROF achieving the best peak signal-to-noise ratio (PSNR) is not proportional to σ^2 (as the identity $\lambda = 2\beta\sigma^2$ would suggest) but behaves roughly like a linear function for large values of σ (in [30] a regression yields the estimate $\lambda_{\mathrm{opt}}(\sigma) \approx 2.92\sigma^2/(1 + 1.03\sigma)$). This is why we propose taking (λ, σ) as the parameters of the TV-LSE model. In section 4.3, the role of these parameters is further discussed and illustrated by numerical experiments.

Theorem 3.4 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 notation, satisfies $S_{\mathrm{LSE}}^{\lambda,\sigma}(v) = \sigma S_{\mathrm{LSE}}^{\lambda/\sigma,1}(\frac{v}{\sigma})$.

Theorem 3.4. *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*

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

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

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

where $\bar{v}\mathbb{1}$ is the constant image equal to the average of v. Moving λ such that $\beta = \lambda/(2\sigma^2)$ is kept constant, we have

$$\begin{aligned} &\text{(v)} \quad \hat{u}_{\mathrm{LSE}}(2\beta\sigma^2,\sigma) \xrightarrow[\sigma\to 0]{} v, \\ &\text{(vi)} \quad \hat{u}_{\mathrm{LSE}}(2\beta\sigma^2,\sigma) \xrightarrow[\sigma\to +\infty]{} \bar{v}\mathbb{1}. \end{aligned}$$

Proof. As $E_{v,\lambda}$ is strongly convex (15), the probability distribution with density $\frac{1}{Z}\exp(-\frac{E_{v,\lambda}}{2\sigma^2})$ (where Z is a normalizing constant depending on σ) weakly converges when $\sigma \to 0$ to the Dirac distribution located at $\hat{u}_{\mathrm{ROF}}(\lambda) = \arg\min_u E_{v,\lambda}(u)$, whose expectation is $\hat{u}_{\mathrm{ROF}}(\lambda)$, which proves (i).

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

$$\hat{u}_{\mathrm{LSE}}(\lambda,\sigma) = v + \frac{\displaystyle\int_{\mathbb{R}^\Omega} \sigma w e^{-\frac{1}{2}(\|w\|^2 + \frac{\lambda}{\sigma}TV(w+\frac{v}{\sigma}))}\,dw}{\displaystyle\int_{\mathbb{R}^\Omega} e^{-\frac{1}{2}(\|w\|^2 + \frac{\lambda}{\sigma}TV(w+\frac{v}{\sigma}))}\,dw} = v + \frac{N}{D}. \tag{26}$$

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

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

Hence N can be split into

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

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 $\sigma \geq 1$,

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

and

$$\|f_\sigma(w)\| \leq \|w\| e^{-\frac{\|w\|^2}{2}} (TV(w) + \alpha\|v\|), \tag{27}$$

where α is the Lipschitz-continuity coefficient of TV. As the right-hand term of (27) 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 $\hat{u}_{\mathrm{LSE}}(\lambda, \sigma)$ toward v and proves (ii).

The proof of (iii) is a simple application of Lebesgue's dominated convergence theorem on both integrals of (24).

For (iv), let us assume that 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 3.2), we simply have to show that $\hat{u}_{\mathrm{LSE}}(\lambda, \sigma)$ converges to 0 when λ goes to ∞. Using Lemma 3.1 and making the change of variable $u = z/\lambda$, we obtain

$$\hat{u}_{\mathrm{LSE}}(\lambda, \sigma) = \frac{1}{\lambda} \frac{\displaystyle\int_{\mathcal{E}_0} ze^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))}\, dz}{\displaystyle\int_{\mathcal{E}_0} e^{-\frac{1}{2\sigma^2}(\|\frac{z}{\lambda}-v\|^2+TV(z))}\, dz}. \tag{28}$$

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

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

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 $z \in \mathcal{E}_0$, $TV(z) \geq C\|z\|_1$ (this can be considered as a discrete version of the Poincaré inequality [1]). Thus thanks to Lebesgue's dominated convergence theorem, each integral in (28) converges to a positive value when $\lambda \to +\infty$, and dividing by λ yields the desired limit $\hat{u}_{\mathrm{LSE}}(\lambda, \sigma) \to 0$.

To prove (v) it is enough to apply Lebesgue's dominated convergence theorem to the integrals that appear in (26).

For (vi), we use Lemma 3.1 and Lebesgue's dominated convergence theorem as in the proof of (iv) above. ■

3.3. TV-LSE as a proximity operator and several consequences. Proximity operators [52, 62] are mappings of a Hilbert space into itself, which 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, nonidentically equal to $+\infty$).

Definition 3.5 (see [52, 62]). *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 nonempty set C ($f = 0$ on C and $f = +\infty$ elsewhere), prox_f simply reduces to the projection on C, and that $\mathrm{prox}_{\frac{\lambda}{2}TV}$ corresponds to the ROF denoising operator.

Note 1 (see [52, 62]). Whenever f is 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, 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 $u = \mathrm{prox}_f(z)$ and $v = \mathrm{prox}_{f^*}(z)$.

Definition 3.6 (see [52, 62]). *The primitive function associated to prox_f is the function $\Phi \in \Gamma_0(\mathbb{R}^n)$ defined by*

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

The function $f = \frac{\lambda}{2\sigma^2}TV$ is an element of $\Gamma_0(\mathbb{R}^n)$ whose domain $\{u \in \mathbb{R}^n \mid f(u) < \infty\}$ has a nonempty interior. In addition, 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 as

$$\forall v \in \mathbb{R}^n, \quad S_{\mathrm{LSE}}(v) = \frac{\int u\, G_\sigma(u-v)\, p(u)\, du}{\int G_\sigma(u-v)\, p(u)\, du}. \tag{29}$$

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

Lemma 3.7. *$S_{\mathrm{LSE}} : \mathbb{R}^\Omega \to \mathbb{R}^\Omega$ is differentiable, and its differential dS_{LSE} is a symmetric positive-definite matrix at every point.*

The proof can be found in Appendix B.

Lemma 3.8. *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 has the closed form*

$$\varphi : v \in \mathbb{R}^n \mapsto \frac{1}{2}\|v\|^2 + \sigma^2 \log(p * G_\sigma)(v). \tag{30}$$

Proof. The function φ defined by (30) is $\mathcal{C}^\infty$ since the convolution of p with a Gaussian kernel is $\mathcal{C}^\infty$. Moreover, we have

$$\nabla\varphi(v) = v + \sigma^2 \nabla_v \log \int G_\sigma(v-u)\, p(u)\, du = v + \sigma^2 \frac{\int \nabla_v G_\sigma(v-u)\, p(u)\, du}{\int G_\sigma(v-u)\, p(u)\, du}, \tag{31}$$

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 (the concavity of the second term $\sigma^2 \log(p * G_\sigma)$ follows from the proof of Theorem 3.9 below). 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 3.7, $dS_{\mathrm{LSE}}(v)$ is positive-definite, which ends the proof. ■

Theorem 3.9. *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 [60]. Hence $\sigma^2 \log(p * G_\sigma)$ is concave, and the function φ (defined in Lemma 3.8) is less convex than $v \mapsto \frac{1}{2}\|v\|^2$ (that is, the mapping $v \mapsto \frac{1}{2}\|v\|^2 - \varphi(v)$ is convex). Then, applying [52, 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 by $g^* \in \Gamma_0(\mathbb{R}^n)$ the Legendre–Fenchel transform of g, we have $\nabla\varphi = \mathrm{prox}_{g^*}$ [52, 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 with which S_{LSE} is associated.

Definition 3.10 (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 $S_{\mathrm{LSE}} = \mathrm{prox}_{\frac{1}{2}TV_\sigma}$.*

The existence of such a function TV_σ is given by Theorem 3.9, while the uniqueness is a consequence of [52, Proposition 8.a]. Thus, and still for $\lambda = 1$, $S_{\mathrm{LSE}}(v)$ 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 notation for $S_{\mathrm{LSE}}^{\lambda,\sigma}$, since we have

$$\forall v \in \mathbb{R}^\Omega, \quad S_{\mathrm{LSE}}^{\lambda,\sigma}(v) = \frac{1}{\lambda} S_{\mathrm{LSE}}^{1,\frac{\sigma}{\lambda}}\left(\frac{1}{\lambda} v\right)$$

and the scaling property of the proximity operators,

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

it follows that

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

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

Corollary 3.11. *S_{LSE} is nonexpansive, that is,*

$$\forall v_1, v_2 \in \mathbb{R}^n, \quad \|S_{\mathrm{LSE}}(v_2) - S_{\mathrm{LSE}}(v_1)\| \leq \|v_2 - v_1\|, \tag{32}$$

and monotone in the sense of Minty [50], that is,

$$\forall v_1, v_2 \in \mathbb{R}^n, \quad \langle S_{\mathrm{LSE}}(v_2) - S_{\mathrm{LSE}}(v_1), v_2 - v_1\rangle \geq \|S_{\mathrm{LSE}}(v_2) - S_{\mathrm{LSE}}(v_1)\|^2. \tag{33}$$

Proof. The nonexpansiveness property is a consequence of [52, Proposition 5.b], and the monotonicity a consequence of [50] or [52, 5.a] (these properties are condensed in [62, p. 340]). ■

3.3.2. No staircasing effect with TV-LSE. We first show that S_{LSE} is a $\mathcal{C}^\infty$-diffeomorphism from $\mathbb{R}^n$ onto itself.

Lemma 3.12. *S_{LSE} is injective.*

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

$$\forall t \in \mathbb{R}, \quad \psi(t) = \langle S_{\mathrm{LSE}}((1-t)v_1 + tv_2), v_2 - v_1\rangle$$

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

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

must vanish at a certain point $t_0 \in [0,1]$. But $dS_{\mathrm{LSE}}((1-t_0)v_1 + t_0 v_2)$ is a positive-definite matrix (see Lemma 3.7), and consequently $\psi'(t) > 0$ unless $v_1 = v_2$. ■

Lemma 3.13. *Let I denote the identity of $\mathbb{R}^n$. The operator $S_{\mathrm{LSE}} - I$ is bounded, and S_{LSE} is onto.*

The proof follows from the Lipschitz-continuity of the discrete TV operator and is detailed in Appendix C.

Theorem 3.14. *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 $S_{\mathrm{LSE}} = \nabla\varphi$ with φ in $\mathcal{C}^\infty$ (see Lemma 3.8). Now, adding the fact that dS_{LSE} is invertible at every point (Lemma 3.7) and that S_{LSE} is injective (Lemma 3.12), we obtain by the global inversion theorem that S_{LSE} is a $\mathcal{C}^\infty$-diffeomorphism from $\mathbb{R}^n$ to $S_{\mathrm{LSE}}(\mathbb{R}^n)$. We conclude the proof by using the fact that $S_{\mathrm{LSE}}(\mathbb{R}^n) = \mathbb{R}^n$ (Lemma 3.13). ■

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, nondestruction of information), but it also allows us to state the main result of this section.

Theorem 3.15 (S_{LSE} induces no staircasing). *If V is a random image whose p.d.f. is absolutely continuous with respect to the Lebesgue measure, then for any distinct pixels $\mathbf{x}$ and $\mathbf{y}$, one has*

$$\mathbb{P}\left\{S_{\mathrm{LSE}}(V)(\mathbf{x}) = S_{\mathrm{LSE}}(V)(\mathbf{y})\right\} = 0. \tag{34}$$

A consequence of this property is that two neighboring pixels (say, for the 4- or the 8-connectedness) have a probability zero of having the same value in $S_{\mathrm{LSE}}(V)$. Thus, almost surely $\hat{u}_{\mathrm{LSE}}$ contains no constant region, which means that there is no staircasing in the sense of [55], 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 TV distribution (that is, $V \sim \frac{1}{Z} e^{-\lambda TV(V)}$), then the assumption on V in Theorem 3.15 is met, and $\hat{u}_{\mathrm{LSE}}$ almost surely contains no staircasing. Note that it does not state that edges should be blurred out. In section 3.3.3 (through a theoretical argument) and section 4 (through denoising experiments), we show that it is indeed not the case.

Note incidentally that (34) implies that any original image in which two pixels share the same gray value cannot be exactly restored. This is not really an issue since "exact restoration" does not make much sense in the numerical world (numerical solutions, and also physical images, are known only up to some precision), and of course such an image can be arbitrarily well approximated using an image with distinct gray values.

Proof of Theorem 3.15. Let p_V be the probability measure associated with 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 the Lebesgue measure, the probability $p_V(A)$ is null. Now

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

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 [69, Théorème 1.1]. In particular, S_{LSE}^{-1} transforms negligible sets into negligible sets [69, Lemma 2.1], and $p_V(S_{\mathrm{LSE}}^{-1}(A)) = 0$. ■

3.3.3. Properties of TV_σ and recovery of edges. In this section, we study the potential TV_σ introduced in Definition 3.10. 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 $p_{\mathrm{LSE}} = \frac{1}{Z}\exp(-\frac{1}{2\sigma^2}TV_\sigma)$, or, equivalently, as the minimizer of a variational formulation including the classical squared L^2 data-fidelity term and the potential TV_σ, as was pointed out in [46, section 3.5] and later in [34] in a more general framework. Here we specifically investigate some properties of TV_σ that are particularly useful in comparing the TV_σ and TV potentials.

Proposition 3.16. *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

$$\frac{1}{2}\nabla TV_\sigma = S_{\mathrm{LSE}}^{-1} - I, \tag{35}$$

and the right-hand term is $\mathcal{C}^\infty$ thanks to Theorem 3.14, 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 [55] 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 3.17. *TV_σ is Lipschitz-continuous, and denoting by $\|\cdot\|_{\mathrm{Lip}}$ the Lipschitz constant of an operator, we have*

$$\forall \sigma > 0, \quad \|TV_\sigma\|_{\mathrm{Lip}} \leq \|TV\|_{\mathrm{Lip}}.$$

Proof. TV_σ is differentiable and S_{LSE} is invertible, so thanks to (35), we get

$$\begin{aligned}\|TV_\sigma\|_{\mathrm{Lip}} &= \sup_u \|\nabla TV_\sigma(u)\| = 2\sup_v \|S_{\mathrm{LSE}}(v) - v\| = 2\sup_v \|\sigma^2 \nabla \log(p * G_\sigma)(v)\| \\ &= 2\sigma^2 \|\log(p * G_\sigma)\|_{\mathrm{Lip}}.\end{aligned}$$

It remains to compute $\|\log(p * G_\sigma)\|_{\mathrm{Lip}}$. But since $p = \frac{1}{Z} e^{-\frac{TV}{2\sigma^2}}$, letting $\kappa = \|TV\|_{\mathrm{Lip}}$, we have for every u, v, and v' in $\mathbb{R}^n$

$$p(v-u) \leq p(v'-u) e^{\frac{\kappa}{2\sigma^2}\|v-v'\|}.$$

Hence, for every v and v',

$$p * G_\sigma(v) = \int p(v-u) G_\sigma(u)\, du \leq \int p(v'-u) e^{\frac{\kappa}{2\sigma^2}\|v-v'\|} G_\sigma(u)\, du = p * G_\sigma(v') e^{\frac{\kappa}{2\sigma^2}\|v-v'\|},$$

which means that $\|\log(p * G_\sigma)\|_{\mathrm{Lip}} \leq \frac{\kappa}{2\sigma^2}$ and that $\|TV_\sigma\|_{\mathrm{Lip}} \leq \kappa$, which concludes the proof. ■

Let us consider a consequence of Corollary 3.17. 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

$$2(\hat{u} - v) + \nabla TV_\sigma(\hat{u}) = 0. \tag{36}$$

Subtracting (36) 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 $\|TV\|_{\mathrm{Lip}}$, we have

$$|\hat{u}(\mathbf{x}) - \hat{u}(\mathbf{y})| \geq |v(\mathbf{x}) - v(\mathbf{y})| - \|TV\|_{\mathrm{Lip}}. \tag{37}$$

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

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

Corollary 3.18. *The potential TV_σ is linked to TV by the equality*

$$\left(I + \frac{1}{2}\nabla TV_\sigma\right)^{-1} = I + \sigma^2 \frac{\nabla(p * G_\sigma)}{p * G_\sigma} \tag{38}$$

or, equivalently, by

$$\frac{1}{2}\nabla TV_\sigma = \left(I + \sigma^2 \nabla \log(e^{-\frac{TV}{2\sigma^2}} * G_\sigma)\right)^{-1} - I. \tag{39}$$

Proof. Rewriting (31) gives

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

Now, because of (35), we can write

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

Grouping these two equations yields (38), and (39) immediately follows from $p = \frac{1}{Z}e^{-\frac{TV}{2\sigma^2}}$. ∎

There is probably no simple closed formula for TV_σ, but (39) is a natural starting point to derive approximations of ∇TV_σ. For instance, it seems that when σ goes to 0, ∇TV_σ converges to ∇TV at each point where TV is differentiable. Obtaining a higher order Taylor expansion of the right-hand side of (39) would be most helpful in getting 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 into a minimization algorithm to efficiently compute approximations of the TV-LSE operator.

Another natural question that arises from the definition of TV_σ is, Would it be interesting to iterate the $TV \mapsto TV_\sigma$ process? Let us make this idea more explicit: call S_0 the ROF denoising operator (associated to the TV-based prior $\boldsymbol{\pi}_0$); then by induction define S_{i+1} (for any integer i) as the conditional mean of $\boldsymbol{\pi}_i$, and assume that it can be interpreted as the mode of a prior $\boldsymbol{\pi}_{i+1}$ (as we did to define TV_σ from TV-LSE). By definition, S_1 is nothing but the TV-LSE operator (S_{LSE}), and the study of the sequence $(S_i)_{i\geq 0}$ could be an interesting problem. Now, since the distribution $\boldsymbol{\pi}_1$ is associated to the smooth functional TV_σ, its mode and its expectation are likely to be very close to each other (for a second-order—hence symmetric—approximation around the mode, they would be equal), so that $S_2 \approx S_1$, which makes us believe that more iterations of the process would probably result in minor alterations of S_1. In a sense, the TV-LSE operator could reconcile Bayesian and frequentist point of views, since the MAP and LSE approaches lead to very similar operators for the prior associated to TV_σ (this is another way of saying that $S_2 \approx S_1$).

4. Experiments.

4.1. An algorithm for TV-LSE. As we saw in (24), the denoised image $\hat{u}_{\mathrm{LSE}}$ can be written as

$$\hat{u}_{\mathrm{LSE}} = \int u\, \boldsymbol{\pi}(du) = \int u\, \pi(u)\, du, \quad \text{where} \quad \pi(u) = \frac{1}{Z} e^{-\frac{1}{2\sigma^2}E_{v,\lambda}(u)} \tag{40}$$

is the density of the posterior distribution $\boldsymbol{\pi}$. Hence, the computation of $\hat{u}_{\mathrm{LSE}}$ 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 give only a quick and intuitive explanation of the algorithm described in [47]. A more complete publication, devoted to the detailed description and study of this algorithm, is currently in preparation; for the time being, the interested reader can find more details in [46].

The principle of the MCMC algorithm is the following: if we were able to draw independent and identically distributed samples from the posterior distribution $\boldsymbol{\pi}$ (40), a good approximation of the posterior mean $\hat{u}_{\mathrm{LSE}}$ could be obtained, thanks to the law of large numbers, by averaging all of 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} depends only 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 [47]).

Let us describe the construction of (U_n) in more detail. 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 it reaches 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 below, which returns an estimate of $S_{\mathrm{LSE}}(u)$. It makes use of the function $E^{\mathbf{x}}_{v,\lambda}(u,t)$, which is defined as follows: denote by $u_{\mathbf{x},t} \in \mathbb{R}^{\Omega}$ 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 $E^{\mathbf{x}}_{v,\lambda}(u,t)$ captures in the formula for $E_{v,\lambda}(u_{\mathbf{x},t})$ (see (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

$$\begin{aligned} \forall (x,y) \in \Omega, \quad E^{(x,y)}_{v,\lambda}(u,t) = {} & (t - v(x,y))^2 \\ & + \lambda\sqrt{(u(x-1,y)-t)^2 + (u(x-1,y) - u(x-1,y+1))^2} \\ & + \lambda\sqrt{(u(x,y-1)-t)^2 + (u(x,y-1) - u(x+1,y-1))^2} \\ & + \lambda\sqrt{(t - u(x+1,y))^2 + (t - u(x,y+1))^2}, \end{aligned} \tag{41}$$

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. Principle of Metropolis–Hastings algorithm to compute $\hat{u}_{\mathrm{LSE}}$.

$n \leftarrow 0$, $S \sim \mu_0$
draw a white noise image U
repeat
  draw $\mathbf{x} \sim \mathcal{U}(\Omega)$ (uniform distribution on Ω)
  $t \leftarrow U(\mathbf{x})$
  draw $t' \sim \mathcal{U}([t-\alpha, t+\alpha])$
  let $U(\mathbf{x}) \leftarrow t'$ with probability $\min\big(1, \exp\big(-\frac{E^{\mathbf{x}}_{v,\lambda}(u,t') - E^{\mathbf{x}}_{v,\lambda}(u,t)}{2\sigma^2}\big)\big)$ (see (41))
  $S \leftarrow S + U$
  $n \leftarrow n + 1$
until convergence criterion is satisfied
return $\frac{1}{n}S$.

In practice, a more elaborate version of Algorithm 1 is used, as described in [46, 47]. The convergence criterion is based on the use of two independent chains, (U_k) and $(\tilde{U}_k)$, and on the fact that due to the large dimension (see section 2.3), the estimation error can be accurately predicted by the distance between the two chains. Indeed, one has (see [46, section 2.3.3] and [47, section 3.2])

$$\left\| \hat{u}_{\mathrm{LSE}} - \frac{S_n + \tilde{S}_n}{2} \right\| \approx \frac{1}{2}\|S_n - \tilde{S}_n\|, \quad \text{where} \quad S_n = \frac{1}{n}\sum_{k=1}^{n} U_k \quad \text{and} \quad \tilde{S}_n = \frac{1}{n}\sum_{k=1}^{n} \tilde{U}_k.$$

Also, a so-called burn-in procedure is used, which speeds up the convergence of the algorithm: instead of averaging all the $(U_k)_{1\le k\le n}$, it is preferable to skip the first b iterations and begin the averaging from iteration $b+1$, once the chain has attained an approximately stationary regime. An elegant procedure, again relying on the high dimensionality of the image space, permits one to optimize the parameter b during the iterations (see [46, section 2.4]).

As for the parameter α of Algorithm 1, it can be automatically set by using a fast preliminary scaling procedure based on the control of the acceptation rate of the $U(\mathbf{x}) \leftarrow t'$ decision in the algorithm. More details, as well as a proof of convergence of the algorithm, can be found in [46] and [47].

Algorithm 1, optimized as described above, is able to compute the TV-LSE denoised version $\tilde{u}$ of a 256×256 image with precision 1 (that is, $\frac{1}{|\Omega|}\sum_{\mathbf{x}\in\Omega}(\tilde{u}(\mathbf{x}) - \hat{u}_{\mathrm{LSE}}(\mathbf{x}))^2 \le 1$) in approximately 1 minute on a single 3 GHz processor ($\sigma = 10$, $\lambda = 40$). Note that this algorithm can be very easily and efficiently parallelized on multicore hardware by running (and averaging) several independent chains.

4.2. Comparison to the ROF model and the staircasing effect. In Figures 2–4, 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 hand, several similarities between the denoised signals or images can be noticed.

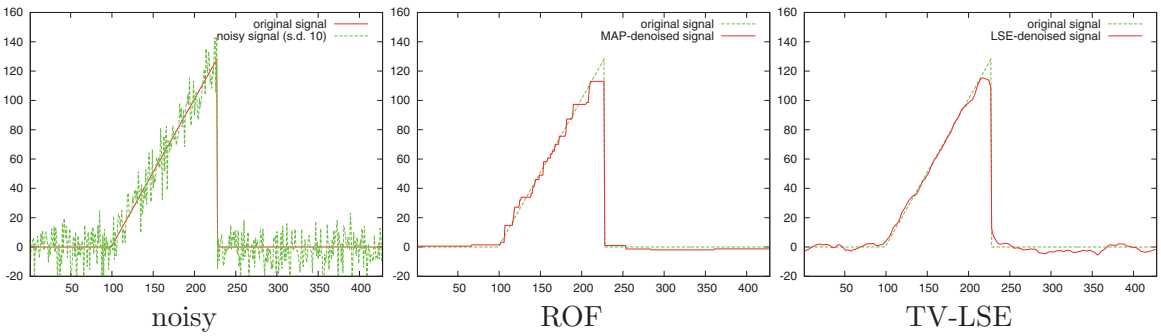

Figure 2. Denoising of a simple synthetic signal. *A triangle-shaped signal (left figure, red curve) is corrupted by an additive white Gaussian noise, and the resulting signal (left, green curve) is then denoised using the ROF (middle) and TV-LSE (right) methods. In the ROF result, the noise has been 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 to equalize the method noise level (L^2-distance from the noisy signal to the result).*

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 hand, the proposed TV-LSE model shows some differences with respect to the ROF model, the most striking of which is the avoidance of the staircasing effect, proved in Theorem 3.15. 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 moves 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 3.12), according to which two versions of the noisy image (two different noise realizations) cannot lead to the same denoised result: there must remain some trace of the initial noise in the denoised image. In Figure 5, we can observe that the histograms of the horizontal derivatives of the ROF denoised images contain a Dirac mass in zero, as was mentioned in section 2.2, while TV-LSE denoised images avoid this artifact, as predicted by Theorem 3.15.

4.3. Role of the hyperparameters. As clearly appears in (24), 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 6 and 7 show how the TV-LSE denoised image changes when λ is tuned while maintaining a fixed value of σ (Figure 6), or when σ is tuned with a fixed value of λ (Figure 7). One can see in Figure 6 that fixing $\sigma > 0$ and letting λ go to 0 makes the image look

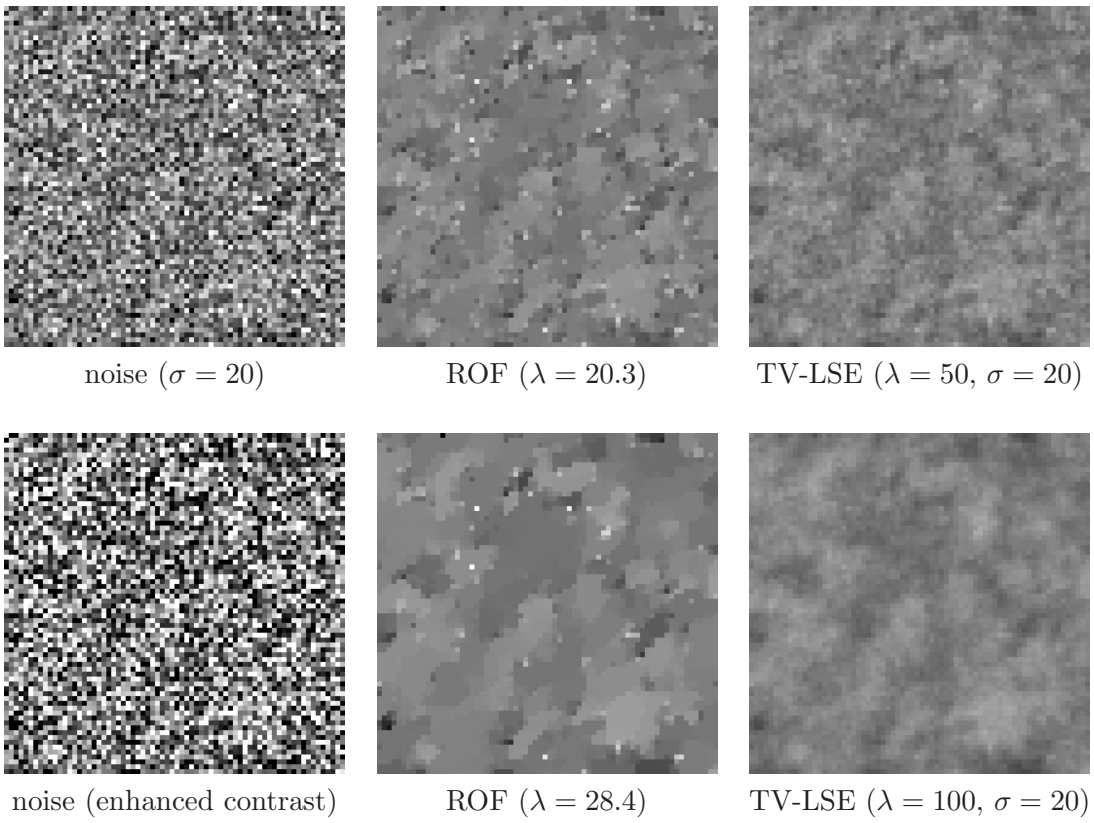

Figure 3. Denoising of a pure noise image. *A constant image is corrupted by a white Gaussian noise with standard deviation $\sigma = 20$ (top left image and bottom left image after an affine contrast change). In columns 2 and 3 we show, respectively, the results of ROF and TV-LSE methods on this image, 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 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 an 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 almost acts 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 behaviors 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.*

like the noisy initial image, and increasing λ makes the image smoother until it becomes a constant. One can also see in Figure 7 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 of these observations agree with the asymptotic results of section 3.2.

The λ parameter is useful since it permits one 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 (6) motivating the introduction of TV-LSE. Thus, considering σ and β as the two hyperparameters of the model allows us to better

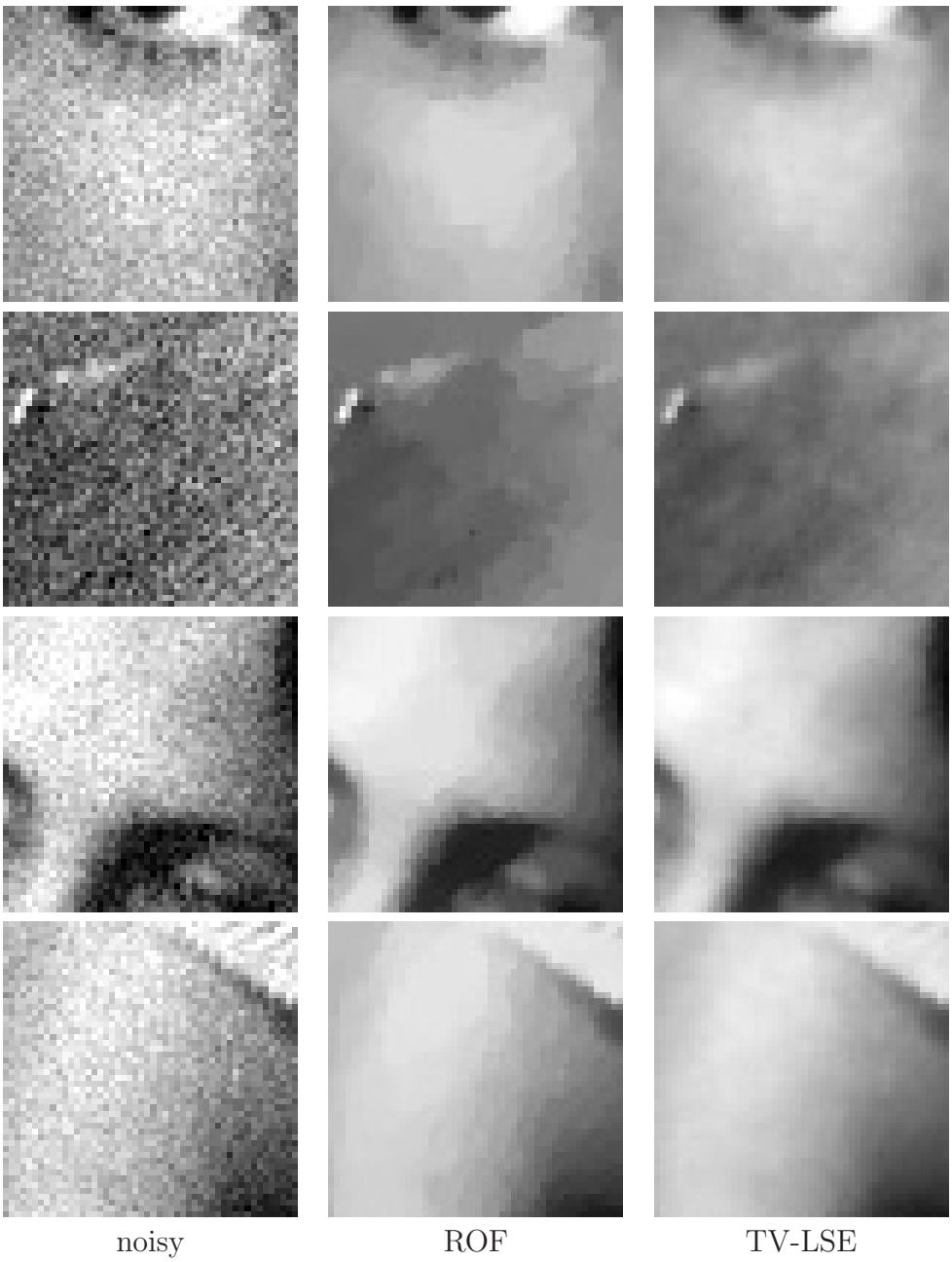

Figure 4. No staircasing effect with TV-LSE. *We experimentally check that the TV-LSE method does not create staircasing artifacts. The left column shows parts of the classical Lena and Barbara images, after they have been corrupted with an additive white Gaussian noise ($\sigma = 10$). The right column 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: $\lambda_{\mathrm{ROF}} = 25.6$, $\lambda_{\mathrm{ROF}} = 20.3$, $\lambda_{\mathrm{ROF}} = 29.0$, $\lambda_{\mathrm{ROF}} = 26.9$). 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.*

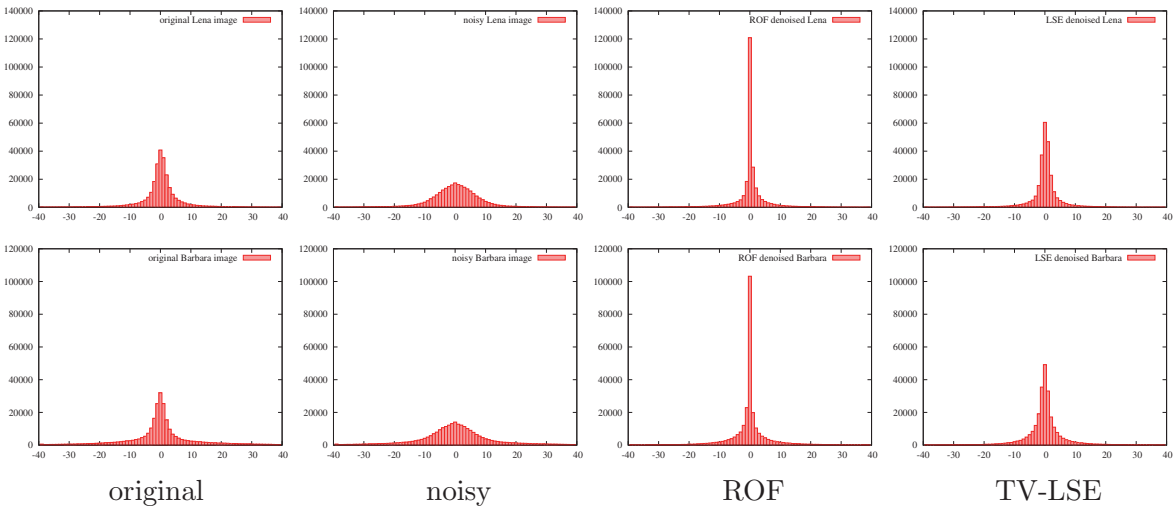

Figure 5. The staircasing effect revealed by gradient histograms. *These plots display the histogram of the discrete horizontal derivative of several versions of the Lena (top row) and Barbara (bottom row) images. The columns correspond, from left to right, to the original (noise-free) image, the noisy version ($\sigma = 10$), and the noisy version denoised by ROF and TV-LSE, respectively, with the same level of denoising (measured by the norm of the estimated noise image). The staircasing effect is responsible for the high central peak of the ROF plot, whereas the TV-LSE plot looks like a generalized Laplace distribution, which would typically be observed on a natural (staircasing-free) image.*

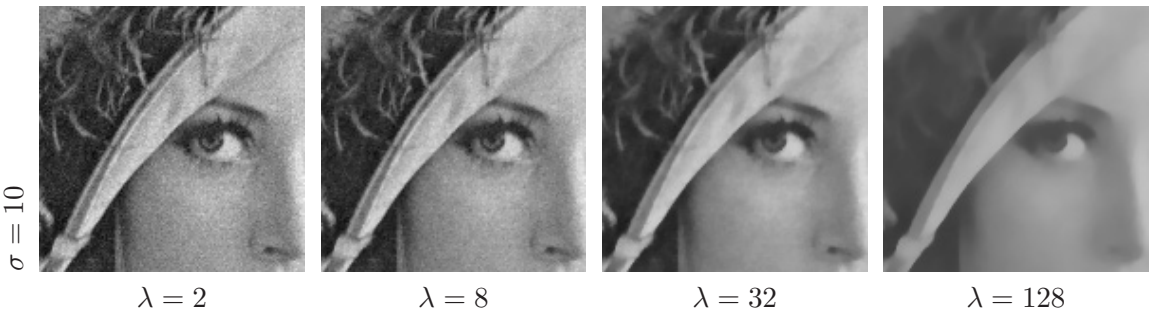

Figure 6. *A noisy image is processed by TV-LSE with $\sigma = 10$ (which corresponds to the standard deviation of the noise) and increasing values of λ. When λ is small (left), the denoised image $\hat{u}_{\mathrm{LSE}}$ is very close to the noisy image v. As λ increases, the noise gradually disappears, 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).*

dissociate the noise and regularization parameters. In Figure 8 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: β really acts as the regularizing parameter. Notice that, inversely, fixing σ and increasing β would be equivalent to the case of Figure 6 (fixed σ and increasing values of λ).

To compare precisely the advantage of TV-LSE over ROF in terms of image denoising (see Figure 9), we fixed the level of denoising, measured by the L^2-norm of the residual image

Figure 7. *A noisy image is processed by TV-LSE with $\lambda = 30$ and increasing values of σ. When σ is small (left), the denoised image $\hat{u}_{\mathrm{LSE}}$ is very close to the ROF-denoised image $\hat{u}_{\mathrm{ROF}}(\lambda)$, 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, $\hat{u}_{\mathrm{LSE}}$ looks more and more like the noisy image, which is consistent with the convergence $\hat{u}_{\mathrm{LSE}}(\sigma, \lambda) \to v$ when $\sigma \to \infty$.*

Figure 8. *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.*

$v - \hat{u}_{\mathrm{LSE}}$ (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 3.15, which states 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 3.15 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 9. Indeed, this limit image would define a filter controlled by a single parameter, the method noise level. In practice, we observe 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 9. *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 10 gives a two-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.

4.4. A SURE criterion to select the hyperparameters σ and λ. Contrary to the ROF model that only involves one parameter (λ), the TV-LSE model depends on two parameters (λ and σ). Hence, considering the relative slowness of Algorithm 1, it may be interesting to have an automatic way of setting these parameters. Of course, σ can be set equal to the noise standard deviation, which is theoretically sound in the Bayesian framework of section 2.1, but as our theoretical convergence results prove (section 3.2), having σ going to 0 makes TV-LSE converge to ROF denoising, so that the tuning of σ comes into question. In the end, as we discussed in section 3.2, it seems more interesting to consider σ as a tunable parameter, on the same level as λ.

If one wants to tune (σ, λ) so as to minimize the L^2-distance between the TV-LSE denoised image and the original noise-free image, knowledge of the latter is required, unless an unbiased risk estimator can be used. It turns out that Stein's unbiased risk estimator (SURE) [70] is easily computable for TV-LSE. The SURE for a denoising operator S and a noisy image $v \in \mathbb{R}^{\Omega}$ is

$$\mathrm{SURE}(S)(v) = \|S(v) - v\|^2 + 2\sigma_0^2 \sum_{\mathbf{x}\in\Omega} \frac{\partial S(v)(\mathbf{x})}{\partial v(\mathbf{x})} - \sigma_0^2|\Omega|, \tag{42}$$

where σ_0 is the standard deviation of the noise. The interesting property of SURE is that its expectation among all realizations of noise is the same as that of $\mathrm{MSE}(v) := \|S(v) - u\|^2$, where u is the (unknown) noise-free image (see [73] for a short proof and further references). In practice, SURE and MSE have the same order of magnitude for a given noise, and it is enough to set the parameters so as to minimize SURE to obtain a near-to-optimal PSNR.

Proposition 4.1. *The SURE for the operator S_{LSE} with parameters λ and σ is*

$$\mathrm{SURE}(S_{\mathrm{LSE}}^{\lambda,\sigma})(v) = \|\mathbb{E}_{\boldsymbol{\pi}}(U) - v\|^2 + \frac{2\sigma_0^2}{\sigma^2}\mathbb{E}_{\boldsymbol{\pi}}\|U - \mathbb{E}_{\boldsymbol{\pi}}(U)\|^2 - \sigma_0^2|\Omega|, \tag{43}$$

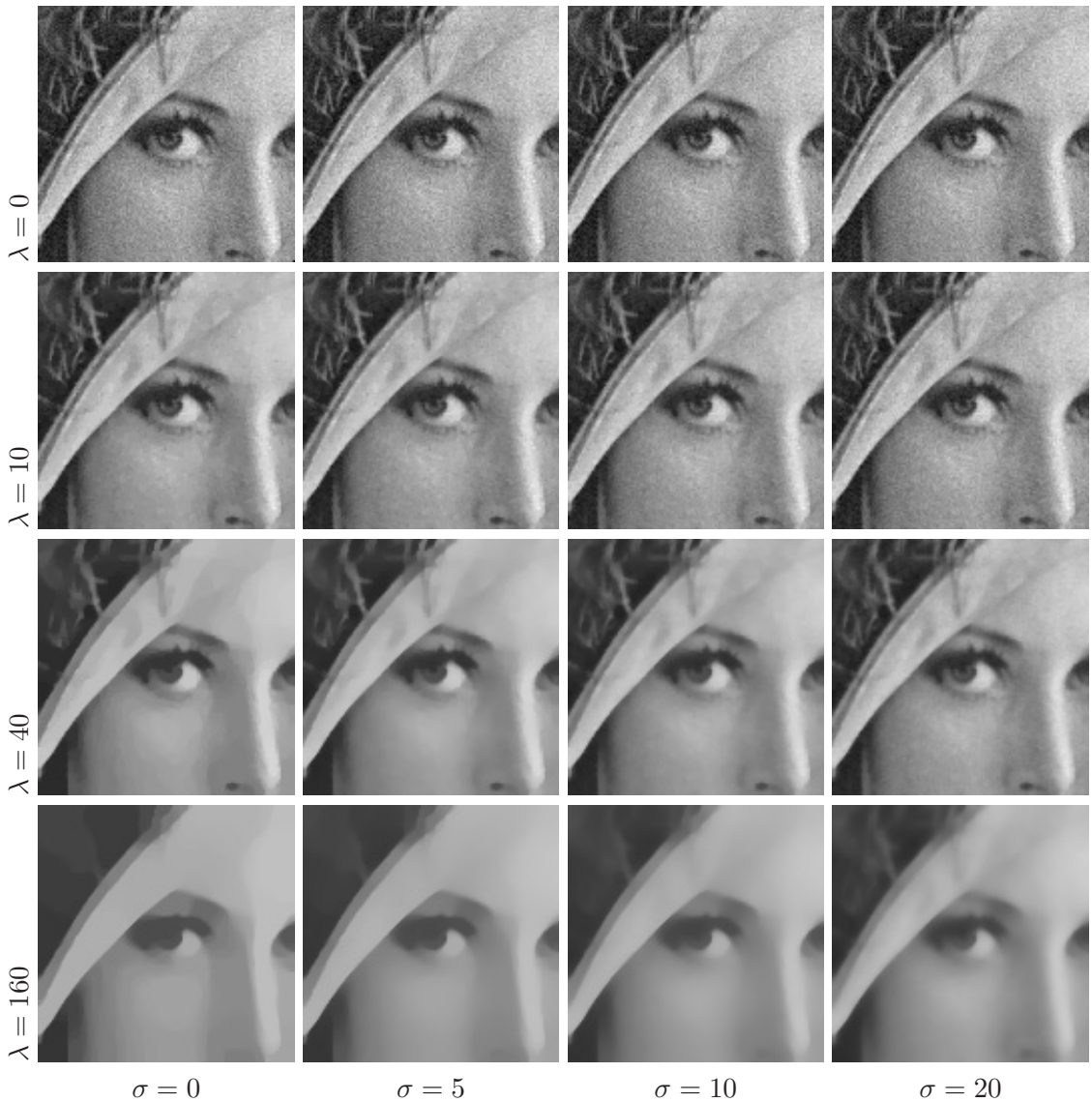

Figure 10. Effect of the two parameters λ and σ on TV-LSE. *A noisy version of the Lena image (Gaussian white noise with standard deviation equal to* 10*) is processed with TV-LSE for various values of* λ *and* σ*. First row:* $\lambda = 0$ *(the TV-LSE image is equal to the noisy image); second row:*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

where $\boldsymbol{\pi}$ *is the posterior distribution (depending on* λ *and* σ*) and* σ_0 *is the standard deviation of the noise.*

The proof of Proposition 4.1 is postponed to Appendix D.

Thanks to the usual property

$$\mathbb{E}_{\boldsymbol{\pi}}\|U - \mathbb{E}_{\boldsymbol{\pi}}(U)\|^2 = \mathbb{E}_{\boldsymbol{\pi}}\|U\|^2 - \|\mathbb{E}_{\boldsymbol{\pi}}(U)\|^2,$$

SURE($S_{\mathrm{LSE}}^{\lambda,\sigma}$) can be computed within Algorithm 1 for a pair (σ, λ) with no extra loop: it suffices to keep in memory the sum of the squares of the chains (the convergence is only a little slower for SURE than for S_{LSE}). In Figure 11, the value of the SURE criterion is plotted for different values of σ and λ, and a comparison with the oracle MSE $= \|S_{\mathrm{LSE}}(v) - u\|^2$ is proposed. It is apparent first that minimizing SURE is an efficient tool for minimizing MSE as they attain their minimum for similar pairs (σ, λ). Note that the minimum of MSE (and SURE) is reached neither for $\sigma = 0$ (which would have demonstrated the superiority of ROF over TV-LSE) nor for $\sigma = \sigma_0$, the true standard deviation of the noise (which is in favor of considering σ as a full parameter, as we do). Note that the SURE criterion for ROF cannot be directly derived from (43) because the associated distribution $\boldsymbol{\pi}$ has variance 0, but another way to compute the SURE criterion of the ROF model can be found in [72].

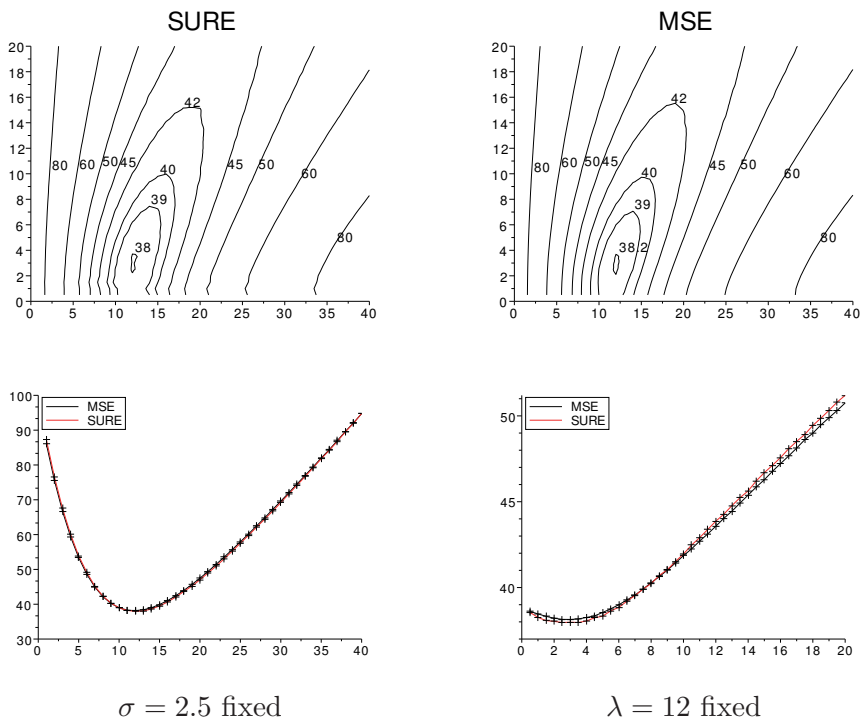

Figure 11. Comparison of SURE and MSE. *Top left: The SURE criterion for TV-LSE is computed for a noisy subpart of the Lena image, in function of λ (x-axis) and σ (y-axis), and some of its level lines are displayed. The similarity to the level lines of the MSE oracle (top right) is remarkable. This shows that the optimal values (in the PSNR sense) of the parameters λ and σ can be efficiently approximated by looking for the values that minimize the SURE criterion. In the bottom row, we can observe slices corresponding, respectively, to $\sigma = 2.5$ and a variable λ (left), and to $\lambda = 12$ and a variable σ (right), which are also very similar.*

4.5. Comparison to other TV-based denoising methods. In this section, we propose comparing 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 TV 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 try to blindly rank the different methods using classical metrics like the PSNR or the structural similarity index which are poor predictors of the visual quality of the results. Indeed, Table 1 show that except for the TV-L^1 method (which performs significantly worse), all considered methods achieve similar PSNR levels.

Table 1

Table of PSNR values obtained for the denoising of the Lena image for different values of σ_0, the standard deviation of the noise. For each method, the optimal parameters are used.

	Noisy	ROF	TV-bary	TV-L^1	TV-LSE	TV-Huber	TV-ε	Local-TV
$\sigma_0 = 10$	28.13	34.21	34.22	32.90	34.26	34.25	34.25	34.21
$\sigma_0 = 20$	22.10	31.05	31.07	30.04	31.10	31.09	31.09	31.05

Let us now focus on the visual comparison of the different TV-based denoising methods being considered. Given a noisy image v, we propose comparing $\hat{u}_{\mathrm{LSE}}(\sigma, \lambda)$, the result of TV-LSE applied to v with parameters σ and λ, to the following:

- ROF denoising, alias TV-MAP: the denoised image is denoted by $\hat{u}_{\mathrm{ROF}}(\lambda_{\mathrm{ROF}})$. The parameter λ_{ROF} is tuned in such a way that the denoising level $\|v - \hat{u}_{\mathrm{ROF}}(\lambda_{\mathrm{ROF}})\|$ 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 $\hat{u}_{\mathrm{ROF}}(\lambda)$ with the same value of λ (that is, for which both methods deal with the same energy $E_{v,\lambda}$), we propose combining $\hat{u}_{\mathrm{ROF}}(\lambda)$ linearly with the noisy image v via

$$\hat{u}_{\mathrm{bary}} = t\,\hat{u}_{\mathrm{ROF}}(\lambda) + (1-t)\,v \quad \text{with } t = \frac{\|v - \hat{u}_{\mathrm{ROF}}(\lambda)\|}{\|v - \hat{u}_{\mathrm{LSE}}(\sigma, \lambda)\|}.$$

  We obtain a barycenter of $\hat{u}_{\mathrm{ROF}}(\lambda)$ and v which has the desired denoising level. The choice of this method is also justified by the observation that the quality of denoising often increases both visually and in PSNR when deviating the ROF estimate toward 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 TV and embedding it in the usual variational framework leads to a staircasing-free denoising model [55]. More precisely, we can define generalizations of TV on $\mathbb{R}^\Omega$ by

$$TV_f(u) = \sum_{x \in \Omega} f(|Du|) \tag{44}$$

  for specific smooth functions $f : \mathbb{R} \to \mathbb{R}$ that approximate the absolute value function, and then denoise an image v by minimizing

$$E_f(u) = \|u - v\|^2 + \lambda\, TV_f(u). \tag{45}$$

  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 referred to here as 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 (44) and (45), discussed in [74], for instance, is the so-called Huber norm

$$f_\alpha : x \mapsto \begin{cases} \frac{1}{2\alpha}x^2 & \text{if } |x| \leq \alpha, \\ |x| - \frac{\alpha}{2} & \text{if } |x| > \alpha. \end{cases}$$

  This leads to a denoising model referred to here as 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} [19]. Like ε in TV-ε denoising, α corresponds to a minimal gradient for discontinuity and is set to 10. The regularization parameter $\lambda = \lambda_{\text{Huber}}$ 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 [26, 54].
- Local-TV: it has been proved in [48] 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 [48]. 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.

Figures 12–14 zoom in on different parts of the Lena image processed with all of 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

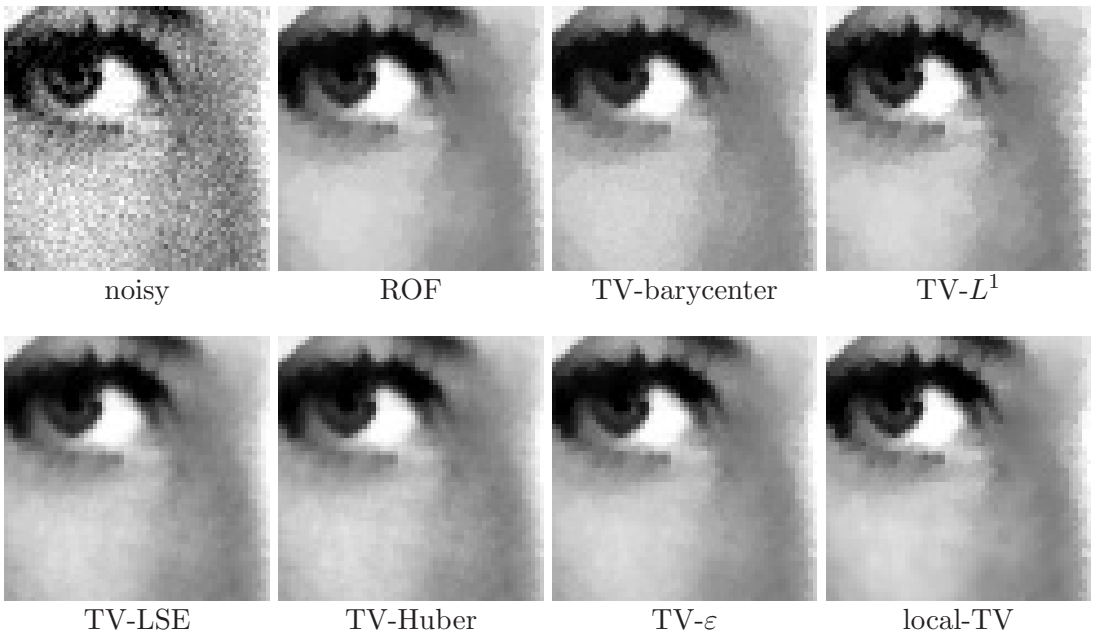

Figure 12. Comparison of TV-LSE denoising to other TV-based denoising methods. *The Lena image is corrupted with an additive Gaussian noise with standard deviation equal to* 10*, and the resulting noisy image (detail on top, left) is first processed with TV-LSE using the parameters* $(\sigma, \lambda) = (10, 30)$*, 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* $\lambda_{\mathrm{ROF}} = 17.03$ *for ROF,* $t = 0.87$ *for TV-barycenter,* $\lambda_{L^1} = 0.80$ *for TV-*L^1*,* $\lambda_{Huber} = 28.78$ *for TV-Huber,* $\lambda_\varepsilon = 23.19$ *for TV-*ε*, and* $\lambda_{loc} = 15.54$ *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 Figures* 13 *and* 14*.*

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: no staircasing, a faithful reconstruction of contrasted edges, and good overall quality. TV-ε also avoids staircasing and is able to reconstruct edges, but it 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 3.10), which seems, according to (39), 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

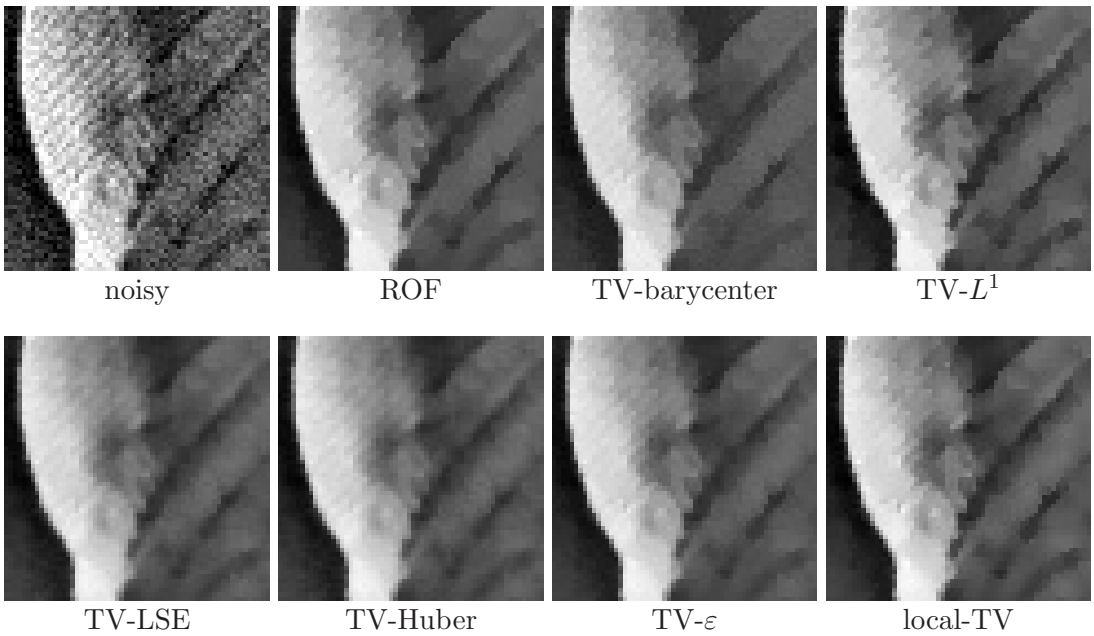

Figure 13. *A second detail of Lena, denoised with various TV-based methods, as in Figure 12. The conclusions are similar: notice in particular how the stripes (top left corner of each subimage) are better restored with the TV-LSE and TV-Huber methods.*

to determine, among all regularized versions 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 [57]. As mentioned in the introduction, the posterior expectation often goes along with a better preservation of local statistics: this is somehow the case for the gradient norm of the denoised images, which, in the TV-LSE variant, avoids the strong peak in 0 that is observed with the ROF model.

These theoretical properties have a direct consequence in the visual quality of the denoised images, which show a nice combination of sharp edges (the most interesting property of the TV functional) and the absence of staircase (piecewise constant) regions. In this sense, the TV-LSE model favorably compares to other TV-based denoising methods, as was shown in section 4. Note that the relative amount of staircasing was evaluated only visually (see Figures 12–14 in particular) and in terms of gradient histograms (Figure 5), but it would be very interesting to derive a specific metric dedicated to staircasing evaluation, in order to obtain quantitative results. Numerical experiments also revealed that the results of the TV-LSE model can be, for

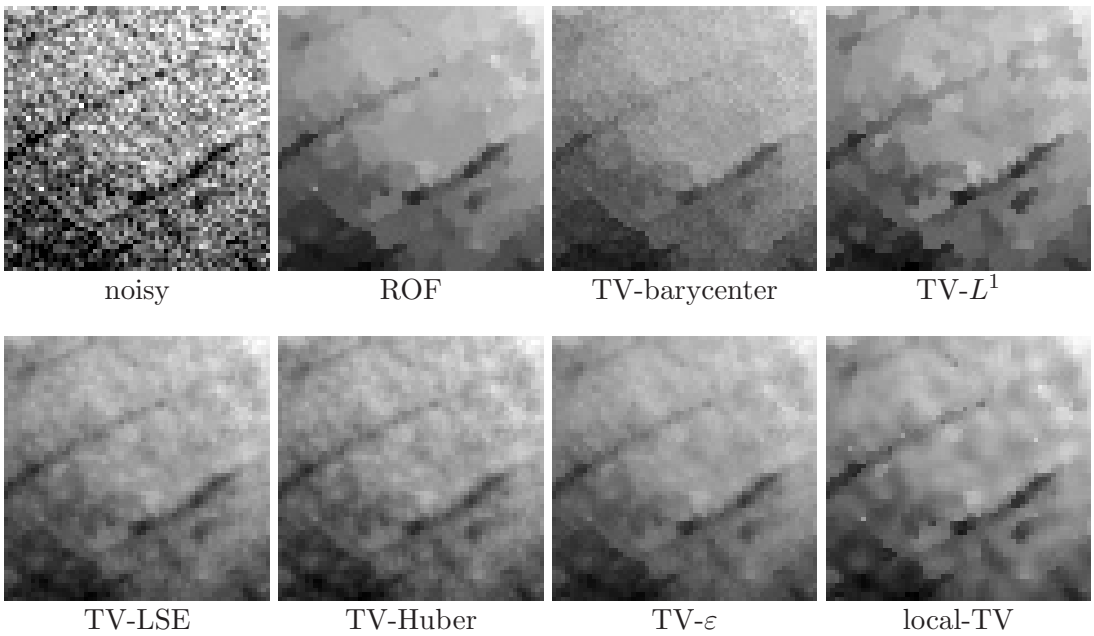

Figure 14. *A third detail of Lena, denoised with various TV-based methods, as in Figure 12.*

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, we believe that the theoretical and numerical framework introduced here opens interesting perspectives, not only for other restoration tasks such as deblurring, zooming, and inpainting, but also because a very similar algorithm could be used to compute the LSE variant associated with other (nonnecessarily convex) functionals (even the nonlocal TV denoising [32] could be reformulated in a TV-LSE setting) or to explore other statistics (median, maximum of marginal distribution, etc.) of the posterior distribution. The appealing case of concave priors seems particularly interesting, but even though initial experiments tend to show that the algorithm used in the present paper still works in some cases, the mathematical framework should be widely adapted.

Appendix A. Mild assumptions for the 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

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

However, the only requirements we really need in the results of section 3 are the following (which 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 nonnegative, convex, and Lipschitz continuous (so that its domain $\{u \in \mathbb{R}^\Omega, TV(u) < +\infty\}$ has a nonempty 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 the Poincaré inequality; i.e., there exists $C > 0$ such that

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

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 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 seminorm $\|\cdot\|_{\dot{B}^s_{p,q}}$, fits the assumptions.

Appendix B. Proof of Lemma 3.7.

Lemma B.1. *Let $P \in \mathbb{R}[X_1, \ldots, X_n]$ be a polynomial. Let p be a bounded p.d.f. Let $F_P : \mathbb{R}^n \to \mathbb{R}$ be such that*

$$F_P : v \mapsto \int_{\mathbb{R}^n} P(u_1, \ldots, u_n)\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u)\, du. \tag{46}$$

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

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

Proof. In this proof, when $u \in \mathbb{R}^n$, we shall write $P(u)$ for $P(u_1, \ldots, u_n)$ 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

$$g : (u, v) \mapsto P(u)\, e^{-\frac{\|u-v\|^2}{2\sigma^2}}\, p(u). \tag{47}$$

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

$$|t| < \varepsilon \quad \Rightarrow \quad \|u - v - th\|^2 \geq \frac{1}{2}\|u - v\|^2 - \varepsilon^2. \tag{48}$$

Let $v \in \mathbb{R}^n$ and $\varepsilon > 0$. Let us denote by $B(v, \varepsilon)$ the set of v' satisfying $\|v' - v\| \leq \varepsilon$. The mapping $g(u, \cdot)$ has an upper bound on $B(v, \varepsilon)$, thanks to (48) given by

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

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 let $\varepsilon > 0$. 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 (48),

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

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 orders 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 is written as

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

which is the desired result. ■

Proof of Lemma 3.7. S_{LSE} is the division of two functions of type F_P (46), with $P = X$ for the numerator and $P = 1$ for the denominator (leading to a positive value). Thanks to Lemma B.1, F_P is continuous and differentiable in both cases, and finally S_{LSE} benefits from this regularity, too.

Again, thanks to Lemma B.1,

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

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 p.d.f. $q_v(u) = \frac{1}{Z} e^{-\frac{\|u-v\|^2}{2\sigma^2}} p(u)$. 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 nonnegative 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 a nonempty 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. ■

Appendix C. Proof of Lemma 3.13. 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, proving 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

$$\lim_{\|v\| \to \infty} \frac{|\langle S_{\mathrm{LSE}}(v), v\rangle|}{\|v\|} = +\infty, \tag{49}$$

so we can apply [12, Corollary 16]: since S_{LSE} is continuous and satisfies (49) 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 (33)), we conclude that S_{LSE} is onto. ■

Appendix D. Proof of Proposition 4.1. Recalling that $S_{\mathrm{LSE}}(v) = \mathbb{E}_{\boldsymbol{\pi}}(U)$, it is sufficient to prove the equality of the middle terms of (42) and (43). Thanks to (31), $S_{\mathrm{LSE}}(v)$ can be written as

$$S_{\mathrm{LSE}}(v) = v + \sigma^2 \nabla \log(p * G_\sigma)(v),$$

with $p = \frac{1}{Z} e^{-\frac{\lambda}{2\sigma^2} TV}$, so that

$$\sum_{\mathbf{x} \in \Omega} \frac{\partial S_{\mathrm{LSE}}(v)}{\partial v(\mathbf{x})}(\mathbf{x}) = |\Omega| + \sigma^2 \Delta \log(p * G_\sigma)(v). \tag{50}$$

Now

$$\begin{aligned}\Delta \log(p * G_\sigma)(v) &= \sum_{\mathbf{x} \in \Omega} \left(\frac{\partial^2}{\partial v(\mathbf{x})^2} \log(p * G_\sigma) \right)(v) \\ &= \sum_{\mathbf{x} \in \Omega} \left[\frac{p * \frac{\partial^2 G_\sigma}{\partial v(\mathbf{x})^2}(v)}{p * G_\sigma(v)} - \left(\frac{p * \frac{\partial G_\sigma}{\partial v(\mathbf{x})}(v)}{p * G_\sigma(v)} \right)^2 \right],\end{aligned}$$

with

$$\frac{\partial G_\sigma}{\partial v(\mathbf{x})}(v) = -\frac{v(\mathbf{x})}{\sigma^2} G_\sigma(v) \quad \text{and} \quad \frac{\partial^2 G_\sigma}{\partial v(\mathbf{x})^2}(v) = \frac{v(\mathbf{x})^2 - \sigma^2}{\sigma^4} G_\sigma(v),$$

which we rewrite using the projection function $\delta_{\mathbf{x}} : v \mapsto v(\mathbf{x})$ as

$$\frac{\partial G_\sigma}{\partial v(\mathbf{x})} = -\frac{1}{\sigma^2} \delta_{\mathbf{x}} G_\sigma \quad \text{and} \quad \frac{\partial^2 G_\sigma}{\partial v(\mathbf{x})^2} = \frac{1}{\sigma^4} (\delta_{\mathbf{x}}^2 - \sigma^2) G_\sigma,$$

so that

$$\Delta \log(p * G_\sigma)(v) = \frac{1}{\sigma^4} \sum_{\mathbf{x} \in \Omega} \left[\frac{p * (\delta_{\mathbf{x}}^2 - \sigma^2) G_\sigma(v)}{p * G_\sigma(v)} - \left(\frac{p * \delta_{\mathbf{x}} G_\sigma(v)}{p * G_\sigma(v)} \right)^2 \right].$$

Now we have

$$\begin{aligned}\frac{p * \delta_{\mathbf{x}} G_\sigma(v)}{p * G_\sigma(v)} &= \frac{\int p(u)(v(\mathbf{x}) - u(\mathbf{x})) G_\sigma(u - v)\, du}{\int p(u) G_\sigma(u - v)\, du} = v(\mathbf{x}) - \frac{\int p(u) u(\mathbf{x}) G_\sigma(u - v)\, du}{\int p(u) G_\sigma(u - v)\, du} \\ &= v(\mathbf{x}) - S_{\mathrm{LSE}}(v)(\mathbf{x}) = v(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})]\end{aligned}$$

and

$$\begin{aligned}\frac{p * (\delta_{\mathbf{x}}^2 - \sigma^2) G_\sigma(v)}{p * G_\sigma(v)} &= \frac{\int p(u)(v(\mathbf{x})^2 - 2v(\mathbf{x})u(\mathbf{x}) + u(\mathbf{x})^2 - \sigma^2) G_\sigma(u - v)\, du}{\int p(u) G_\sigma(u - v)\, du} \\ &= v(\mathbf{x})^2 - 2v(\mathbf{x}) \mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})] + \mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})^2] - \sigma^2.\end{aligned}$$

Consequently,

$$\Delta \log(p * G_\sigma)(v) = \frac{1}{\sigma^4} \sum_{\mathbf{x}\in\Omega} \left(v(\mathbf{x})^2 - 2v(\mathbf{x})\mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})] + \mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})^2] - \sigma^2 - (v(\mathbf{x}) - \mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})])^2 \right),$$

which can be simplified to

$$\Delta \log(p * G_\sigma)(v) = \frac{1}{\sigma^4} \sum_{\mathbf{x}\in\Omega} \left(\mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})^2] - (\mathbb{E}_{\boldsymbol{\pi}}[U(\mathbf{x})])^2 - \sigma^2 \right).$$

Combining this with (50), we obtain (43) from (42), and this concludes the proof. ■

Acknowledgment. We thank the anonymous referees for their useful comments which allowed a significant improvement of the paper.

REFERENCES

[1] L. Ambrosio, N. Fusco, and D. Pallara, *Functions of Bounded Variation and Free Discontinuity Problems*, The Clarendon Press, Oxford University Press, New York, 2000.

[2] G. Aubert and P. Kornprobst, *Mathematical Problems in Image Processing: Partial Differential Equations and the Calculus of Variations*, 2nd ed., Appl. Math. Sci. 147, Springer-Verlag, New York, 2006.

[3] G. Aubert and L. Vese, *A variational method in image recovery*, SIAM J. Numer. Anal., 34 (1997), pp. 1948–1979.

[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 (2005), pp. 71–88.

[5] J.-F. Aujol, G. Gilboa, T. Chan, and S. Osher, *Structure-texture image decomposition—Modeling, algorithms, and parameter selection*, Int. J. Comput. Vision, 67 (2006), pp. 111–136.

[6] M. Bergounioux and L. Piffet, *A second-order model for image denoising*, Set-Valued Var. Anal., 18 (2010), pp. 277–306.

[7] J. Besag, *Digital image processing: Towards Bayesian image analysis*, J. Appl. Stat., 16 (1989), pp. 395–407.

[8] G. Blanchet and L. Moisan, *An explicit sharpness index related to global phase coherence*, in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), 2012, pp. 1065–1068.

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

[10] K. Bredies, K. Kunisch, and T. Pock, *Total generalized variation*, SIAM J. Imaging Sci., 3 (2010), pp. 492–526.

[11] K. Bredies, K. Kunisch, and T. Valkonen, *Properties of L^1-TGV^2: The one-dimensional case*, J. Math. Anal. Appl., 398 (2013), pp. 438–454.

[12] H. R. Brezis, *Les opérateurs monotones*, Séminaire Choquet: 1965166, Initiation à l'Analyse, Fasc. 2, Exp. 10, Secrétariat mathématique, Paris, 1968.

[13] A. Buades, B. Coll, and J.-M. Morel, *The staircasing effect in neighborhood filters and its solution*, IEEE Trans. Image Process., 15 (2006), pp. 1499–1505.

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

[15] V. Caselles, A. Chambolle, and M. Novaga, *Total variation in imaging*, in Handbook of Mathematical Methods in Imaging, Springer, New York, 2011, pp. 1016–1057.

[16] L. Chaari, *Parallel Magnetic Resonance Imaging Reconstruction Problems Using Wavelet Representations*, Ph.D. thesis, Université Paris Est, Paris, France, 2010.

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

[18] 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, Berlin, 2010, pp. 263–340.

[19] A. Chambolle and T. Pock, *A first-order primal-dual algorithm for convex problems with applications to imaging*, J. Math. Imaging Vision, 40 (2011), pp. 120–145.

[20] T. F. Chan, S. Esedoḡlu, and M. Nikolova, *Algorithms for finding global minimizers of image segmentation and denoising models*, SIAM J. Appl. Math., 66 (2006), pp. 1632–1648.

[21] T. F. Chan, S. Esedoglu, and F. E. Park, *Image decomposition combining staircase reduction and texture extraction*, J. Vis. Commun. Image Represent., 18 (2007), pp. 464–486.

[22] 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 Proceedings of the 17th IEEE International Conference on Image Processing, 2010, pp. 4137–4140.

[23] T. F. Chan and J. Shen, *Mathematical models for local nontexture inpaintings*, SIAM J. Appl. Math., 62 (2002), pp. 1019–1043.

[24] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian, *Image denoising with block-matching and 3D filtering*, in Proc. SPIE 6064, Image Processing: Algorithms and Systems, Neural Networks, and Machine Learning, SPIE, Bellingham, WA, 606414.

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

[26] V. Duval, J.-F. Aujol, and Y. Gousseau, *The TVL1 model: A geometric point of view*, Multiscale Model. Simul., 8 (2009), pp. 154–189.

[27] C. Fox and G. K. Nicholls, *Exact MAP states and expectations from perfect sampling: Greig, Porteous and Seheult revisited*, in Bayesian Inference and Maximum Entropy Methods in Science and Engineering: 20th International Workshop, AIP Conf. Proc. 568, American Institute of Physics, Melville, NY, 2000, pp. 252–263.

[28] D. Geman, *Random fields and inverse problems in imaging*, in École d'Été de Probabilités de Saint-Flour XVIII—1988, Lecture Notes in Math. 1427, Springer-Verlag, Berlin, 1990, pp. 113–193.

[29] 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, Morgan Kaufmann, Los Altos, CA, 1987, pp. 564–584.

[30] P. Getreuer, *Rudin-Osher-Fatemi Total Variation Denoising Using Split Bregman*, Image Processing On Line, 2012, http://dx.doi.org/10.5201/ipol.2012.g-tvd.

[31] C. J. Geyer, *Practical Markov chain Monte Carlo*, Statist. Sci., 7 (1992), pp. 473–483.

[32] G. Gilboa, J. Darbon, S. Osher, and T. Chan, *Nonlocal Convex Functionals for Image Regularization*, UCLA CAM Report 06-57, UCLA, Los Angeles, CA, 2006.

[33] M. Grasmair, *The equivalence of the taut string algorithm and BV-regularization*, J. Math. Imaging Vision, 27 (2007), pp. 59–66.

[34] R. Gribonval, *Should penalized least squares regression be interpreted as maximum a posteriori estimation?*, IEEE Trans. Signal Process., 59 (2011), pp. 2405–2410.

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

[36] G. Huang and D. Mumford, *Statistics of natural images and models*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1999, pp. 541–547.

[37] K. Jalalzai, *Regularization of Inverse Problems in Image Processing*, Ph.D. thesis, École Polytechnique, Palaiseau, France, 2012.

[38] K. Jalalzai and A. Chambolle, *Enhancement of blurred and noisy images based on an original variant of the total variation*, in Scale Space and Variational Methods in Computer Vision, Lecture Notes in Comput. Sci. 5567, Springer, Berlin, Heidelberg, 2009, pp. 368–376.

[39] S. Kindermann, S. Osher, and P. W. Jones, *Deblurring and denoising of images by nonlocal functionals*, Multiscale Model. Simul., 4 (2005), pp. 1091–1115.

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

[41] M. Lassas, E. Saksman, and S. Siltanen, *Discretization-invariant Bayesian inversion and Besov space priors*, Inverse Probl. Imaging, 3 (2009), pp. 87–122.

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

[43] M. Ledoux, *Measure Concentration, Transportation Cost, and Functional Inequalities*, Lectures presented at the Instructional Conference on Combinatorial Aspects of Mathematical Analysis, Edinburgh, Scotland, 2002, and the Summer School on Singular Phenomena and Scaling in Mathematical Models, Bonn, Germany, 2003.

[44] M. Ledoux, *The Concentration of Measure Phenomenon*, Math. Surveys Monogr., American Mathematical Society, Providence, RI, 2005.

[45] S. Lefkimmiatis, A. Bourquard, and M. Unser, *Hessian-based norm regularization for image restoration with biomedical applications*, IEEE Trans. Image Process., 21 (2012), pp. 983–995.

[46] C. Louchet, *Variational and Bayesian Models for Image Denoising: From Total Variation towards Non-Local Means*, Ph.D. thesis, Université Paris Descartes, Paris, France, 2008.

[47] C. Louchet and L. Moisan, *Total variation denoising using posterior expectation*, in Proceedings of the 16th European Signal Processing Conference (EUSIPCO 2008), EURASIP, 2008.

[48] C. Louchet and L. Moisan, *Total variation as a local filter*, SIAM J. Imaging Sci., 4 (2011), pp. 651–694.

[49] B. Luo, J.-F. Aujol, and Y. Gousseau, *Local scale measure from the topographic map and application to remote sensing images*, Multiscale Model. Simul., 8 (2009), pp. 1–29.

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

[51] J.-M. Mirebeau and A. Cohen, *Anisotropic smoothness classes: From finite element approximation to image models*, J. Math. Imaging Vision, 38 (2010), pp. 52–69.

[52] J.-J. Moreau, *Proximité et dualité dans un espace hilbertien*, Bull. Soc. Math. France, 93 (1965), pp. 273–299.

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

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

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

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

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

[58] G. Papandreou and A. L. Yuille, *Perturb-and-MAP random fields: Using discrete optimization to learn and sample from energy models*, in Proceedings of the 2011 International Conference on Computer Vision (ICCV), IEEE Computer Society, Washington, DC, 2011, pp. 193–200.

[59] P. Pletscher, S. Nowozin, P. Kohli, and C. Rother, *Putting MAP back on the map*, in Proceedings of the 33rd Annual Symposium of the German Association for Pattern Recognition (DAGM), Springer-Verlag, Berlin, Heidelberg, 2011, pp. 111–121.

[60] A. Prékopa, *On logarithmic concave measures and functions*, Acta Sci. Math. (Szeged), 34 (1973), pp. 335–343.

[61] W. Ring, *Structural properties of solutions of total variation regularization problems*, M2AN Math. Model. Numer. Anal., 34 (2000), pp. 799–810.

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

[63] S. Roth and M. J. Black, *Fields of experts*, Int. J. Comput. Vision, 82 (2009), pp. 205–229.

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

[65] H. Rue and M. A. Hurn, *Loss functions for Bayesian image analysis*, Scand. J. Statist., 24 (1997), pp. 103–114.

[66] J. Salmon, *Agrégation d'estimateurs et méthodes à patch pour le débruitage d'images numériques*, Ph.D. thesis, Université Paris-Diderot, Paris, France, 2010.

[67] U. Schmidt, Q. Gao, and S. Roth, *A generative perspective on MRFs in low-level vision*, in Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, 2010, pp. 1751–1758.

[68] C. E. Shannon, *Communication in the presence of noise*, Proc. I.R.E., 37 (1949), pp. 10–21.
[69] H. V. T. Si, *Une remarque sur la formule du changement de variables dans R^n*, Bull. Soc. Roy. Sci. Liège, 73 (2004), pp. 21–25.
[70] C. Stein, *Estimation of the mean of a multivariate normal distribution*, Ann. Statist., 9 (1981), pp. 1135–1151.
[71] J. Tropp, *Just relax: Convex programming methods for identifying sparse signals in noise*, IEEE Trans. Inform. Theory, 52 (2006), pp. 1030–1051.
[72] S. Vaiter, C. A. Deledalle, G. Peyré, C. Dossal, and J. Fadili, *Local behavior of sparse analysis regularization: Applications to risk estimation*, Appl. Comput. Harmon. Anal., 35 (2013), pp. 433–451.
[73] D. Vandeville, *SURE-based non-local means*, IEEE Signal Process. Lett., 16 (2009), pp. 973–976.
[74] P. Weiss, L. Blanc-Féraud, and G. Aubert, *Efficient schemes for total variation minimization under constraints in image processing*, SIAM J. Sci. Comput., 31 (2009), pp. 2047–2080.
[75] O. J. Woodford, C. Rother, and V. Kolmogorov, *A global perspective on MAP inference for low-level vision*, Proceedings of the 2009 IEEE 12th International Conference on Computer Vision (ICCV), 2009, pp. 2319–2326.
[76] C. Zach, T. Pock, and H. Bischof, *A duality based approach for realtime TV-L^1 optical flow*, in Proceedings of the 29th DAGM Symposium on Pattern Recognition, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 214–223.