

# Mathematical Foundations of Data Sciences



Gabriel Peyré  
CNRS & DMA  
École Normale Supérieure  
[gabriel.peyre@ens.fr](mailto:gabriel.peyre@ens.fr)  
<https://mathematical-tours.github.io>  
[www.numerical-tours.com](http://www.numerical-tours.com)

March 19, 2020



## Chapter 7

# Variational Priors and Regularization

### 7.1 Sobolev and Total Variation Priors

The simplest prior are obtained by integrating local differential quantity over the image. They corresponds to norms in functional spaces that imposes some smoothness on the signal of the image. We detail here the Sobolev and the total variation priors, that are the most popular in image processing.

#### 7.1.1 Continuous Priors

In the following, we consider either continuous functions  $f \in L^2([0, 1]^2)$  or discrete vectors  $f \in \mathbb{R}^N$ , and consider continuous priors and there discrete counterparts in Section 7.1.2.

**Sobolev prior.** The prior energy  $J(f) \in \mathbb{R}$  is intended to be low for images in a class  $f \in \Theta$ . The class of uniformly smooth functions detailed in Section 4.2.1 corresponds to functions in Sobolev spaces. A simple prior derived from this Sobolev class is thus

$$J_{\text{Sob}}(f) = \frac{1}{2} \|f\|_{\text{Sob}}^2 = \frac{1}{2} \int \|\nabla f(x)\|^2 dx, \quad (7.1)$$

where  $\nabla f$  is the gradient in the sense of distributions.

**Total variation prior.** To take into account discontinuities in images, one considers a total variation energy, introduced in Section 4.2.3. It was introduced for image denoising by Rudin, Osher and Fatemi [22]

The total variation of a smooth image  $f$  is defined as

$$J_{\text{TV}}(f) = \|f\|_{\text{TV}} = \int \|\nabla_x f\| dx. \quad (7.2)$$

This energy extends to non-smooth functions of bounded variations  $f \in \text{BV}([0, 1]^2)$ . This class contains indicators functions  $f = 1_\Omega$  of sets  $\Omega$  with a bounded perimeter  $|\partial\Omega|$ .

The total variation norm can be computed alternatively using the co-area formula (4.12), which shows in particular that  $\|1_\Omega\|_{\text{TV}} = |\partial\Omega|$ .

#### 7.1.2 Discrete Priors

An analog image  $f \in L^2([0, 1]^2)$  is discretized through an acquisition device to obtain a discrete image  $f \in \mathbb{R}^N$ . Image processing algorithms work on these discrete data, and we thus need to define discrete priors for finite dimensional images.

**Discrete gradient.** Discrete Sobolev and total variation priors are obtained by computing finite differences approximations of derivatives, using for instance forward differences

$$\begin{aligned}\delta_1 f_{n_1, n_2} &= f_{n_1+1, n_2} - f_{n_1, n_2} \\ \delta_2 f_{n_1, n_2} &= f_{n_1, n_2+1} - f_{n_1, n_2},\end{aligned}$$

and one can use higher order schemes to process more precisely smooth functions. One should be careful with boundary conditions, and we consider here for simplicity periodic boundary conditions, which correspond to computing the indexes  $n_i + 1$  modulo  $N$ . More advanced symmetric boundary conditions can be used as well to avoid boundary artifacts.

A discrete gradient is defined as

$$\nabla f_n = (\delta_1 f_n, \delta_2 f_n) \in \mathbb{R}^2$$

which corresponds to a mapping from images to vector fields

$$\nabla : \mathbb{R}^N \longrightarrow \mathbb{R}^{N \times 2}.$$

Figure 7.1 shows examples of gradient vectors. They point in the direction of the largest slope of the function discretized by  $f$ . Figure 7.2 shows gradients and their norms displayed as an image. Regions of high gradients correspond to large intensity variations, and thus typically to edges or textures.



Figure 7.1: Discrete gradient vectors.

**Discrete divergence.** One can also use backward differences,

$$\begin{aligned}\tilde{\delta}_1 f_{n_1, n_2} &= f_{n_1, n_2} - f_{n_1-1, n_2} \\ \tilde{\delta}_2 f_{n_1, n_2} &= f_{n_1, n_2} - f_{n_1, n_2-1}.\end{aligned}$$

They are dual to the forward differences, so that

$$\delta_i^* = -\tilde{\delta}_i,$$

which means that

$$\forall f, g \in \mathbb{R}^N, \quad \langle \delta_i f, g \rangle = -\langle f, \tilde{\delta}_i g \rangle,$$

which is a discrete version of the integration by part formula

$$\int_0^1 f' g = - \int_0^1 f g'$$

for smooth periodic functions on  $[0, 1]$ .

A divergence operator is defined using backward differences,

$$\text{div}(v)_n = \tilde{\delta}_1 v_{1,n} + \tilde{\delta}_2 v_{2,n},$$

and corresponds to a mapping from vector fields to images

$$\text{div} : \mathbb{R}^{N \times 2} \longrightarrow \mathbb{R}^N.$$

It is related to the dual of the gradient

$$\text{div} = -\nabla^*$$

which means that

$$\forall f \in \mathbb{R}^N, \forall v \in \mathbb{R}^{N \times 2}, \quad \langle \nabla f, v \rangle_{\mathbb{R}^{N \times 2}} = -\langle f, \text{div}(v) \rangle_{\mathbb{R}^N}$$

which corresponds to a discrete version of the divergence theorem.

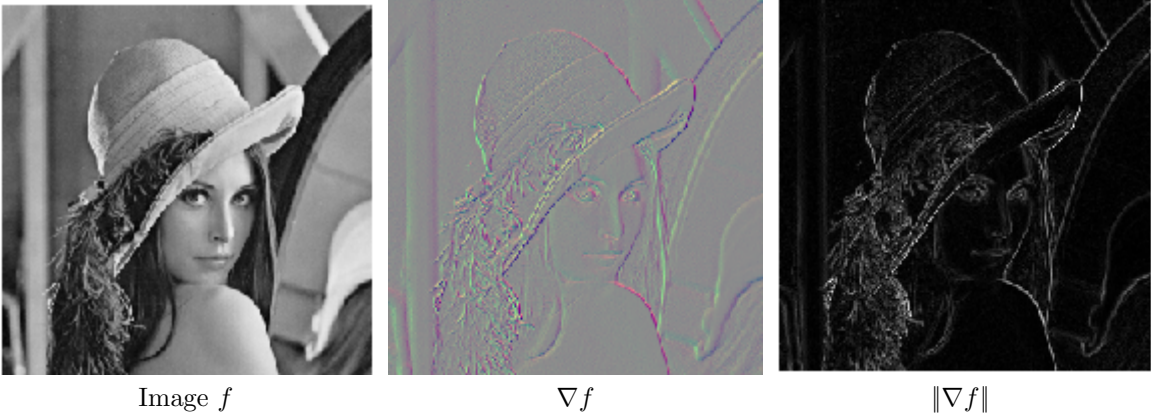


Figure 7.2: Discrete operators.

**Discrete laplacian.** A general definition of a Laplacian is

$$\Delta f = \text{div}(\nabla f),$$

which corresponds to a semi-definite negative operator.

For discrete images, and using the previously defined gradient and divergence, it is a local high pass filter

$$\Delta f_n = \sum_{p \in V_4(n)} f_p - 4f_n, \quad (7.3)$$

that approximates the continuous second order derivative

$$\frac{\partial^2 f}{\partial x_1^2}(x) + \frac{\partial^2 f}{\partial x_2^2} \approx N^2 \Delta f_n \quad \text{for } x = n/N.$$

Laplacian operators thus correspond to filterings. A continuous Laplacian is equivalently defined over the Fourier domain in diagonal form as

$$g = \Delta f \implies \hat{g}(\omega) = \|\omega\|^2 \hat{f}(\omega)$$

and the discrete Laplacian (7.3) as

$$g = \Delta f \implies \hat{g}_\omega = \rho_\omega^2 \hat{f}(\omega) \quad \text{where} \quad \rho_\omega^2 = \sin\left(\frac{\pi}{N}\omega_1\right)^2 + \sin\left(\frac{\pi}{N}\omega_2\right)^2. \quad (7.4)$$

**Discrete energies.** A discrete Sobolev energy is obtained by using the  $\ell^2$  norm of the discrete gradient vector field

$$J_{\text{Sob}}(f) = \frac{1}{2} \sum_n (\delta_1 f_n)^2 + (\delta_2 f_n)^2 = \frac{1}{2} \|\nabla f\|^2. \quad (7.5)$$

Similarly, a discrete TV energy is defined as the  $\ell^1$  norm of the gradient field

$$J_{\text{TV}}(f) = \sum_n \sqrt{(\delta_1 f_n)^2 + (\delta_2 f_n)^2} = \|\nabla f\|_1 \quad (7.6)$$

where the  $\ell^1$  norm of a vector field  $v \in \mathbb{R}^{N \times 2}$  is

$$\|v\|_1 = \sum_n \|v_n\| \quad (7.7)$$

where  $v_n \in \mathbb{R}^2$ .

## 7.2 PDE and Energy Minimization

Image smoothing is obtained by minimizing the prior using a gradient descent.

### 7.2.1 General Flows

The gradient of the prior  $J : \mathbb{R}^N \rightarrow \mathbb{R}$  is a vector  $\text{grad } J(f)$ . It describes locally up to the first order the variation of the prior

$$J(f + \varepsilon) = J(f) + \langle \varepsilon, \text{grad } J(f) \rangle + o(\|\varepsilon\|).$$

If  $J$  is a smooth function of the image  $f$ , a discrete energy minimization is obtained through a gradient descent

$$f^{(k+1)} = f^{(k)} - \tau \text{grad } J(f^{(k)}), \quad (7.8)$$

where the step size  $\tau$  must be small enough to guarantee convergence.

For infinitesimal step size  $\tau$ , one replaces the discrete parameter  $k$  by a continuous time, and the flow

$$t > 0 \mapsto f_t \in \mathbb{R}^N$$

solves the following partial differential equation

$$\frac{\partial f_t}{\partial t} = -\text{grad } J(f_t) \quad \text{and} \quad f_0 = f. \quad (7.9)$$

The gradient descent can be seen as an explicit discretization in time of this PDE at times  $t_k = k\tau$ .

### 7.2.2 Heat Flow

The heat flow corresponds to the instantiation of the generic PDE (7.9) to the case of the Sobolev energies  $J_{\text{Sob}}(f)$  defined for continuous function in (7.1) and for discrete images in (7.5).

Since it is a quadratic energy, its gradient is easily computed

$$J(f + \varepsilon) = \frac{1}{2} \|\nabla f + \nabla \varepsilon\|^2 = J(f) - \langle \Delta f, \varepsilon \rangle + o(\|\varepsilon\|^2),$$

so that

$$\text{grad } J_{\text{Sob}}(f) = -\Delta f.$$

Figure 7.4, left, shows an example of Laplacian. It is typically large (positive or negative) near edges.

The heat flow is thus

$$\frac{\partial f_t}{\partial t}(x) = -(\text{grad } J(f_t))(x) = \Delta f_t(x) \quad \text{and} \quad f_0 = f. \quad (7.10)$$



Figure 7.3: Display of  $f_t$  for increasing time  $t$  for heat flow (top row) and TV flow (bottom row).

**Continuous in space.** For continuous images and an unbounded domain  $\mathbb{R}^2$ , the PDE (7.10) has an explicit solution as a convolution with a Gaussian kernel of increasing variance as time increases

$$f_t = f \star h_t \quad \text{where} \quad h_t(x) = \frac{1}{4\pi t} e^{-\frac{\|x\|^2}{4t}}. \quad (7.11)$$

This shows the regularizing property of the heat flow, that operates a blurring to make the image more regular as time evolves.

**Discrete in space.** The discrete Sobolev energy (7.5) minimization defined a PDE flow that is discrete in space

$$\frac{\partial f_{n,t}}{\partial t} = -(\text{grad } J(f_t))_n = (\Delta f_t)_n.$$

It can be further discretized in time as done in (7.8) and leads to a fully discrete flow

$$f_n^{(k+1)} = f_n^{(k)} + \tau \left( \sum_{p \in V_4(n)} f_p - 4f_n \right) = (f \star h)_n$$

where  $V_4(n)$  are the four neighbor to a pixel  $n$ . The flow thus corresponds to iterative convolutions

$$f^{(k)} = f \star h \star \dots \star h = f \star^k h.$$

where  $h$  is a discrete filter.

It can be shown to be stable and convergent if  $\tau < 1/4$ .

### 7.2.3 Total Variation Flows

**Total variation gradient.** The total variation energy  $J_{\text{TV}}$ , both continuous (7.2) and discrete (7.6) is not a smooth function of the image. For instance, the discrete  $J_{\text{TV}}$  is non-differentiable at an image  $f$  such that there exists a pixel  $n$  where  $\nabla f_n = 0$ .

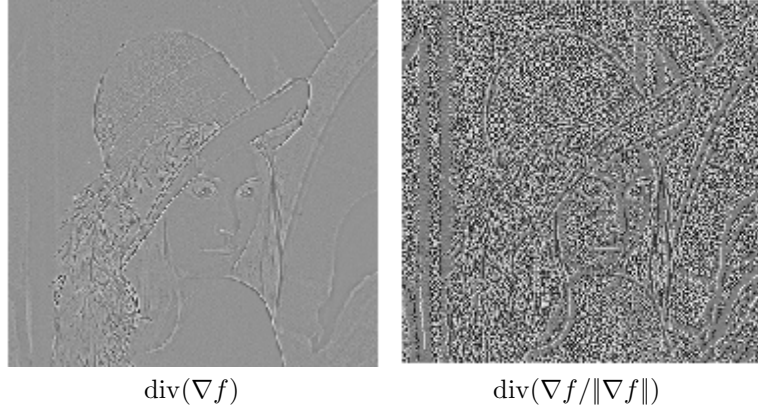


Figure 7.4: Discrete Laplacian and discrete TV gradient.

If  $\nabla f_n \neq 0$ , one can compute the gradient of the TV energy specialized at pixel  $n$  as

$$(\text{grad } J(f))_n = -\text{div} \left( \frac{\nabla f}{\|\nabla f\|} \right)_n$$

which exhibits a division by zero singularity for a point with vanishing gradient. Figure 7.4 shows an example of TV gradient, which appears noisy in smooth areas, because  $\|\nabla f_n\|$  is small in such regions.

This non-differentiability makes impossible the definition of a gradient descent and a TV flow.

**Regularized total variation.** To avoid this issue, one can modify the TV energy, and define a smoothed TV prior

$$J_{\text{TV}}^\varepsilon(f) = \sum_n \sqrt{\varepsilon^2 + \|\nabla f_n\|^2} \quad (7.12)$$

where  $\varepsilon > 0$  is a small regularization parameter. Figure 7.5 shows this effect of this regularization on the absolute value function.

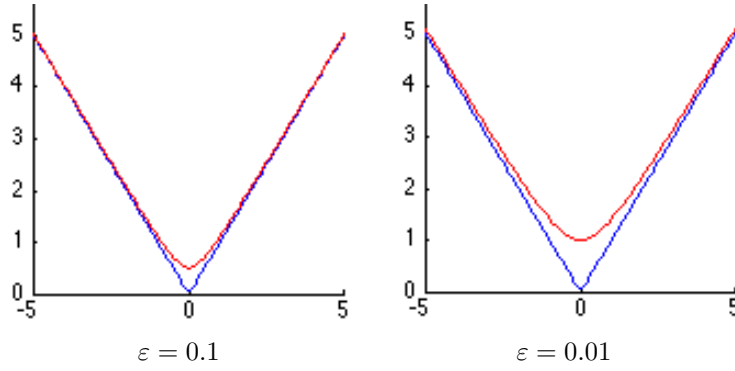


Figure 7.5: Regularized absolute value  $x \mapsto \sqrt{x^2 + \varepsilon^2}$ .

This smoothed TV energy is a differentiable function of the image, and its gradient is

$$\text{grad } J_{\text{TV}}^\varepsilon(f) = -\text{div} \left( \frac{\nabla f}{\sqrt{\varepsilon^2 + \|\nabla f\|^2}} \right). \quad (7.13)$$



One can see that this smoothing interpolate between TV and Sobolev, as

$$\text{grad}_f^\varepsilon \sim -\Delta/\varepsilon \quad \text{when} \quad \varepsilon \rightarrow +\infty.$$

Figure 7.6 shows the evolution of this gradient for several value of the smoothing parameter.

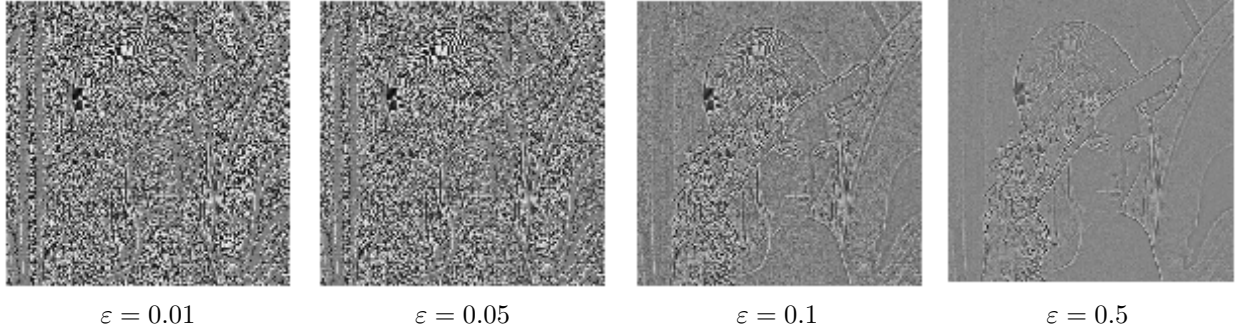


Figure 7.6: Regularized gradient norm  $\sqrt{\|\nabla f(x)\|^2 + \varepsilon^2}$ .

**Regularized total variation flow.** The smoothed total variation flow is then defined as

$$\frac{\partial f_t}{\partial t} = \text{div} \left( \frac{\nabla f_t}{\sqrt{\varepsilon^2 + \|\nabla f_t\|^2}} \right). \quad (7.14)$$

Choosing a small  $\varepsilon$  makes the flow closer to a minimization of the total variation, but makes the computation unstable.

In practice, the flow is computed with a discrete gradient descent (7.8). For the smoothed total variation flow to converge, one needs to impose that  $\tau < \varepsilon/4$ , which shows that being more faithful to the TV energy requires smaller time steps and thus slower algorithms.

Figure 7.3 shows a comparison between the heat flow and the total variation flow for a small value of  $\varepsilon$ . This shows that the TV flow smooth less the edges than heat diffusion, which is consistent with the ability of the TV energy to better characterize sharp edges.

#### 7.2.4 PDE Flows for Denoising

PDE flows can be used to remove noise from an observation  $f = f_0 + w$ . As detailed in Section 6.1.2 a simple noise model assumes that each pixel is corrupted with a Gaussian noise  $w_n \sim \mathcal{N}(0, \sigma)$ , and that these perturbations are independent (white noise).

The denoising is obtained using the PDE flow within initial image  $f$  at time  $t = 0$

$$\frac{\partial f_t}{\partial t} = -\text{grad}_{f_t} J \quad \text{and} \quad f_{t=0} = f.$$

An estimator  $\tilde{f} = f_{t_0}$  is obtained for a well chose  $t = t_0$ . Figure 7.7 shows examples of Sobolev and TV flows for denoising.

Since  $f_t$  converges to a constant image when  $t \rightarrow +\infty$ , the choice of  $t_0$  corresponds to a tradeoff between removing enough noise and not smoothing too much the edges in the image. This is usually a difficult task. During simulation, if one has access to the clean image  $f_0$ , one can monitor the denoising error  $\|f_0 - f_t\|$  and choose the  $t = t_0$  that minimizes this error. Figure 7.8, top row, shows an example of this oracle estimation of the best stopping time.



Figure 7.7: Denoising using  $f_t$  displayed for various time  $t$  for Sobolev (top) and TV (bottom) flows.

## 7.3 Regularization for Denoising

Instead of performing a gradient descent flow for denoising as detailed in Section 7.2.4 and select a stopping time, one can formulate an optimization problem. The estimator is then defined as a solution of this optimization. This setup has the advantage as being well defined mathematically even for non-smooth priors such as the TV prior  $J_{TV}$  or the sparsity prior  $J_1$ . Furthermore, this regularization framework is also useful to solve general inverse problems as detailed in Chapter ??.

### 7.3.1 Regularization

Given some noisy image  $f = f_0 + w$  of  $N$  pixels and a prior  $J$ , we consider the convex optimization problem

$$f_\lambda^* \in \operatorname{argmin}_{g \in \mathbb{R}^N} \frac{1}{2} \|f - g\|^2 + \lambda J(g), \quad (7.15)$$

where  $\lambda > 0$  is a Lagrange multiplier parameter that weights the influence of the data fitting term  $\|f - g\|^2$  and the regularization term  $J(g)$ .

If one has at his disposal a clean original image  $f_0$ , one can minimize the denoising error  $\|f_\lambda^* - f_0\|$ , but it is rarely the case. In practice, this parameter should be adapted to the noise level and to the regularity of the unknown image  $f_0$ , which might be a non trivial task.

We note that since we did not impose  $J$  to be convex, the problem (7.15) might have several optimal solutions.

An estimator is thus defined as

$$\tilde{f} = f_\lambda^*$$

for a well chosen  $\lambda$ .

If  $J$  is differentiable and convex, one can compute the solution of (7.15) through a gradient descent

$$f^{(k+1)} = f^{(k)} - \tau \left( f^{(k)} - \lambda \operatorname{grad} J(f^{(k)}) \right) \quad (7.16)$$

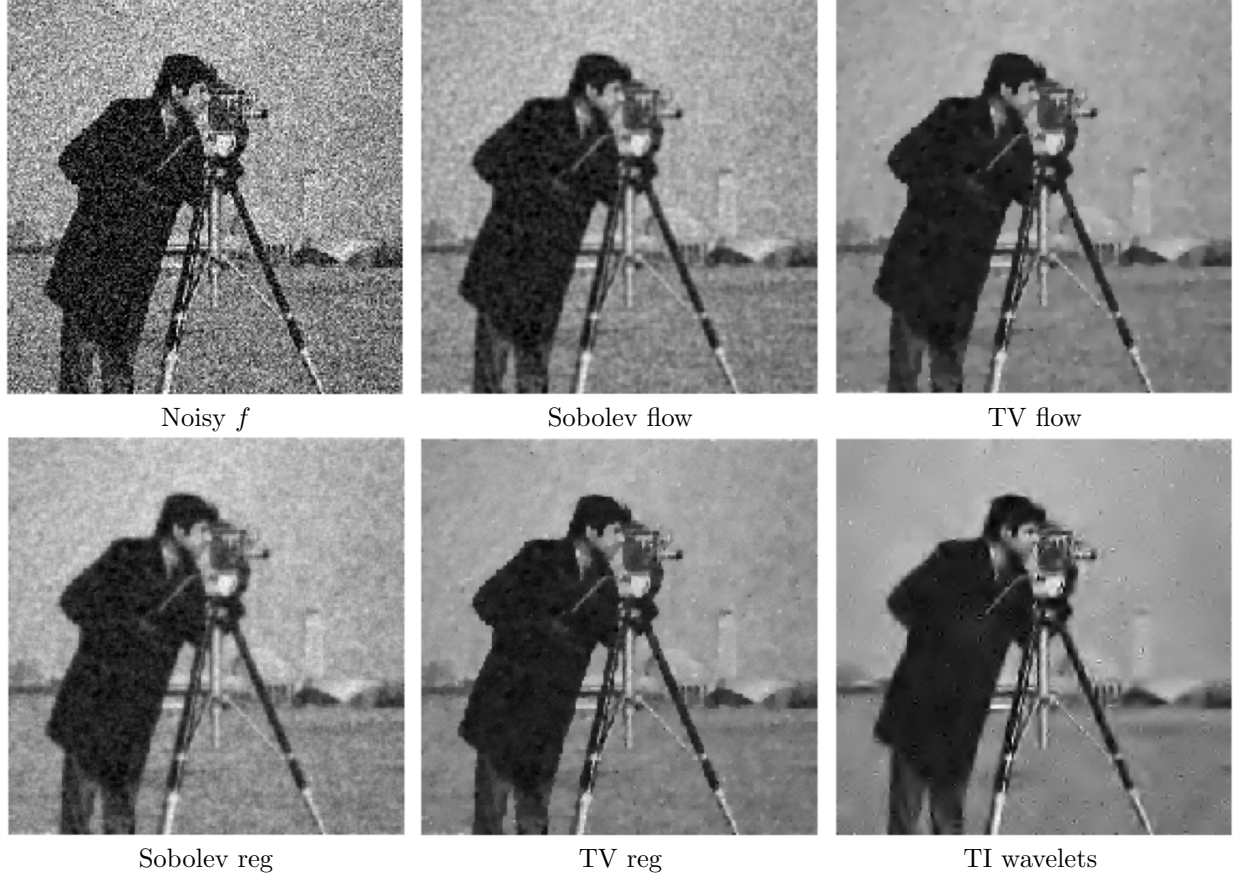


Figure 7.8: Denoising using PDE flows and regularization.

where the descent step size  $\tau > 0$  should be small enough. This gradient descent is similar to the time-discretized minimization flow (7.8), excepted that the data fitting term prevent the flow to converge to a constant image.

Unfortunately, priors such as the total variation  $J_{TV}$  or the sparsity  $J_1$  are non-differentiable. Some priors such as the ideal sparsity  $J_0$  might even be non-convex. This makes the simple gradient descent not usable to solve for (7.15). In the following Section we show how to compute  $f_\lambda^*$  for several priors.

### 7.3.2 Sobolev Regularization

The discrete Sobolev prior defined in (7.5) is differentiable, and the gradient descent (7.16) reads

$$f^{(k+1)} = (1 - \tau)f^{(k)} + \tau f - \tau \lambda \Delta J(f^{(k)}).$$

Since  $J(f) = \|\nabla f\|^2$  is quadratic, one can use a conjugate gradient descent, which converges faster.

The solution  $f_\lambda^*$  can be computed in closed form as the solution of a linear system

$$f_\lambda^* = (\text{Id}_N - \lambda \Delta)^{-1} f,$$

which shows that the regularization (7.15) is computing an estimator that depends linearly on the observations  $f$ .

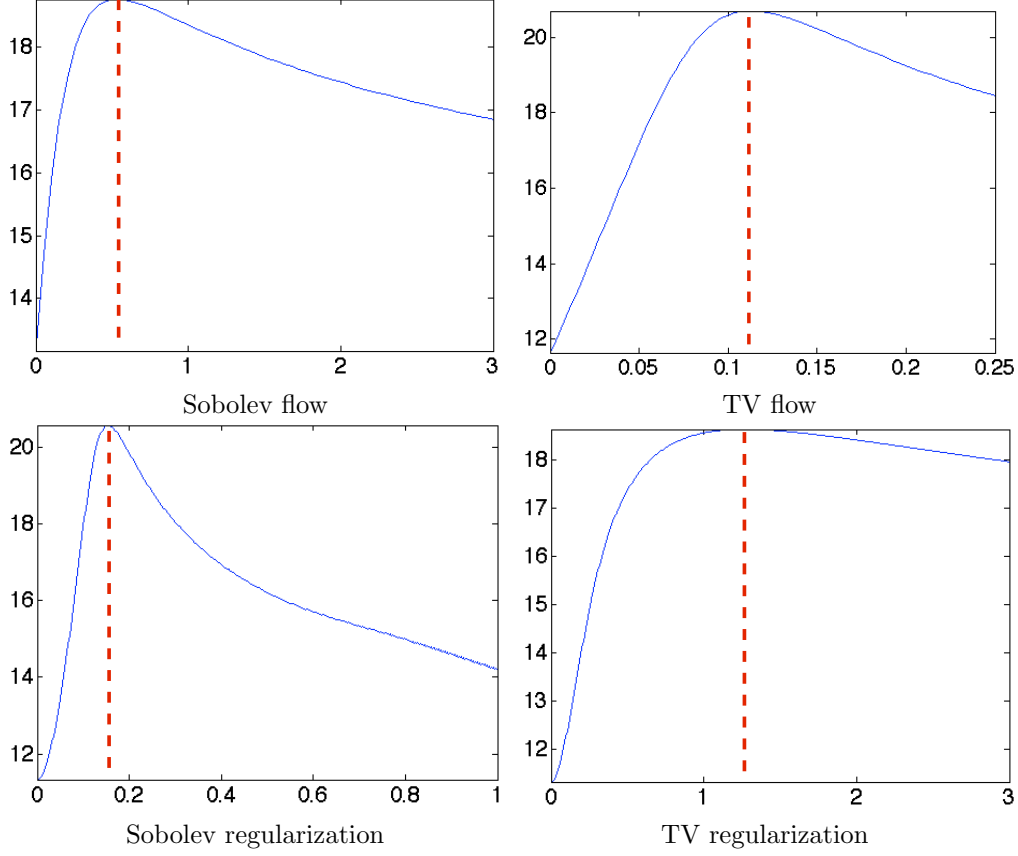


Figure 7.9: SNR as a function of time  $t$  for flows (top) and  $\lambda$  for regularization (bottom).

If the differential operators are computed with periodic boundary conditions, this linear system can be solved exactly over the Fourier domain

$$(\hat{f}_\lambda^*)_\omega = \frac{1}{1 + \lambda \rho_\omega^2} \hat{f}_\omega \quad (7.17)$$

where  $\rho_\omega$  depends on the discretization of the Laplacian, see for instance (7.4).

Equation (7.17) shows that denoising using Sobolev regularization corresponds to a low pass filtering, whose strength is controlled by  $\lambda$ . This should be related to the solution (7.11) of the heat equation, which also corresponds to a filtering, but using a Gaussian low pass kernel parameterized by its variance  $t^2$ .

This Sobolev regularization denoising is a particular case of the linear estimator considered in Section 6.2. The selection of the parameter  $\lambda$  is related to the selection of an optimal filter as considered in Section 6.2.2, but with the restriction that the filter is computed in a parametric family.

### 7.3.3 TV Regularization

The total variation prior  $J_{TV}$  defined in (7.6) is non-differentiable. One can either use a smoothed approximation of the prior, or use an optimization algorithm not based on a gradient descent.

The TV prior can be approximated to obtain the prior  $J_{TV}^\varepsilon(g)$  defined in (7.12), which is differentiable with respect to  $g$ . Using the gradient of this prior defined in (7.13), one obtains the following instantiation

of the gradient descent (7.16)

$$f^{(k+1)} = (1 - \tau)f^{(k)} + \tau f + \lambda \tau \operatorname{div} \left( \frac{\nabla f_t}{\sqrt{\varepsilon^2 + \|\nabla f_t\|^2}} \right). \quad (7.18)$$

which converge to the unique minimizer  $f_\lambda^*$  of (7.15).

Section 12.4.1 details a better alternative which does not require introducing this  $\varepsilon$  smoothing.



# Bibliography

- [1] Amir Beck. *Introduction to Nonlinear Optimization: Theory, Algorithms, and Applications with MATLAB*. SIAM, 2014.
- [2] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, and Jonathan Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.
- [3] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
- [4] E. Candès and D. Donoho. New tight frames of curvelets and optimal representations of objects with piecewise  $C^2$  singularities. *Commun. on Pure and Appl. Math.*, 57(2):219–266, 2004.
- [5] E. J. Candès, L. Demanet, D. L. Donoho, and L. Ying. Fast discrete curvelet transforms. *SIAM Multiscale Modeling and Simulation*, 5:861–899, 2005.
- [6] A. Chambolle. An algorithm for total variation minimization and applications. *J. Math. Imaging Vis.*, 20:89–97, 2004.
- [7] Antonin Chambolle, Vicent Caselles, Daniel Cremers, Matteo Novaga, and Thomas Pock. An introduction to total variation for image analysis. *Theoretical foundations and numerical methods for sparse recovery*, 9(263-340):227, 2010.
- [8] Antonin Chambolle and Thomas Pock. An introduction to continuous optimization for imaging. *Acta Numerica*, 25:161–319, 2016.
- [9] S.S. Chen, D.L. Donoho, and M.A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20(1):33–61, 1999.
- [10] Philippe G Ciarlet. Introduction à l’analyse numérique matricielle et à l’optimisation. 1982.
- [11] P. L. Combettes and V. R. Wajs. Signal recovery by proximal forward-backward splitting. *SIAM Multiscale Modeling and Simulation*, 4(4), 2005.
- [12] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Commun. on Pure and Appl. Math.*, 57:1413–1541, 2004.
- [13] D. Donoho and I. Johnstone. Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, 81:425–455, Dec 1994.
- [14] Heinz Werner Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems*, volume 375. Springer Science & Business Media, 1996.
- [15] M. Figueiredo and R. Nowak. An EM Algorithm for Wavelet-Based Image Restoration. *IEEE Trans. Image Proc.*, 12(8):906–916, 2003.
- [16] Simon Foucart and Holger Rauhut. *A mathematical introduction to compressive sensing*, volume 1. Birkhäuser Basel, 2013.

- [17] Stephane Mallat. *A wavelet tour of signal processing: the sparse way*. Academic press, 2008.
- [18] D. Mumford and J. Shah. Optimal approximation by piecewise smooth functions and associated variational problems. *Commun. on Pure and Appl. Math.*, 42:577–685, 1989.
- [19] Neal Parikh, Stephen Boyd, et al. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.
- [20] Gabriel Peyré. *L’algèbre discrète de la transformée de Fourier*. Ellipses, 2004.
- [21] J. Portilla, V. Strela, M.J. Wainwright, and Simoncelli E.P. Image denoising using scale mixtures of Gaussians in the wavelet domain. *IEEE Trans. Image Proc.*, 12(11):1338–1351, November 2003.
- [22] L. I. Rudin, S. Osher, and E. Fatemi. Nonlinear total variation based noise removal algorithms. *Phys. D*, 60(1-4):259–268, 1992.
- [23] Otmar Scherzer, Markus Grasmair, Harald Grossauer, Markus Haltmeier, Frank Lenzen, and L Sirovich. *Variational methods in imaging*. Springer, 2009.
- [24] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [25] Jean-Luc Starck, Fionn Murtagh, and Jalal Fadili. *Sparse image and signal processing: Wavelets and related geometric multiscale analysis*. Cambridge university press, 2015.