

Image Segmentation with Stabilized Inverse Diffusion Equations

■ 3.1 Introduction.

IN this chapter, we introduce the Stabilized Inverse Diffusion Equations (SIDEs), as well as illustrate their speed and robustness, in comparison with some of the methods reviewed in Chapter 2. As we mentioned in the previous chapter, the starting point for the development of SIDEs were image restoration and segmentation procedures based on PDEs of evolution [1, 12, 46, 49, 50, 55–57, 72]. We observed that the numerical schemes for solving such equations do not necessarily exhibit the behavior of the equations themselves. We therefore concentrate in this thesis on semi-discrete scale spaces (i.e., continuous in scale and discrete in space). More specifically, SIDEs, which are the main focus and contribution of this thesis, are a new family of semi-discrete evolution equations which stably sharpen edges and suppress noise. We will see that SIDEs may be viewed as a conceptually limiting case of Perona-Malik diffusions which were reviewed in the previous chapter. SIDEs have discontinuous right-hand sides and act as inverse diffusions “almost everywhere”, with stabilization resulting from the presence of discontinuities in the vector field defined by the evolution. The scale space of such an equation is a family of segmentations of the original image, with larger values of the scale parameter t corresponding to segmentations at coarser scales. Moreover, in contrast to continuous evolutions, the ones introduced here naturally define a sequence of logical “stopping times”, i.e. points along the evolution endowed with useful information, and corresponding to times at which the evolution hits a discontinuity surface of its defining vector field.

In the next section we begin by describing a convenient mechanical analog for the visualization of many spatially-discrete evolution equations, including discretized linear or nonlinear diffusions such as that of Perona and Malik, as well as the discontinuous equations that we introduce in Section 3.3. The implementation of such a discontinuous equation naturally results in a recursive region merging algorithm. Because of the discontinuous right-hand side of SIDEs, some care must be taken in defining solutions, but as we show in Section 3.4, once this is done, the resulting evolutions have a number of important properties. Moreover, as we have indicated, they lead to very effective

algorithms for edge enhancement and segmentation, something that is demonstrated in Section 3.5. In particular, as we will see, they can produce sharp enhancement of edges in high noise as well as accurate segmentations of very noisy imagery such as SAR and ultrasound imagery subject to severe speckle. In Section 3.6, we point out its principal differences from Koepfler, Lopez, and Morel's [36] region merging procedure for minimizing the Mumford-Shah functional [44]. The rest of that section is devoted to exploring the links with other important work in the field reviewed in Chapter 2: the total variation approach [6, 56, 58]; shock filters of Osher and Rudin [46]; the robust variational formulation of D. Geman and Reynolds [21]; and the stochastic modeling approach of Zhu and Mumford [75].

■ 3.2 A Spring-Mass Model for Certain Evolution Equations.

As we indicated in the introduction, the focus of this chapter is on discrete-space, continuous-time evolutions of the following general form:

$$\begin{aligned}\dot{\mathbf{u}}(t) &= \mathcal{F}(\mathbf{u})(t), \\ \mathbf{u}(0) &= \mathbf{u}^0,\end{aligned}\tag{3.1}$$

where \mathbf{u} is either a discrete sequence consisting of N samples ($\mathbf{u} = (u_1, \dots, u_N)^T \in \mathbb{R}^N$), or an N -by- N image whose j -th entry in the i -th row is u_{ij} ($\mathbf{u} \in \mathbb{R}^{N^2}$). The initial condition \mathbf{u}^0 corresponds to the original signal or image to be processed, and $\mathbf{u}(t)$ then represents the evolution of this signal/image at time (scale) t , resulting in a scale-space family for $0 \leq t < \infty$.

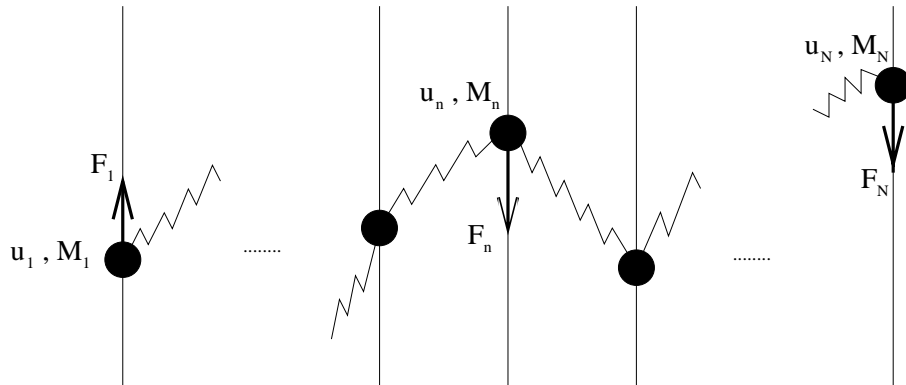


Figure 3.1. A spring-mass model.

The nonlinear operators \mathcal{F} of interest in this chapter can be conveniently visualized through the following simple mechanical model. For the sake of simplicity in visualization, let us first suppose that $\mathbf{u} \in \mathbb{R}^N$ is a one-dimensional (1-D) sequence, and interpret $\mathbf{u}(t) = (u_1(t), \dots, u_N(t))^T$ in (3.1) as the vector of vertical positions of the N particles of masses M_1, \dots, M_N , depicted in Figure 3.1. The particles are forced to

move along N vertical lines. Each particle is connected by springs to its two neighbors (except the first and last particles, which are only connected to one neighbor.) Every spring whose vertical extent is v has energy $E(v)$, i.e., the energy of the spring between the n -th and $(n + 1)$ -st particles is $E(u_{n+1} - u_n)$. We impose the usual requirements on this energy function:

$$\begin{aligned} E(v) &\geq 0, \\ E(0) &= 0, \\ E'(v) &\geq 0 \text{ for } v > 0, \\ E(v) &= E(-v). \end{aligned} \tag{3.2}$$

Then the derivative of $E(v)$, which we refer to as “the force function” and denote by $F(v)$, satisfies

$$\begin{aligned} F(0) &= 0, \\ F(v) &\geq 0 \text{ for } v > 0, \\ F(v) &= -F(-v). \end{aligned} \tag{3.3}$$

We also call $F(v)$ a “force function” and $E(v)$ an “energy” if $-E(v)$ satisfies (3.2) and $-F(v)$ satisfies (3.3). We make the movement of the particles non-conservative by stopping it after a small period of time Δt and re-starting with zero velocity. (Note that this will make our equation non-hyperbolic.) It is assumed that during one such step, the total force $F_n = -F(u_n - u_{n+1}) - F(u_n - u_{n-1})$, acting on the n -th particle, stays approximately constant. The displacement during one iteration is proportional to the product of acceleration and the square of the time interval:

$$u_n(t + \Delta t) - u_n(t) = \frac{(\Delta t)^2}{2} \frac{F_n}{M_n}.$$

Letting $\Delta t \rightarrow 0$, while fixing $\frac{2M_n}{\Delta t} = m_n$, where m_n is a positive constant, leads to

$$\dot{u}_n = \frac{1}{m_n} (F(u_{n+1} - u_n) - F(u_n - u_{n-1})), \quad n = 1, 2, \dots, N, \tag{3.4}$$

with the conventions $u_0 = u_1$ and $u_{N+1} = u_N$ imposed by the absence of springs to the left of the first particle and to the right of the last particle. We will refer to m_n as “the mass of the n -th particle” in the remainder of the thesis. Note that Equation (3.4) is a (weighted) gradient descent equation for the following global energy:

$$\mathcal{E}(\mathbf{u}) = \sum_{i=1}^{N-1} E(u_{i+1} - u_i). \tag{3.5}$$

The examples below, where $m_n = 1$, clearly illustrate these notions.

Example 3.1. *Linear heat equation.*

A linear force function $F(v) = v$ leads to the semi-discrete linear heat equation

$$\dot{u}_n = u_{n+1} - 2u_n + u_{n-1}.$$

This corresponds to a simple discretization of the 1-D linear heat equation and results in evolutions which produce increasingly low-pass filtered and smoothed versions of the original signal \mathbf{u}^0 . ■

In general, $F(v)$ is called a “diffusion force” if, in addition to (3.3), it is monotonically increasing:

$$v_1 < v_2 \Rightarrow F(v_1) < F(v_2), \quad (3.6)$$

which is illustrated in Figure 3.2(a). We shall call the corresponding energy a “dif-

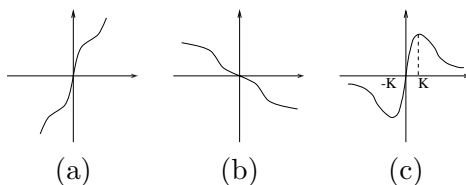


Figure 3.2. Force functions: (a) diffusion; (b) inverse diffusion; (c) Perona-Malik.

fusion energy” and the corresponding evolution (3.4) a “diffusion”. The evolution in Example 3.1 is clearly a diffusion. We call $F(v)$ an “inverse diffusion force” if $-F(v)$ satisfies Equations (3.3) and (3.6), as illustrated in Figure 3.2(b). The corresponding evolution (3.4) is called an “inverse diffusion”. Inverse diffusions have the characteristic of enhancing abrupt differences in \mathbf{u} corresponding to “edges” in the 1-D sequence. Such pure inverse diffusions, however, lead to unstable evolutions (in the sense that they greatly amplify arbitrarily small noise). The following example, which is prototypical of the examples considered by Perona and Malik, defines a stable evolution that captures at least some of the edge enhancing characteristics of inverse diffusions.

Example 3.2. *Perona-Malik equations.*

Taking $F(v) = v \exp(-(\frac{v}{K})^2)$, as illustrated in Figure 3.2(c), yields a 1-D semi-discrete (continuous in scale and discrete in space) version of the Perona-Malik equation (see equations (3.3), (3.4), and (3.12) in [50]). In general, given a positive constant K , a force $F(v)$ will be called “Perona-Malik force of thickness K ” if, in addition to (3.3), it satisfies the following conditions:

$$\begin{aligned} F(v) \text{ has a unique maximum at } v = K, \\ F(v_1) = F(v_2) \Rightarrow (|v_1| - K)(|v_2| - K) < 0. \end{aligned} \quad (3.7)$$

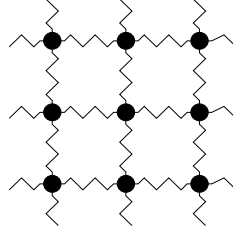


Figure 3.3. Spring-mass model in 2-D (view from above).

We shall call the corresponding energy a “Perona-Malik energy” and the corresponding evolution equation a “Perona-Malik equation of thickness K ”. As Perona and Malik demonstrate (and as can also be inferred from the results in the present thesis), evolutions with such a force function act like inverse diffusions in the regions of high gradient and like usual diffusions elsewhere. They are stable and capable of achieving some level of edge enhancement depending on the exact form of $F(v)$. ■

Finally, to extend the mechanical model of Figure 3.1 to images, we simply replace the sequence of vertical lines along which the particles move with an N -by- N square grid of such lines, as shown in Figure 3.3. The particle at location (i, j) is connected by springs to its four neighbors: $(i - 1, j)$, $(i, j + 1)$, $(i + 1, j)$, $(i, j - 1)$, except for the particles in the four corners of the square (which only have two neighbors each), and the rest of the particles on the boundary of the square (which have three neighbors). This arrangement is reminiscent of (and, in fact, was suggested by) the resistive network of Figure 8 in [49]. The analog of Equation (3.4) for images is then:

$$\begin{aligned} \dot{u}_{ij} &= \frac{1}{m_{ij}} (F(u_{i+1,j} - u_{ij}) - F(u_{ij} - u_{i-1,j})) \\ &+ F(u_{i,j+1} - u_{ij}) - F(u_{ij} - u_{i,j-1}), \end{aligned} \quad (3.8)$$

with $i = 1, 2, \dots, N$, $j = 1, 2, \dots, N$, and the conventions $u_{0,j} = u_{1,j}$, $u_{N+1,j} = u_{N,j}$, $u_{i,0} = u_{i,1}$ and $u_{i,N+1} = u_{i,N}$ imposed by the absence of springs outside of $1 \leq i \leq N$, $1 \leq j \leq N$.

■ 3.3 Stabilized Inverse Diffusion Equations (SIDs): The Definition.

In this section, we introduce a discontinuous force function, resulting in a system (3.4) that has discontinuous right-hand side (RHS). Such equations received much attention in control theory because of the wide usage of relay switches in automatic control systems [17, 67]. More recently, deliberate introduction of discontinuities has been used in control applications to drive the state vector onto lower-dimensional surfaces in the state space [67]. As we will see, this objective of driving a trajectory onto a lower-dimensional surface also has value in image analysis and in particular in image segmentation. Segmenting a signal or image, represented as a high-dimensional vector

\mathbf{u} , consists of evolving it so that it is driven onto a comparatively low-dimensional subspace which corresponds to a segmentation of the signal or image domain into a small number of regions.

The type of force function of interest to us here is illustrated in Figure 3.4. More precisely, we wish to consider force functions $F(v)$ which, in addition to (3.3), satisfy the following conditions:

$$\begin{aligned} F'(v) &\leq 0 \quad \text{for } v \neq 0, \\ F(0^+) &> 0 \\ F(v_1) = F(v_2) &\Leftrightarrow v_1 = v_2. \end{aligned} \tag{3.9}$$

Contrasting this form of a force function to the Perona-Malik function in Figure 3.2,

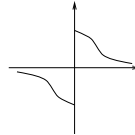


Figure 3.4. Force function for a stabilized inverse diffusion equation.

we see that in a sense one can view the discontinuous force function as a limiting form of the continuous force function in Figure 3.2(c), as $K \rightarrow 0$. However, because of the discontinuity at the origin of the force function in Figure 3.4, there is a question of how one defines solutions of Equation (3.4) for such a force function. Indeed, if Equation (3.4) evolves toward a point of discontinuity of its RHS, the value of the RHS of (3.4) apparently depends on the direction from which this point is approached (because $F(0^+) \neq F(0^-)$), making further evolution non-unique. We therefore need a special definition of how the trajectory of the evolution proceeds at these discontinuity points.¹ For this definition to be useful, the resulting evolution must satisfy well-posedness properties: the existence and uniqueness of solutions, as well as stability of solutions with respect to the initial data. In the rest of this section we describe how to define solutions to (3.4) for force functions (3.9). Assuming the resulting evolutions to be well-posed, we demonstrate that they have the desired qualitative properties, namely that they both are stable and also act as inverse diffusions and hence enhance edges. We address the issue of well-posedness and other properties in Section 3.4.

Consider the evolution (3.4) with $F(v)$ as in Figure 3.4 and Equation (3.9) and with all of the masses m_n equal to 1. Notice that the RHS of (3.4) has a discontinuity at a point \mathbf{u} if and only if $u_i = u_{i+1}$ for some i between 1 and $N - 1$. It is when a trajectory reaches such a point \mathbf{u} that we need the following definition. In terms of the spring-mass model of Figure 3.1, once the vertical positions u_i and u_{i+1} of two neighboring particles become equal, the spring connecting them is replaced by a rigid link. In other words,

¹Having such a definition is crucial because, as we will show in Section 3.4, equation (3.4) will reach a discontinuity point of its RHS in finite time, starting with any initial condition.

the two particles are simply merged into a single particle which is twice as heavy (see Figure 3.5), yielding the following modification of (3.4) for $n = i$ and $n = i + 1$:

$$\dot{u}_i = \dot{u}_{i+1} = \frac{1}{2}(F(u_{i+2} - u_{i+1}) - F(u_i - u_{i-1})).$$

(The differential equations for $n \neq i, i + 1$ do not change.) Similarly, if m consecutive

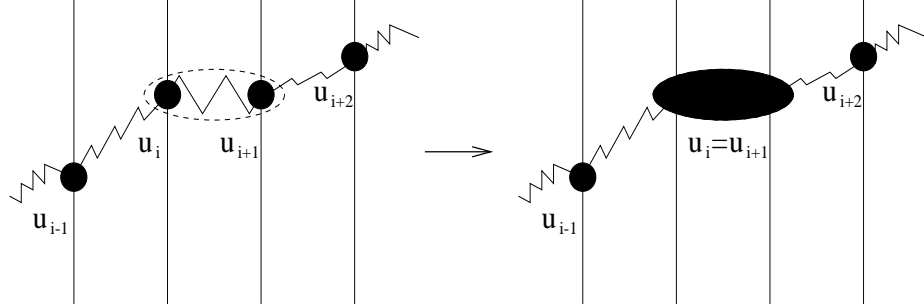


Figure 3.5. A horizontal spring is replaced by a rigid link.

particles reach equal vertical position, they are merged into one particle of mass m ($1 \leq m \leq N$):

$$\begin{aligned} \dot{u}_n &= \dots = \dot{u}_{n+m-1} = \\ &= \frac{1}{m}(F(u_{n+m} - u_{n+m-1}) - F(u_n - u_{n-1})) \end{aligned} \quad (3.10)$$

if

$$u_{n-1} \neq u_n = u_{n+1} = \dots = u_{n+m-2} = u_{n+m-1} \neq u_{n+m}.$$

Notice that this system is the same as (3.4), but with possibly unequal masses. It is convenient to re-write this equation so as to explicitly indicate the reduction in the number of state variables:

$$\begin{aligned} \dot{u}_{n_i} &= \frac{1}{m_{n_i}}(F(u_{n_{i+1}} - u_{n_{i+1}-1}) - F(u_{n_i} - u_{n_i-1})), \end{aligned} \quad (3.11)$$

$$u_{n_i} = u_{n_i+1} = \dots = u_{n_i+m_{n_i}-1},$$

where $i = 1, \dots, p$,

$$1 = n_1 < n_2 < \dots < n_{p-1} < n_p \leq N,$$

$$n_{i+1} = n_i + m_{n_i}.$$

The compound particle described by the vertical position u_{n_i} and mass m_{n_i} consists of m_{n_i} unit-mass particles $u_{n_i}, u_{n_i+1}, \dots, u_{n_i+m_{n_i}-1}$ that have been merged, as shown in Figure 3.5. The evolution can then naturally be thought of as a sequence of stages: during each stage, the right-hand side of (3.11) is continuous. Once the solution hits a discontinuity surface of the right-hand side, the state reduction and re-assignment

of m_{n_i} 's, described above, takes place. The solution then proceeds according to the modified equation until it hits the next discontinuity surface, etc.

Notice that such an evolution automatically produces a multiscale segmentation of the original signal if one views each compound particle as a region of the signal. Viewed as a segmentation algorithm, this evolution can be summarized as follows:

1. Start with the trivial initial segmentation: each sample is a distinct region.
2. Evolve (3.11) until the values in two or more neighboring regions become equal.
3. Merge the neighboring regions whose values are equal.
4. Go to step 2.

The same algorithm can be used for 2-D images, which is immediate upon re-writing Equation (3.11):

$$\dot{u}_{n_i} = \frac{1}{m_{n_i}} \sum_{n_j \in A_{n_i}} F(u_{n_j} - u_{n_i}) p_{ij}, \quad (3.12)$$

where

m_{n_i} is again the mass of the compound particle n_i (= the number of pixels in the region n_i);

A_{n_i} is the set of the indices of all the neighbors of n_i , i.e., of all the compound particles that are connected to n_i by springs;

p_{ij} is the number of springs between regions n_i and n_j (always 1 in 1-D, but can be larger in 2-D).

Just as in 1-D, two neighboring regions n_1 and n_2 are merged by replacing them with one region n of mass $m_n = m_{n_1} + m_{n_2}$ and the set of neighbors $A_n = A_{n_1} \cup A_{n_2} \setminus \{n_1, n_2\}$.

We close this section by describing one of the basic and most important properties of these evolutions, namely that the evolution is stable but nevertheless behaves like an inverse diffusion. Notice that a force function $F(v)$ satisfying (3.9) can be represented as the sum of an inverse diffusion force $F_{id}(v)$ and a positive multiple of $\text{sgn}(v)$: $F(v) = F_{id}(v) + C \text{sgn}(v)$, where $C = F(0^+)$ and $-F_{id}(v)$ satisfies (3.3) and (3.6). Therefore, if $u_{n_{i+1}} - u_{n_i}$ and $u_{n_i} - u_{n_{i-1}}$ are of the same sign (which means that u_{n_i} is not a local extremum of the sequence $(u_{n_1}, \dots, u_{n_p})$), then (3.11) can be written as

$$\dot{u}_{n_i} = \frac{1}{m_{n_i}} (F_{id}(u_{n_{i+1}} - u_{n_i}) - F_{id}(u_{n_i} - u_{n_{i-1}})). \quad (3.13)$$

If $u_{n_i} > u_{n_{i+1}}$ and $u_{n_i} > u_{n_{i-1}}$ (i.e., u_{n_i} is a local maximum), then (3.11) is

$$\dot{u}_{n_i} = \frac{1}{m_{n_i}} (F_{id}(u_{n_{i+1}} - u_{n_i}) - F_{id}(u_{n_i} - u_{n_{i-1}}) - 2C). \quad (3.14)$$

If $u_{n_i} < u_{n_{i+1}}$ and $u_{n_i} < u_{n_{i-1}}$ (i.e., u_{n_i} is a local minimum), then (3.11) is

$$\dot{u}_{n_i} = \frac{1}{m_{n_i}}(F_{id}(u_{n_{i+1}} - u_{n_i}) - F_{id}(u_{n_i} - u_{n_{i-1}}) + 2C). \quad (3.15)$$

Equation (3.13) says that the evolution is a pure inverse diffusion at the points which are not local extrema. It is not, however, a *global* inverse diffusion, since pure inverse diffusions drive local maxima to $+\infty$ and local minima to $-\infty$ and thus are unstable. In contrast, equations (3.14) and (3.15) show that at local extrema, the evolution introduced in this chapter is an inverse diffusion plus a stabilizing term which guarantees that the local maxima do not increase and the local minima do not decrease. Indeed, $|F_{id}(v)| \leq F(0^+) = C$ for any v and for any SIDE force function F , and therefore the RHS of (3.14) is negative, and the RHS of (3.15) is positive. For this reason, we call the new evolution (3.11), (3.12) a “stabilized inverse diffusion equation” (“SIDE”), a force function satisfying (3.9) a “SIDE force”, and the corresponding energy a “SIDE energy”. In Chapter 4, we will analyze a simpler version of this equation, which results from dropping the inverse diffusion term. In this particular case, the local extrema move with constant speed and all the other samples are stationary, which makes the analysis of the equation more tractable.

■ 3.4 Properties of SIDEs.

■ 3.4.1 Basic Properties in 1-D.

The SIDEs described in the two preceding sections enjoy a number of interesting properties which validate and explain their adaptability to segmentation problems. We first examine the SIDEs in one spatial dimension for which we can make the strongest statements.

We define the n_i -th discontinuity hyperplane of a SIDE (3.11) by $S_{n_i} = \{\mathbf{u} \in \mathbb{R}^p : u_{n_i} = u_{n_{i+1}}\}$, $i = 1, \dots, p-1$. Sometimes it is more convenient to work with the vector $\mathbf{v} = (v_{n_1}, \dots, v_{n_{p-1}})^T \in \mathbb{R}^{p-1}$ of the first differences of \mathbf{u} : $v_{n_i} = u_{n_{i+1}} - u_{n_i}$, for $i = 1, \dots, p-1$. We abuse notation by also denoting $S_{n_i} = \{\mathbf{v} \in \mathbb{R}^{p-1} : v_{n_i} = 0\}$.

On such hyperplanes, we defined the solution of a SIDE as the solution to a modified, lower-dimensional, equation whose RHS is continuous on S_{n_i} . In what follows, we will assume that the SIDE force function $F(v)$ is sufficiently regular away from zero, so that the ODE (3.11), restricted to the domain of continuity of its RHS, is well-posed. As a result, existence and uniqueness of solutions of SIDEs immediately follow from the existence and uniqueness of solutions of ODEs with continuous RHS. Continuous dependence on the initial data is also guaranteed for a trajectory segment lying inside a region of continuity of the RHS. In order to show, however, that the solutions that we have defined are continuous with respect to initial conditions over *arbitrary* time intervals, we must take into account the presence of discontinuities on the RHS. In particular, what must be shown is that trajectories that start very near a discontinuity surface remain close to one that starts on the surface. More precisely, we need to be

able to show that a trajectory whose initial point is very close to S_{n_i} will, in fact, hit S_{n_i} (see Figure 3.6). In the literature on differential equations and control theory [17, 67], the behavior that SIDEs exhibit and which is illustrated in Figure 3.6 is referred to as “sliding modes”. Specifically, as proven in Appendix A, the behavior of the evolution near discontinuity hyperplanes satisfies the following:

Lemma 3.1 (Lemma on Sliding). *Let σ be a permutation of (n_1, \dots, n_{p-1}) , and m an integer between 1 and $p - 1$, and let S be the set of all points in the intersection of m hyperplanes which do not belong to the remaining $p - m - 1$ hyperplanes:*

$$S = \bigcap_{q=1}^m S_{\sigma(q)} \setminus \left(\bigcup_{q=m+1}^{p-1} S_{\sigma(q)} \right).$$

Then, as \mathbf{v} approaches S from any quadrant,² its velocity is directed towards S :

$$\lim(\dot{v}_{\sigma(q)} \text{ sign}(v_{\sigma(q)})) \leq 0 \text{ for } q = 1, \dots, m,$$

and for at least one q this inequality is strict. ■

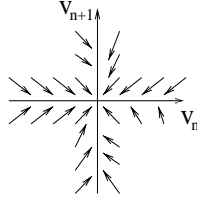


Figure 3.6. Solution field near discontinuity surfaces.

Intuitively, and as illustrated in Figure 3.6, this lemma states that the solution field of a SIDE near any discontinuity surface points toward that surface. As a consequence, a trajectory which hits such a surface may be continuously extended to “slide” along the surface, as shown in [17, 67]. For this reason the discontinuity surfaces are commonly referred to as “sliding surfaces”. For SIDEs, a simple calculation verifies that the dynamics along such a surface, obtained through any of the three classical definitions in [17, 67], correspond exactly to the definition given in the preceding section.

The Lemma on Sliding, together with the well-posedness of SIDEs inside their continuity regions, directly implies the overall well-posedness of 1-D SIDEs: for finite T , the trajectory from $t = 0$ to $t = T$ depends continuously on its initial point. As shown in Property 3.2 to follow, a SIDE reaches a steady state in finite time, which establishes its well-posedness for infinite time intervals.

²In \mathbb{R}^{p-1} , a quadrant containing a vector $\mathbf{a} = (a_1, \dots, a_{p-1})^T$ such that $a_i \neq 0$ for $i = 1, \dots, p - 1$ is the set $Q = \{\mathbf{b} \in \mathbb{R}^{p-1} : b_i a_i > 0 \text{ for } i = 1, \dots, p - 1\}$.

We call u_{n_i} , with $i \in \{2, \dots, p-1\}$ a local maximum (minimum) of the sequence $(u_{n_1}, \dots, u_{n_p})$ if $u_{n_i} > u_{n_{i\pm 1}}$ ($u_{n_i} < u_{n_{i\pm 1}}$). The point u_{n_1} is a local maximum (minimum) if $u_{n_1} > u_{n_2}$ ($u_{n_1} < u_{n_2}$); u_{n_p} is a local maximum (minimum) if $u_{n_p} > u_{n_{p-1}}$ ($u_{n_p} < u_{n_{p-1}}$). Therefore, we immediately have (as we saw in Equations (3.14), (3.15)) that the maxima (minima) are always pulled down (up):

Property 3.1 (maximum principle). *Every local maximum is decreased and every local minimum is increased by a SIDE. Therefore,*

$$|u_i(t)| < \max_n |u_n(0)| \text{ for } t > 0. \quad (3.16)$$

Using this result, we can prove the following:

Property 3.2 (finite evolution time). *A SIDE, started at $\mathbf{u}^0 = (u_1^0, \dots, u_N^0)^T$, reaches its equilibrium (i.e., the point $\mathbf{u} = (u_1, \dots, u_N)^T$ where $u_1 = \dots = u_N = \frac{1}{N} \sum_{i=1}^N u_i^0$) in finite time.*

Proof. The sum of the vertical positions of all unit-mass particles is equal to the sum of the vertical positions of the compound particles, weighted by their masses: $\sum_{n=1}^N u_n = \sum_{i=1}^p u_{n_i} m_{n_i}$. The time derivative of this quantity is zero, as verified by summing up the right-hand sides of (3.11). Therefore, the mean vertical position $\frac{1}{N} \sum_{n=1}^N u_n$ is constant throughout the evolution. Writing (3.11) for $i = 1$, $\dot{u}_{n_1} = \frac{1}{m_{n_1}} F(u_{n_2} - u_{n_1})$, we see that the leftmost compound particle is stationary only if $p = 1$, i.e., if all unit-mass particles have the same vertical position: $u_{n_1} = u_1 = u_2 = \dots = u_N$. Since the mean is conserved, the unique steady state is $u_1 = \dots = u_N = \frac{1}{N} \sum_{i=1}^N u_i^0$. To prove that it is reached in finite time, we again refer to the spring-mass model of Figure 3.1 and use the fact that a SIDE force function assigns larger force to shorter springs. If we put $L = 2 \max_n |u_n(0)|$, then the maximum principle implies that in the system there cannot exist a spring with vertical extent larger than L at any time during the evolution. Therefore, the rate of decrease of the absolute maximum, according to Equation (3.11), is at least $F(L)/N$ (because $F(L)$ is the smallest force possible in the system, and N is the largest mass). Similarly, the absolute minimum always increases at least as quickly. They will meet no later than at $t = \frac{LN}{2F(L)}$, at which point the sequence $\mathbf{u}(t)$ must be a constant sequence. ■

The above property allows us immediately to state the well-posedness results as follows:

Property 3.3 (well-posedness). *For any initial condition \mathbf{u}^{0*} , a SIDE has a unique solution $\mathbf{u}^*(t)$ satisfying $\mathbf{u}^*(0) = \mathbf{u}^{0*}$. Moreover, for any such \mathbf{u}^{0*} and any $\varepsilon > 0$, there exists a $\delta > 0$ such that $|\mathbf{u}^0 - \mathbf{u}^{0*}| \leq \delta$ implies $|\mathbf{u}(t) - \mathbf{u}^*(t)| \leq \varepsilon$ for $t \geq 0$, where $\mathbf{u}(t)$ is the solution of the SIDE with the initial condition \mathbf{u}^0 .* ■

As we pointed out in the introductory section of this chapter, a SIDE evolution defines a natural set of hitting times which intuitively should be of use in characterizing

features in an image. For this to be true, however, we would need some type of continuity of this hitting time sequence. Specifically, let $t_n(\mathbf{u}^0)$ denote the “ n -th hit time”, i.e., the time when the solution starting at \mathbf{u}^0 reaches the sliding hyperplane S_n . By Property 3.2, this is a finite number. Let $\mathbf{u}(t)$ be “a typical solution” if it never reaches two different sliding hyperplanes at the same time: $t_i(\mathbf{u}(0)) \neq t_j(\mathbf{u}(0))$ if $i \neq j$. One of the consequences of the Lemma on Sliding is that a trajectory that hits a single hyperplane S_n does so transversally (that is, cannot be tangent to it). Since trajectories vary continuously, this means that nearby solutions also hit S_n . Therefore, for typical solutions the following holds:

Property 3.4 (stability of hit times). *If $\mathbf{u}(t)$ is a typical solution, all solutions with initial data sufficiently close to $\mathbf{u}(0)$ get onto surfaces S_n in the same order as $\mathbf{u}(t)$. ■*

The sequence in which a trajectory hits surfaces S_n is an important characteristic of the solution. Property 3.4 says that, for a typical solution $\mathbf{u}(t)$, the (strict) ordering of hit times $t_n(\mathbf{u}(0))$ is stable with respect to small disturbances in $\mathbf{u}(0)$:

$$t_{n_1}(\mathbf{u}(0)) < t_{n_2}(\mathbf{u}(0)) < \dots < t_{n_{N-1}}(\mathbf{u}(0)), \quad (3.17)$$

where (n_1, \dots, n_{N-1}) is a permutation of $(1, \dots, N-1)$. For the purposes of segmentation and edge detection, the only interesting output occurs at these $N-1$ time points, since they are the only instants when the segmentation of the initial signal changes (i.e., when regions are merged and edges are erased). While a thorough investigation of how to use these hitting times and in particular how to stop a SIDE so as to obtain the best segmentation is an open one for the general form of the SIDE force function F , we will obtain a partial answer for a certain choice of F in the next chapter. Specifically, when the number of regions is an exponential random variable, and $F(v) = \text{sgn}(v)$, then the stopping rule in 1-D is given by Proposition 4.6 of the next chapter. For other SIDE force functions, the fact that the choice of the output time points is limited to a finite set provides us with both a natural sequence of segmentations of increasing granularity and with, at the very least, some simple stopping rules. For example, if the number of “useful” regions, r , is known or bounded a priori, a natural candidate for a stopping time would be $t_{n_{N-r}}$, i.e., the time when exactly r regions remain. In the next section we illustrate the effectiveness of such a rule in the simplest case, namely when $r = 2$ so that we are seeking a partition of the field of interest into two regions.

We already mentioned that our definition of solutions on sliding surfaces for SIDES in one spatial dimension coincides with all three classical definitions of solutions for a general equation with discontinuous right-hand side, which are presented on pages 50-56 of Filippov’s book [17]. We use a result on page 95 of [17] to infer the following:

Property 3.5 (continuous dependence on the RHS). *Let us consider a SIDE force function $F_S(v)$, and let $p_K(v)$ be a smoothing kernel of width K :*

$$p_K(v) \geq 0, \quad \text{supp}(p_K) = [-K; K], \quad \int p_K(v) dv = 1.$$

Let $F_K(v) = \int F_S(w)p_K(v-w)dw$ be a regularized version of $F_S(v)$. Consider system (3.4) with $m_n = 1$ and $F(v) = F_K(v)$. Then for any ε , there is a K such that the solution of this system stays closer than ε to the solution of the SIDE with the same initial condition and force $F_S(v)$. ■

We note that if the smoothing kernel $p_K(v)$ is appropriately chosen, then the resulting $F_K(v)$ will be a Perona-Malik force function of thickness K . (For example, one easy choice for $p_K(v)$ is a multiple of the indicator function of the interval $[-K;K]$.) Thus, semi-discrete Perona-Malik evolutions with small K are regularizations of SIDEs, and consequently a SIDE in 1-D can be viewed as a limiting case of a Perona-Malik-type evolution. However, as we will see in the experimental section, the SIDE evolutions appear to have some advantages over such regularized evolutions even in 1-D.

■ 3.4.2 Energy Dissipation in 1-D.

It was mentioned in the previous chapter that the SIDE (3.11) is the gradient descent equation for the global energy

$$\mathcal{E}(\mathbf{u}) = \sum_{n=1}^{N-1} E(u_{n+1} - u_n), \tag{3.18}$$

where E is the SIDE energy function (Figure 2.6), i.e., an antiderivative of the SIDE force function. Note that the standard definition of the gradient cannot be used here. Indeed, non-differentiability of E at the origin makes the directional derivatives of $\mathcal{E}(\mathbf{u})$ in the directions orthogonal to a sliding surface S undefined for $\mathbf{u} \in S$. But once $\mathbf{u}(t)$ hits a sliding surface, it stays there for all future times, and therefore we do not have to be concerned with the partial derivatives of $\mathcal{E}(\mathbf{u})$ in the directions which do not lie in the sliding surface. This leads to the definition of the gradient as the vector of partial derivatives taken with respect to the directions which belong to the sliding surface.

Definition 3.1. Suppose that S is the intersection of all the sliding hyperplanes of the SIDE (3.11) to which the vector \mathbf{u} belongs. Suppose further that $\{\mathbf{f}_i\}_{i=1}^p$ is an orthonormal basis for S . Then the gradient of \mathcal{E} with respect to S , $\nabla_S \mathcal{E}$, is defined as the weighted sum of the basis vectors, with the weights equal to the corresponding directional derivatives:

$$\nabla_S \mathcal{E}(\mathbf{u}) \stackrel{\text{def}}{=} \sum_{i=1}^p \frac{\partial \mathcal{E}(\mathbf{u})}{\partial \mathbf{f}_i} \mathbf{f}_i. \tag{3.19}$$

We will show in this section that at any moment t , the RHS of the SIDE (3.11) is the negative gradient of $\mathcal{E}(\mathbf{u}(t))$, taken with respect to the intersection S of all the sliding surfaces to which $\mathbf{u}(t)$ belongs. An auxiliary result is needed in order to show this.

Lemma 3.2. *Suppose that, as in Equation (3.11), \mathbf{u} is a signal with p distinct regions of masses m_1, \dots, m_p :*

$$\begin{aligned} 1 = n_1 < n_2, \dots < n_{p-1} < n_p \leq N & \quad \text{are such that} \\ n_{i+1} = n_i + m_{n_i}, & \quad \text{and} \\ u_{n_{i-1}} \neq u_{n_i} = u_{n_{i+1}} = \dots = u_{n_{i+1}-1} \neq u_{n_{i+1}}, & \quad \text{for } i = 1, \dots, p. \end{aligned} \quad (3.20)$$

Let $\{\mathbf{e}_j\}_{j=1}^N$ be the standard basis of \mathbb{R}^N (i.e., the j -th entry of \mathbf{e}_j is 1 and all other entries are zeros), and define

$$\mathbf{f}_i = \frac{1}{\sqrt{m_{n_i}}} \sum_{j=n_i}^{n_{i+1}-1} \mathbf{e}_j, \quad \text{for } i = 1, \dots, p. \quad (3.21)$$

Then $\{\mathbf{f}_i\}_{i=1}^p$ is an orthonormal basis for the sliding surface S defined by (3.20).

Proof. The vector \mathbf{f}_i satisfies Equation (3.20), and therefore it belongs to the sliding surface S . Since \mathbf{e}_j 's are mutually orthogonal, so are \mathbf{f}_i 's. Since there are p distinct \mathbf{f}_i 's, they form a basis for the p -dimensional surface S . The norm of \mathbf{f}_i is

$$\sum_{j=n_i}^{n_{i+1}-1} \left(\frac{1}{\sqrt{m_{n_i}}} \right)^2 = m_{n_i} \frac{1}{m_{n_i}} = 1. \quad \blacksquare$$

Property 3.6 (gradient descent). *The SIDE (3.11) is the gradient descent equation for the global energy (3.18), i.e.,*

$$\dot{\mathbf{u}}(t) = -\nabla_{S(t)} \mathcal{E}(\mathbf{u}(t)), \quad (3.22)$$

where $S(t)$ is the intersection of all sliding hyperplanes to which $\mathbf{u}(t)$ belongs, and $\nabla_{S(t)}$ is the gradient with respect to $S(t)$.

Proof. In order to prove this property, we write out Equation (3.22) in terms of the coefficients of $\dot{\mathbf{u}}$ and $-\nabla_S \mathcal{E}(\mathbf{u})$ with respect to the basis $\{\mathbf{f}_i\}_{i=1}^p$ (3.21). It is immediate from the definition (3.21) of \mathbf{f}_i 's that

$$\mathbf{u} = \sum_{i=1}^p u_{n_i} \sqrt{m_{n_i}} \mathbf{f}_i,$$

and so the i -th coefficient of $\dot{\mathbf{u}}$ in the basis $\{\mathbf{f}_i\}_{i=1}^p$ is

$$\sqrt{m_{n_i}} \dot{u}_{n_i} \quad (3.23)$$

Since the basis $\{\mathbf{f}_i\}_{i=1}^p$ is orthonormal, the i -th coefficient of $-\nabla_S \mathcal{E}(\mathbf{u})$ in this basis is the directional derivative of $-\mathcal{E}$ in the direction \mathbf{f}_i :

$$-\frac{\partial \mathcal{E}}{\partial \mathbf{f}_i} = -\lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \{ \mathcal{E}(\mathbf{u} + \mathbf{f}_i \Delta) - \mathcal{E}(\mathbf{u}) \}$$

$$\begin{aligned}
 &= -\lim_{\Delta \rightarrow 0} \frac{1}{\Delta} \left\{ \left[\sum_{n=1}^{n_i-2} E(u_{n+1} - u_n) + E\left(u_{n_i} + \frac{\Delta}{\sqrt{m_{n_i}}} - u_{n_i-1}\right) \right. \right. \\
 &\quad \left. \left. + \sum_{n=n_i}^{n_{i+1}-2} E\left(u_{n+1} + \frac{\Delta}{\sqrt{m_{n_i}}} - u_n - \frac{\Delta}{\sqrt{m_{n_i}}}\right) \right. \right. \\
 &\quad \left. \left. + E\left(u_{n_{i+1}} - u_{n_{i+1}-1} - \frac{\Delta}{\sqrt{m_{n_i}}}\right) + \sum_{n=n_{i+1}}^{N-1} E(u_{n+1} - u_n) \right] - \sum_{n=1}^{N-1} E(u_{n+1} - u_n) \right\} \\
 &= -\lim_{\Delta \rightarrow 0} \left\{ \frac{1}{\Delta} \left[E\left(u_{n_i} - u_{n_i-1} + \frac{\Delta}{\sqrt{m_{n_i}}}\right) - E(u_{n_i} - u_{n_i-1}) \right] \right. \\
 &\quad \left. + \frac{1}{\Delta} \left[E\left(u_{n_{i+1}} - u_{n_{i+1}-1} - \frac{\Delta}{\sqrt{m_{n_i}}}\right) - E(u_{n_{i+1}} - u_{n_{i+1}-1}) \right] \right\} \\
 &= -\left\{ \frac{1}{\sqrt{m_{n_i}}} E'(u_{n_i} - u_{n_i-1}) - \frac{1}{\sqrt{m_{n_i}}} E'(u_{n_{i+1}} - u_{n_{i+1}-1}) \right\} \\
 &= \frac{1}{\sqrt{m_{n_i}}} (F(u_{n_{i+1}} - u_{n_{i+1}-1}) - F(u_{n_i} - u_{n_i-1})). \tag{3.24}
 \end{aligned}$$

Equating the coefficients (3.23) and (3.24), we get that the gradient descent equation (3.22), written in the basis $\{\mathbf{f}_i\}_{i=1}^p$, is:

$$\dot{u}_{n_i} = \frac{1}{m_{n_i}} (F(u_{n_{i+1}} - u_{n_{i+1}-1}) - F(u_{n_i} - u_{n_i-1})),$$

which is the SIDE (3.11). ■

It is possible to characterize further the process of energy dissipation during the evolution of a SIDE. Namely, between any two consecutive mergings (i.e., hits of a sliding surface), the energy is a concave function of time.

Property 3.7 (energy dissipation). *Consider the SIDE (3.11) and let E be the corresponding SIDE energy function: $E' = F$. Then between any two consecutive mergings during the evolution of the SIDE, the global energy (3.18) is decreasing and concave as a function of time:*

$$\begin{aligned}
 \dot{\mathcal{E}} &< 0 \\
 \ddot{\mathcal{E}} &\leq 0.
 \end{aligned}$$

Proof. To simplify notation, we will denote

$$y_i = u_{n_i} \text{ for } i = 1, \dots, p,$$

and will simply write m_i instead of m_{n_i} . Then the global energy (3.18) is

$$\mathcal{E} = \sum_{i=1}^p E(y_{i+1} - y_i),$$

and therefore the SIDE (3.11) can be re-written as follows:

$$\dot{y}_i = -\frac{1}{m_i} \frac{\partial \mathcal{E}}{\partial y_i}, \quad i = 1, \dots, p.$$

By the chain rule of differentiation, we have:

$$\dot{\mathcal{E}} = \sum_{i=1}^p \frac{\partial \mathcal{E}}{\partial y_i} \dot{y}_i = -\sum_{i=1}^p \left(\frac{\partial \mathcal{E}}{\partial y_i} \right)^2 \frac{1}{m_i} < 0.$$

Differentiating with respect to t one more time and applying the chain rule again yields:

$$\begin{aligned} \ddot{\mathcal{E}} &= -\sum_{i=1}^p 2 \frac{\partial \mathcal{E}}{\partial y_i} \frac{d}{dt} \left(\frac{\partial \mathcal{E}}{\partial y_i} \right) \frac{1}{m_i} \\ &= -\sum_{i=1}^p 2 \frac{\partial \mathcal{E}}{\partial y_i} \left(\sum_{k=1}^p \frac{\partial^2 \mathcal{E}}{\partial y_i \partial y_k} \dot{y}_k \right) \frac{1}{m_i} \\ &= \sum_{i=1}^p 2 \frac{\partial \mathcal{E}}{\partial y_i} \left(\sum_{k=1}^p \frac{\partial^2 \mathcal{E}}{\partial y_i \partial y_k} \frac{1}{m_k} \frac{\partial \mathcal{E}}{\partial y_k} \right) \frac{1}{m_i} \\ &= 2D^T H D, \end{aligned} \tag{3.25}$$

where

$$D = \left(\frac{1}{m_1} \frac{\partial \mathcal{E}}{\partial y_1}, \dots, \frac{1}{m_p} \frac{\partial \mathcal{E}}{\partial y_p} \right)^T,$$

and H is the Hessian matrix of \mathcal{E} , i.e., the matrix of all the mixed second derivatives of \mathcal{E} . The entry in the i -th row and k -th column of H is $\frac{\partial^2 \mathcal{E}}{\partial y_i \partial y_k}$. In other words,

$$H = - \begin{pmatrix} x_1 & -x_1 & 0 & 0 & 0 & \dots & 0 \\ -x_1 & x_1 + x_2 & -x_2 & 0 & 0 & \dots & 0 \\ 0 & -x_2 & x_2 + x_3 & -x_3 & 0 & \dots & 0 \\ \vdots & & & & & & \vdots \\ & & & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & & 0 & -x_{p-2} & x_{p-2} + x_{p-1} & -x_{p-1} \\ 0 & \dots & \dots & & 0 & -x_{p-1} & x_{p-1} \end{pmatrix},$$

where $x_i = -F'(y_{i+1} - y_i)$. Note that, by our definition of y_i 's, $y_{i+1} - y_i \neq 0$, and that $F(y_{i+1} - y_i)$ is monotonically decreasing for $y_{i+1} - y_i \neq 0$. Therefore, $x_i > 0$.

All that remains to show is that H is negative semidefinite, which, combined with (3.25), means that $\ddot{\mathcal{E}} \leq 0$. It is easily verified that $-H$ can be factorized into a lower-

triangular and an upper-triangular matrix as follows:

$$-H = \begin{pmatrix} 1 & 0 & \dots & 0 \\ -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & & & \vdots & \\ & & \ddots & \ddots & \ddots & 0 \\ 0 & \dots & -1 & 1 & 0 \\ 0 & \dots & 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} x_1 & -x_1 & 0 & 0 & \dots & 0 \\ 0 & x_2 & -x_2 & 0 & \dots & 0 \\ 0 & 0 & x_3 & -x_3 & \dots & 0 \\ \vdots & & & & & \vdots \\ & & & & \ddots & \ddots & 0 \\ 0 & \dots & & & x_{p-1} & -x_{p-1} \\ 0 & \dots & & & & 0 \end{pmatrix}.$$

The diagonal entries $x_1, \dots, x_{p-1}, 0$ of the upper-triangular matrix are the pivots ([62], page 32) of $-H$. Since all the pivots are nonnegative, it follows ([62], page 339) that $-H \geq 0 \Rightarrow H \leq 0$, which implies $\ddot{\mathcal{E}} \leq 0$. ■

A typical picture of the energy dissipation is shown in Figure 3.7; the only points where \mathcal{E} might not be concave as a function of time are the merge points.

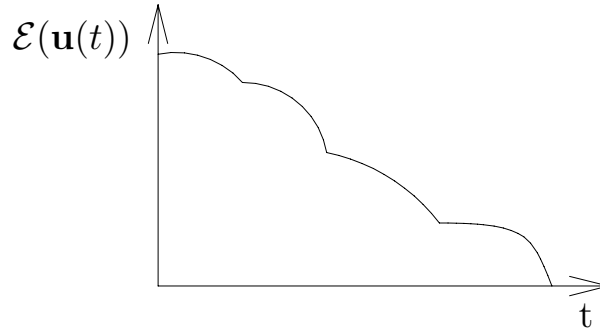


Figure 3.7. Typical picture of the energy dissipation during the evolution of a SIDE.

In addition to being the gradient descent equation for the global energy (3.18), with $E' = F$, we now show that the SIDE (3.11) also reduces $\sum_{n=1}^N E_1(u_{n+1} - u_n)$ if $E_1 \neq E$ is any SIDE energy function.

Property 3.8 (Lyapunov functionals). *Consider the SIDE (3.11), and let E be the corresponding SIDE energy function: $E' = F$. Let E_1 be an arbitrary SIDE energy function (i.e., such function that E_1' satisfies (3.3),(3.9)), and define*

$$\mathcal{E}_1(\mathbf{u}) = \sum_{n=1}^N E_1(u_{n+1} - u_n).$$

Then \mathcal{E}_1 is a Lyapunov functional of the SIDE. In other words, $\mathcal{E}_1(\mathbf{u}(t))$ is a decreasing function of time, until the steady state is reached.

Proof. We again use the notation from the proof of the previous property:

$$\begin{aligned} y_i &= u_{n_i} \text{ for } i = 1, \dots, p; \\ \mathcal{E}_1 &= \sum_{i=1}^p E_1(y_{i+1} - y_i). \end{aligned}$$

Then, by the chain rule,

$$\begin{aligned} \dot{\mathcal{E}} &= \sum_{i=1}^p \frac{\partial \mathcal{E}}{\partial y_i} \dot{y}_i \\ &= -E'_1(y_2 - y_1)\dot{y}_1 + \sum_{i=2}^{p-1} (E'_1(y_i - y_{i-1}) - E'_1(y_{i+1} - y_i))\dot{y}_i + E'_1(y_p - y_{p-1})\dot{y}_p \\ &= - \left[\frac{1}{m_1} E'_1(y_2 - y_1) F(y_2 - y_1) \right. \\ &\quad + \sum_{i=2}^{p-1} \frac{1}{m_i} (E'_1(y_i - y_{i-1}) - E'_1(y_{i+1} - y_i)) (F(y_i - y_{i-1}) - F(y_{i+1} - y_i)) \\ &\quad \left. + \frac{1}{m_p} E'_1(y_p - y_{p-1}) F(y_p - y_{p-1}) \right]. \end{aligned}$$

The first term inside the brackets is positive, since $E'_1(y_2 - y_1)$, $F(y_2 - y_1)$, and $y_2 - y_1$ all have the same sign. Similarly, the last term is positive. Each term in the summation is also positive, because of monotonicity of E'_1 and F . Therefore, $\dot{\mathcal{E}} < 0$. ■

We now analyze another class of Lyapunov functionals, which includes the ℓ^2 norm and negative entropy.

Property 3.9 (Lyapunov functionals, continued). *Suppose that $R : \mathbb{R} \rightarrow \mathbb{R}$ is a function such that its derivative R' is monotonically increasing. Define*

$$\mathcal{R}(\mathbf{u}) = \sum_{n=1}^N R(u_n).$$

Then \mathcal{R} is a Lyapunov functional of Equation (3.4), i.e.,

$$\dot{\mathcal{R}} < 0,$$

until the steady state is reached. In particular, \mathcal{R} is a Lyapunov functional of SDEs.

Proof. Using the notation of the previous proof,

$$\dot{\mathcal{R}} = \frac{d}{dt} \left(\sum_{i=1}^p m_i R(y_i) \right)$$

$$\begin{aligned}
&= \sum_{i=1}^p m_i R'(y_i) \dot{y}_i \\
&= R'(y_1)F(y_2 - y_1) + \sum_{i=2}^{p-1} R'(y_i)(F(y_{i+1} - y_i) - F(y_i - y_{i-1})) - R'(y_p)F(y_p - y_{p-1}) \\
&= \sum_{i=1}^{p-1} (R'(y_i) - R'(y_{i+1}))F(y_{i+1} - y_i). \tag{3.26}
\end{aligned}$$

Since R' is monotonically increasing,

$$\operatorname{sgn}(R'(y_i) - R'(y_{i+1})) = \operatorname{sgn}(y_i - y_{i+1}) = -\operatorname{sgn}(y_{i+1} - y_i);$$

since F is a force function,

$$\operatorname{sgn}(F(y_{i+1} - y_i)) = \operatorname{sgn}(y_{i+1} - y_i).$$

Therefore, the product $(R'(y_i) - R'(y_{i+1}))F(y_{i+1} - y_i)$ is negative, the sum (3.26) is negative, and so $\dot{\mathcal{R}} < 0$. ■

Example 3.3. ℓ^q norms.

Suppose $R(u_n) = |u_n|^q$ with $q > 1$. Then $R'(u_n) = \operatorname{sgn}(u_n)q|u_n|^{q-1}$ is monotonically increasing, which means that the ℓ^q norm

$$(\mathcal{R}(\mathbf{u}))^{\frac{1}{q}} = \left(\sum_{i=1}^q |u_n|^q \right)^{\frac{1}{q}}$$

is a Lyapunov functional of Equation (3.4). This is yet another characterization of the stability of the equations described by the mechanical model of Section 3.2. ■

Example 3.4. Moments.

It is similarly shown that the even central moments of \mathbf{u} are also Lyapunov functionals:

$$\frac{1}{N} \sum_{n=1}^N \left(u_n - \frac{1}{N} \sum_{i=1}^N u_i \right)^{2k}, \quad k = 1, 2, \dots \quad \blacksquare$$

Example 3.5. Entropy.

Suppose the initial condition \mathbf{u}^0 is positive:

$$\min_{1 \leq n \leq N} u_n^0 > 0,$$

and define $R(u_n) = u_n \ln u_n$. Then $R'(u_n) = \ln u_n + 1$ is monotonically increasing, and so the negative entropy $\mathcal{R}(\mathbf{u}) = \sum_{n=1}^N u_n \ln u_n$ is a Lyapunov functional. The fact that the entropy is increased by SIDs and other evolution equations of the form (3.4) is in agreement with the intuitive notion that, as scale increases, the signal is simplified: at coarser scales, the information content is reduced. ■

■ 3.4.3 Properties in 2-D.

The existence and uniqueness of solutions in 2-D again follow easily from our construction of solutions. Property 3.1 (the maximum principle) is easily inferred from the 2-D spring-mass model. (A local maximum (minimum) is a region of a 2-D image whose value is larger (smaller) than the values of its neighbors. Re-phrasing this definition in terms of our spring-mass model, a maximum (minimum) is a particle with all its attached springs directed downward (upward).) Property 3.2 (finite evolution time) also carries over, with the same proof.

There is, however, no analog of the Lemma on Sliding in 2-D for SIDE force functions such as that depicted in Figure 3.4: it is easy to show that the solutions in the vicinity of a discontinuity hyperplane of (3.12) do not necessarily slide onto that hyperplane. Notice, however, that forcing two neighboring regions with equal intensities to merge is conceptually very similar to using a modified force function which is infinite at zero, as depicted in Figure 3.8. Indeed, the fact that $F(0^\pm) = \pm\infty$ means that if the vertical distance between two particles is very small compared to the distances to their other neighbors, they will be driven towards each other and will be merged. Thus, the Lemma on Sliding holds for the force function of Figure 3.8, from which the global continuous dependence on the initial data is again inferred. We do not use this force function in simulations, since its large values near zero present problems with numerical integration of the corresponding equation. What we do use is a SIDE force function, in

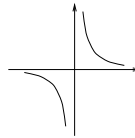


Figure 3.8. A modified force function, for which sliding happens in 2-D, as well as in 1-D.

conjunction with Equation (3.12). Since sliding modes do not necessarily occur on the discontinuity hyperplanes, there is no global continuous dependence on the initial data. In particular, the sequence of hitting times and associated discontinuity planes does not depend continuously on initial conditions, and our SIDE evolution does not correspond to a limiting form of a Perona-Malik evolution in 2-D but in fact represents a decidedly different type of evolutionary behavior. Several factors, however, indicate the value of this new evolution and also suggest that a weaker stability result can be proven. First of all, as shown in the experimental results in the next section, SIDEs can produce excellent segmentations in 2-D images even in the presence of considerable noise. Moreover, thanks to the maximum principle, excessively wild behavior of solutions is impossible, something that is again confirmed by the experiments of the next section. Consequently, the sequence of hit times (3.17) does not seem to be very sensitive to the initial condition in that the presence of noise, while perhaps perturbing the ordering of hitting times and the sliding planes that are hit, seems to introduce perturbations that are, in some sense, “small”.

Finally, we note without giving details that the properties on energy dissipation (3.6 and 3.7) and Property 3.9 on Lyapunov functionals, carry over to 2-D, as well as their proofs—with slight changes to accommodate the fact that a region may have more than two neighbors in 2-D.

■ 3.5 Experiments.

In this section we present examples in both 1-D and 2-D. The purpose of 1-D experiments is to provide the basic intuition for how SIDes work, as well as to contrast SIDes with the methods reviewed in the previous chapter. We do not claim that SIDes are the best for any of these 1-D examples, for which good results can be efficiently obtained using simple algorithms. In 2-D, however, this is no longer true, and SIDes have considerable advantages over the existing methods.

Choosing a SIDE force function best suited for a particular application is an open research question. (It is partly addressed in Chapter 4, by describing the problems for which $F(v) = \text{sgn}(v)$ is the best choice.) For the examples below, we use a very simple, piecewise-linear force function $F(v) = \text{sgn}(v) - \frac{v}{L}$, depicted in Figure 3.9. Note that,

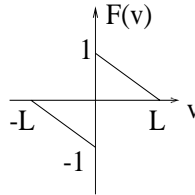


Figure 3.9. The SIDE force function used in the experimental section.

formally, this function does not satisfy our definition (3.3) of a force function, since it is negative for $v > L$. Therefore, in our experiments we always make sure that L is larger than the dynamic range of the signal or image to be processed. In that case, thanks to the maximum principle, we will have $|u_i(t) - u_j(t)| < L$ for any pair of pixels at any time t during evolution, and therefore $F(|u_i(t) - u_j(t)|) > 0$.

As we mentioned before, choosing the appropriate stopping rule is also an open problem. In the examples to follow, we assume that we know the number of regions we are looking for, and stop the evolution when that number of regions is achieved.

■ 3.5.1 Experiment 1: 1-D Unit Step in High Noise Environment.

We first test this SIDE on a unit step function corrupted by additive white Gaussian noise whose standard deviation is equal to the amplitude of the step, and which is depicted in Figure 3.10(a). The remaining parts of this figure display snapshots of the SIDE evolution starting with the noisy data in Figure 3.10(a), i.e., they correspond to the evolution at a selected set of hitting times. The particular members of the scale space which are illustrated are labeled according to the number of remaining regions.

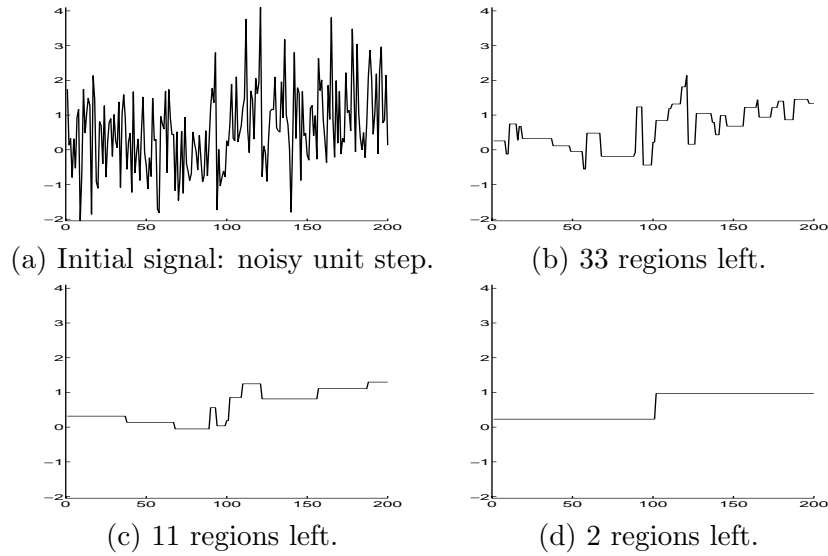


Figure 3.10. Scale space of a SIDE for a noisy unit step at location 100: (a) the original signal; (b)–(d) representatives of the resulting SIDE scale space.

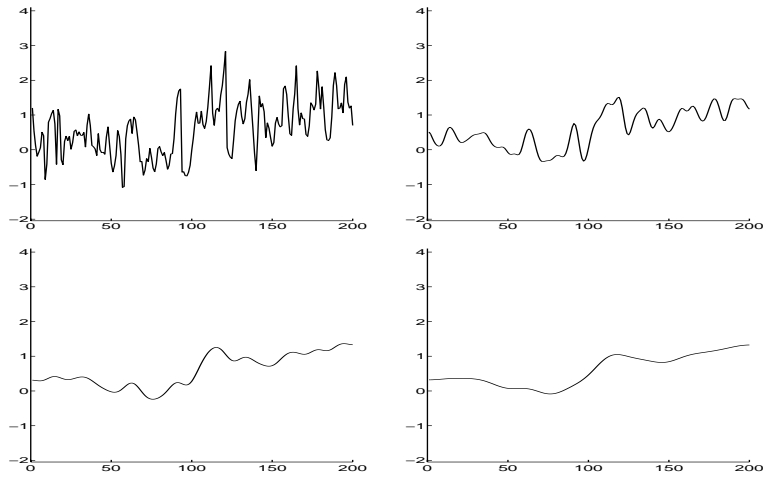


Figure 3.11. Scale space of a Perona-Malik equation with a large K for the noisy step of Figure 3.10.

Note that the last remaining edge, i.e., the edge in Figure 3.10(d) for the hitting time at which there are only two regions left, is located between samples 101 and 102, which is quite close to the position of the original edge (between the 100-th and 101-st samples). In this example, the step in Figure 3.10(d) also has amplitude that is close to that of the original unit step. In general, thanks to the stability of SIDes, the sizes of discontinuities will be diminished through such an evolution, much as they are in other evolution equations. However, from the perspective of segmentation this is irrelevant—

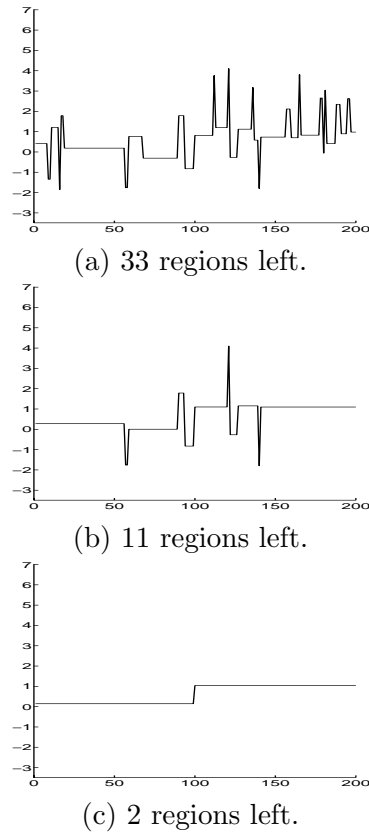


Figure 3.12. Scale space of the region merging algorithm of Koepfler, Lopez, and Morel for the noisy unit step signal of Figure 3.10(a).

i.e., the focus of attention is on detecting and locating the edge, not on estimating its amplitude.

This example also provides us with the opportunity to contrast the behavior of a SIDE evolution with a Perona-Malik evolution and in fact to describe the behavior that originally motivated our work. Specifically, as we noted in the discussion of Property 3.5 of the previous section, a SIDE in 1-D can be approximated with a Perona-Malik equation of a small thickness K . Observe that a Perona-Malik equation of a large thickness K will diffuse the edge before removing all the noise. Consequently, if the objective is segmentation, the desire is to use as small a value of K as possible. Following the procedure prescribed by Perona, Shiota, and Malik in [50], we computed the histogram of the absolute values of the gradient throughout the initial signal, and fixed K at 90% of its integral. The resulting evolution is shown in Figure 3.11. In addition to its good denoising performance, it also blurs the edge, which is clearly undesirable if the objective is a sharp segmentation. The comparison of Figures 3.10 and 3.11 strongly suggests that the smaller K the better. It was precisely this observation that originally

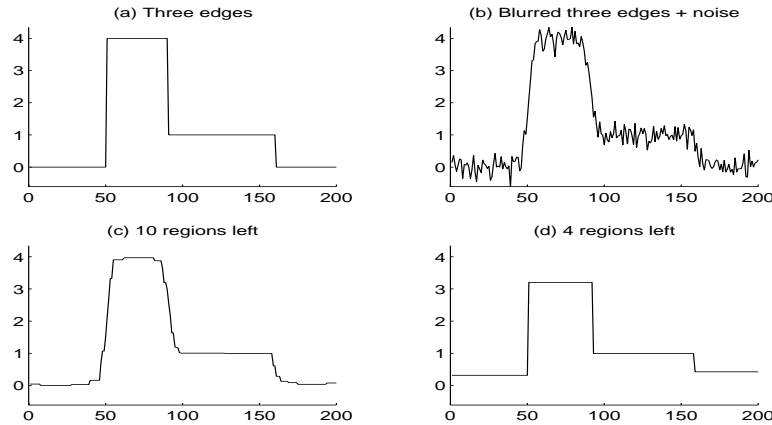


Figure 3.13. Scale space of a SIDE for a noisy blurred 3-edge staircase: (a) noise-free original signal; (b) its blurred version with additive noise; (c),(d) representatives of the resulting SIDE scale space.

motivated the development of SIDEs. However, while in 1-D a SIDE evolution can be viewed precisely as a limit of a Perona-Malik evolution as K goes to 0, there is still an advantage to using the form of the evolution that we have described rather than a Perona-Malik evolution with a very small value of K . Specifically, the presence of explicit reductions in dimensionality during the evolution makes a SIDE implementation more efficient than that described in [50]. Even for this simple example the Perona-Malik evolution that produced the result comparable to that in Figure 3.10 evolved approximately 5 times more slowly than our SIDE evolution. (Both were implemented via forward Euler discretization schemes [14] in MATLAB.) Although a SIDE in 2-D cannot be viewed as a limit of Perona-Malik evolutions, the same comparison in speed of evolution is still true, although in this case the difference in computation time can be orders of magnitude.

In this example, the region merging method of Koepfler, Lopez and Morel [36] works quite well (see Figure 3.12). We will soon see, however, that it is not as robust as SIDEs: its performance worsens dramatically when signals are corrupted with a heavy-tailed noise.

■ 3.5.2 Experiment 2: Edge Enhancement in 1-D.

Our second one-dimensional example shows that SIDEs can stably enhance edges. The staircase signal in the upper left-hand corner of Figure 3.13 was convolved with a Gaussian and corrupted by additive noise. The evolution was stopped when there were only four regions (three edges) left. The locations of the edges are very close to those in the original signal. (Note that the amplitudes of the final signal are quite different from those of the initial condition. This is immaterial, since we are interested in segmentation, not in restoration.)

■ 3.5.3 Experiment 3: Robustness in 1-D.

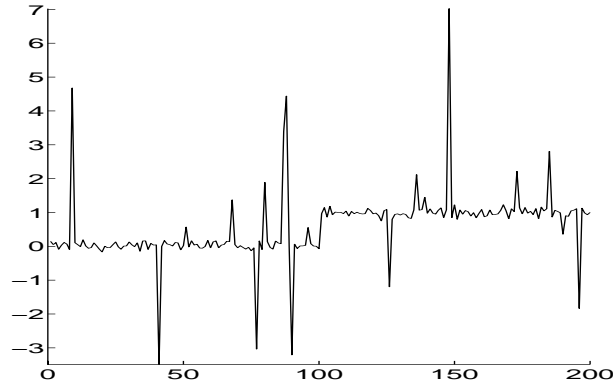


Figure 3.14. A unit step with heavy-tailed noise.

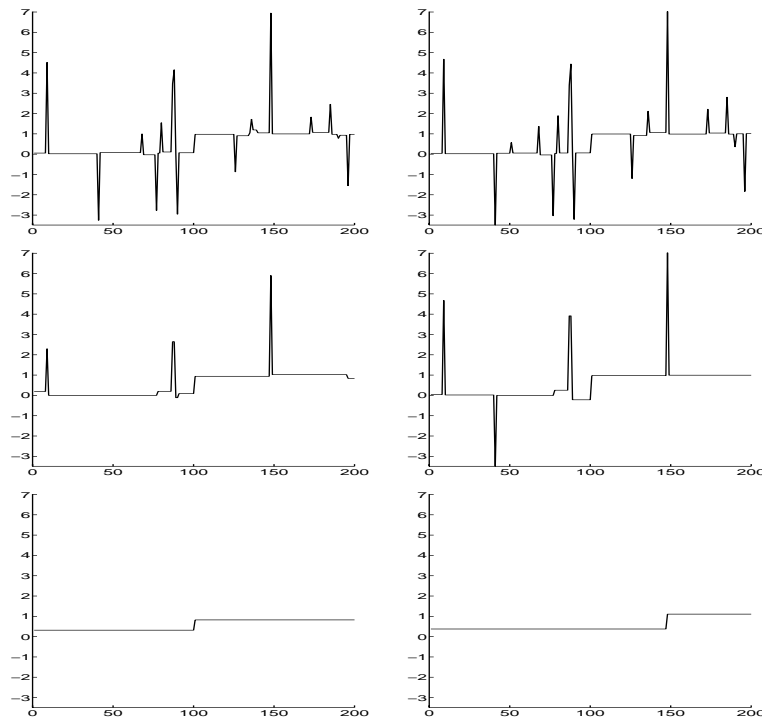


Figure 3.15. Scale spaces for the signal of Figure 3.14: SIDE (left) and Koepfler-Lopez-Morel (right). Top: 33 regions; middle: 11 regions; bottom: 2 regions.

We now compare the robustness of our algorithm to Koepfler, Lopez, and Morel’s [36] region merging minimization of the Mumford-Shah functional [44]. For that purpose, we use Monte-Carlo simulations on a unit step signal corrupted by “heavy-tailed”

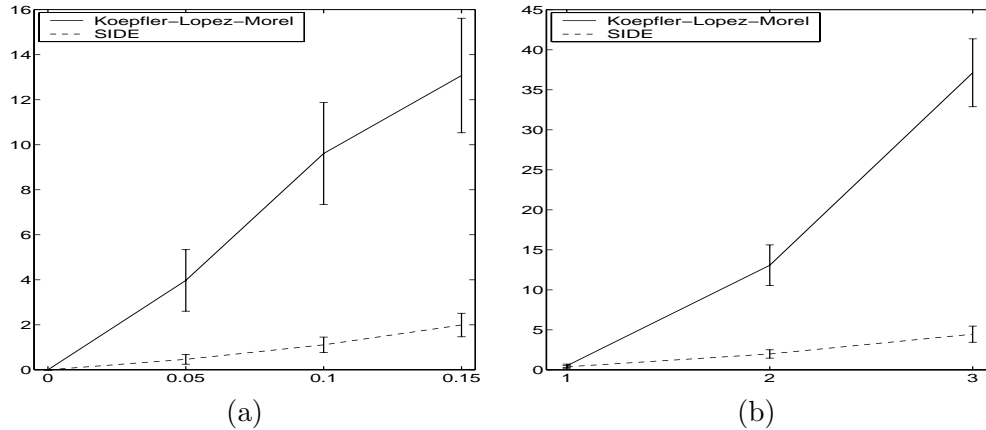


Figure 3.16. Mean absolute errors for Monte-Carlo runs. (Koepfler-Lopez-Morel: solid line; SIDE: broken line.) The error bars are \pm two standard deviations. (a) Different contamination probabilities (0, 0.05, 0.1, and 0.15); contaminating standard deviation is fixed at 2. (b) Contamination probability is fixed at 0.15; different contaminating standard deviations (1, 2, and 3).

noise which is, with high probability $1 - \varepsilon$, normally distributed with $\sigma_1 = 0.1$, and, with low probability ε , normally distributed with a larger standard deviation σ_2 . A typical sample path, for $\varepsilon = 0.1$ and $\sigma_2 = 2$, is shown in Figure 3.14. The SIDE and Koepfler-Lopez-Morel scale spaces for this signal are illustrated in Figure 3.15. During every Monte-Carlo trial, each algorithm was stopped when only two regions remained, and the resulting jump location was taken as the output. When $\sigma_2 = 2$, the mean absolute errors in locating the jump for $\varepsilon = 0$, $\varepsilon = 0.05$, $\varepsilon = 0.1$, and $\varepsilon = 0.15$ are shown in Figure 3.16(a) (the solid line is Koepfler-Lopez-Morel, the broken line is SIDE). The error bars are \pm two standard deviations. Figure 3.16(b) shows the mean absolute errors for different standard deviations σ_2 of the contaminating Gaussian, when ε is fixed at 0.15.

As we anticipated in Chapter 2 and will further discuss in the next section, the quadratic term of the Mumford-Shah energy makes it non-robust to heavy-tailed noise, and the performance degrades considerably as the contamination probability and the variance of the contaminating Gaussian increase. Note that when $\sigma_2 = 3$ and $\varepsilon = 0.15$, using the Koepfler-Lopez-Morel algorithm is not significantly better than guessing the edge location as a random number between 1 and 200. At the same time, our method is very robust, even if the outlier probability is as high as 0.15.

Figure 3.17 shows the scale space generated by a Perona-Malik equation for the step signal with heavy-tailed noise depicted in Figure 3.14. As in Experiment 1, K was fixed at 90% of the histogram of the gradient, in accordance with Perona, Shiota, and Malik [50]. As before, its de-noising performance is good; however, it also introduces blurring and therefore its output does not immediately provide a segmentation. In order to get a good segmentation from this procedure, one needs to devise a stopping rule, so as to stop the evolution at a scale when noise spikes are diffused but the step

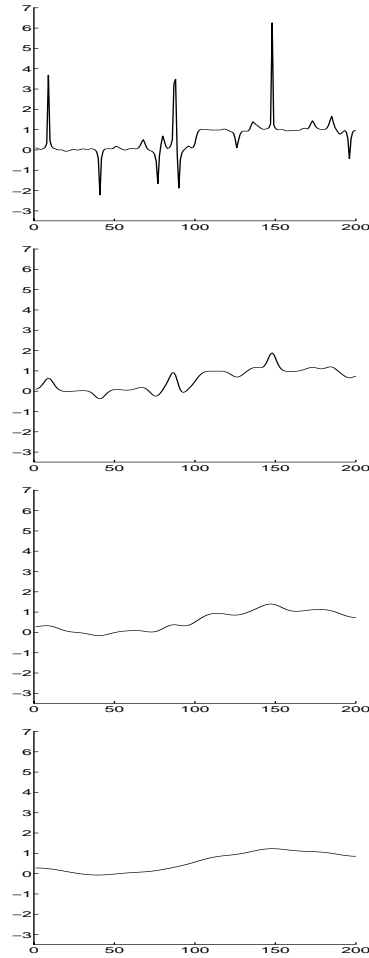


Figure 3.17. Scale space of a Perona-Malik equation with large K for the signal of Figure 3.14.

is not completely diffused (such as in the second plot of Figure 3.14). In addition, one needs to use an edge detector in order to extract the edge from the signal at that scale.

We again emphasize that neither SIDes, nor the Koepfler-Lopez-Morel algorithm, nor any combination of the Perona-Malik equation with a stopping rule and an edge detector, are optimal for this simple 1-D problem, for which near-perfect results can be achieved in a computationally efficient manner by very simple procedures. The purpose of including this example is to provide statistical evidence for our claim of robustness of SIDes. This becomes very important for complicated 2-D problems, such as the ones considered in the next example, where simple techniques no longer work.

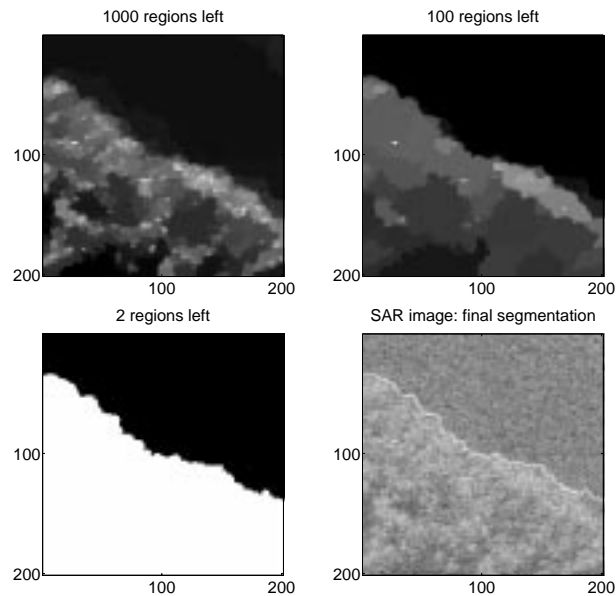


Figure 3.18. Scale space of a SIDE for the SAR image of trees and grass, and the final boundary superimposed on the initial image.

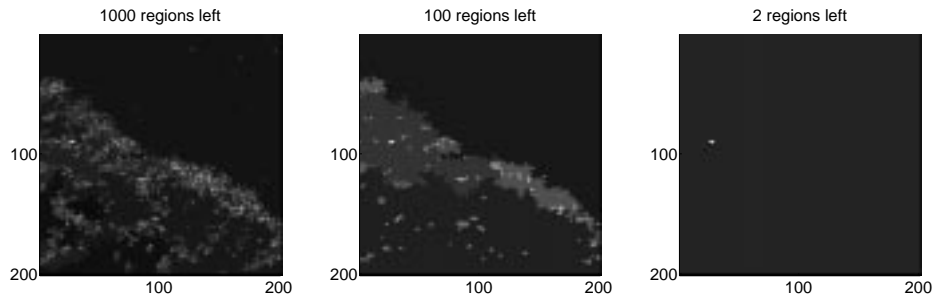


Figure 3.19. Segmentations of the SAR image via the region merging method of Koepfler, Lopez, and Morel.

■ 3.5.4 Experiment 4: SIDE Evolutions in 2-D.

Both the sharpness of boundaries and robustness of SIDEs are also evident in the image experiments we have conducted. These properties are used to advantage in segmenting the SAR image of Figure 1.1 in which only two textures are present (forest and grass). The scale space is shown in Figure 3.18 (with the intensity values of each image scaled so as to take up the whole grayscale range), as well as the resulting boundary superimposed onto the original log-magnitude image. SAR imagery, such as the example shown here, are subject to the phenomenon known as speckle, which is present in any coherent imaging system and which leads to the large amplitude variations and noise evident in the original image. Consequently, the accurate segmentation of such imagery can

be quite challenging and in particular cannot be accomplished using standard edge detection algorithms. For example, the scale space of the region merging algorithm of [36], as implemented in [20] and discussed above, is depicted in Figure 3.19. If evolved until two regions remain, it will find the boundary around a burst of noise. In contrast, the two-region segmentation displayed in Figure 3.18(d) is very accurate.

We note, that, as mentioned in Experiment 1, the SIDE evolutions require far less computation time than Perona-Malik-type evolutions. Since in 2-D a SIDE evolution is not a limiting form of a Perona-Malik evolution, the comparison is not quite as simple. However, in experiments that we have performed in which we have devised Perona-Malik evolutions that produce results as qualitatively similar to those in Figure 3.18 as possible, we have found that the resulting computational effort is roughly 130 times slower for this (201×201) image than our SIDE evolution. (As in 1-D, we used forward Euler discretization schemes [14] to implement both equations.)

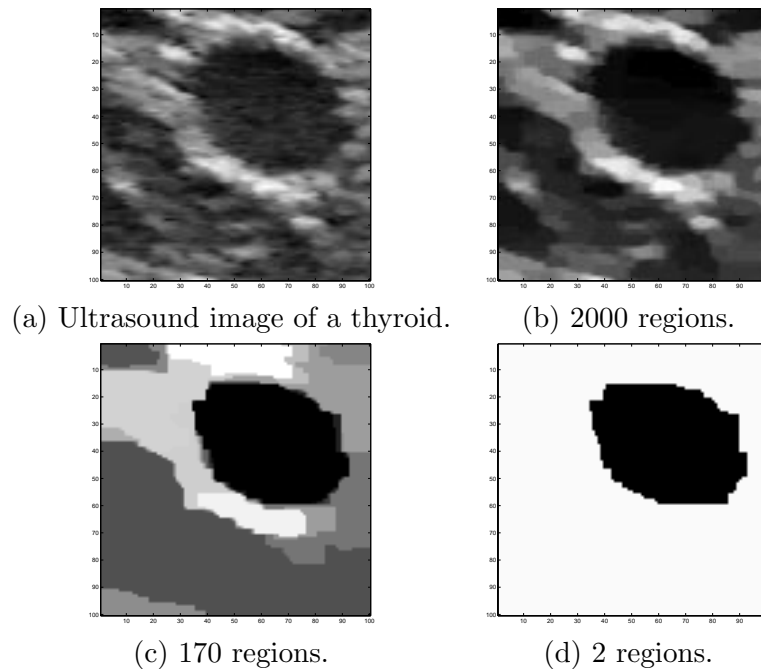


Figure 3.20. Scale space of a SIDE for the ultrasound image of a thyroid.

We close this section by presenting another segmentation example in 2-D. In the ultrasound image of a thyroid (Figure 3.20(a)), both speckle and blurring are evident. This degradation, inherent to ultrasound imaging, makes the problem of segmentation very difficult. Three images from the SIDE scale space are shown in Figure 3.20(b-d). The two-region segmentation (Figure 3.20(d)) is again very accurate.

We note that SIDEs have been used successfully in the context of another medical application in [19], namely segmentation of dermatoscopic images of skin lesions.

■ 3.6 Related Approaches.

■ 3.6.1 Mumford-Shah, Geman-Reynolds, and Zhu-Mumford.

The global energy (3.18) associated with SIDEs is similar to the first term of the image restoration model of D. Geman and Reynolds [21], as well as to Zhu and Mumford’s potential function [75]. As we showed in the experimental section, this not only leads to sharp segmentations, but also allows our method to be more robust to heavy-tailed noise than algorithms which use quadratic energy terms. An important distinction from [21] and [75] is the interpretation of SIDEs as a region merging method, which leads to a much faster numerical implementation. It is also a more *universal* method of edge sharpening, since, unlike the algorithm of Geman and Reynolds, it does not require the knowledge of the blur model.

Since SIDEs are implemented via recursive region merging, it is instructive to compare them with other recursive region merging algorithms, such as Koepfler, Lopez, and Morel’s [36] implementation of Mumford-Shah [44] segmentation, in which merging of neighboring regions occurs if it reduces the energy $\mathcal{E}_{MS}(\mathbf{u}) = (\mathbf{u} - \mathbf{u}^0)^T(\mathbf{u} - \mathbf{u}^0) + \lambda l$. (Here l is the total length of the boundaries, and λ is a scale parameter: a larger λ imposes greater penalty on boundaries, which results in a coarser segmentation \mathbf{u} .) The first term of this functional makes it non-robust to noise outliers. As we have seen, this term quadratically penalizes the difference between the initial image and its approximation \mathbf{u} , thereby causing very large outliers present in the original image \mathbf{u}^0 to re-appear in \mathbf{u} , even for large values of λ . This was clearly illustrated by the examples presented in the experimental section. The absence of such a term from the SIDE energy allows the evolution to diffuse strong bursts of noise, making it robust.

■ 3.6.2 Shock Filters and Total Variation.

Replacing the discrete vector $\mathbf{u}(t)$ with a function $u(t, x)$ of a continuous spatial variable x , and replacing first differences with derivatives in Equation (3.4), we see that, for $m_n = 1$, Equation (3.4) is a discretization of $u_t = \frac{\partial}{\partial x}[F(u_x)]$, where letter subscripts denote corresponding partial derivatives. Expanding the SIDE force function again as $F(v) = F_{id}(v) + C \operatorname{sgn}(v)$, we obtain:

$$u_t = C \frac{\partial}{\partial x}[\operatorname{sgn}(u_x)] + F'_{id}(u_x)u_{xx}. \quad (3.27)$$

The first of the RHS terms is the 1-D version of the gradient descent on total variation. It has very good noise removal properties, but, if used alone, it will ultimately blur the signal. If $F_{id}(v) = -\frac{1}{2}v|v|$, then the second term is equal to the RHS of one of the shock filters introduced by Osher and Rudin in [46]—namely, Equation (2.16) which we considered in Chapter 2. Discretizations of certain shock filters are excellent for edge enhancement, but, as we saw in Chapter 2, they cannot remove noise (and, in fact, some of them are unstable and noise-enhancing.) Thus, SIDEs combine the noise-suppressive properties of the total variation approach with the edge-sharpening features of shock

filters. It should be noted, however, that (3.27) requires careful interpretation, because its RHS contains the signum function of u_x which itself may have both singularities and segments over which it is identically zero. In addition, this strange object is differentiated with respect to x . Thus, the interesting research issue arises of *defining* what one means by a solution to (3.27), in such a manner that the definition results in solutions relevant to the desired image processing applications. This complicated problem is avoided entirely with the introduction of SIDEs, since there one starts with a semi-discrete formulation, in which the issues of the existence and uniqueness of solutions are well understood. The SIDEs are thus a logical extension of Bouman and Sauer's approach of [6, 58] in which images are discrete matrices of numbers, rather than functions of two continuous variables. As described in Chapter 2, Bouman and Sauer proposed minimizing an energy functional consisting of two terms, one of which is the discrete counterpart of the total variation. Their method of quickly computing a local minimum of this non-differentiable functional involved merging pixels and thus anticipated SIDEs.

■ 3.7 Conclusion.

In this chapter we have presented a new approach to edge enhancement and segmentation, and demonstrated its successful application to signals and images with very high levels of noise, as well as to blurry signals. Our approach is based on a new class of evolution equations for the processing of imagery and signals which we have termed stabilized inverse diffusion equations or SIDEs. These evolutions, which have discontinuous right-hand sides, have conceptual and mathematical links to other evolution-based methods in signal and image processing, but they also have their own unique qualitative characteristics and properties. The next chapter is devoted to extensive analysis of a particular version of SIDEs.