VARIATIONAL ANALYSIS, PDE'S AND IMAGE ANALYSIS: THE BIG PICTURE AND A SAMPLING OF DETAILS

N. D. GEORGE*AND K. R. VIXIE†

Los Alamos National Laboratory, MS P365
Los Alamos, NM 87545, USA
Contact E-mail: ndg@lanl.gov

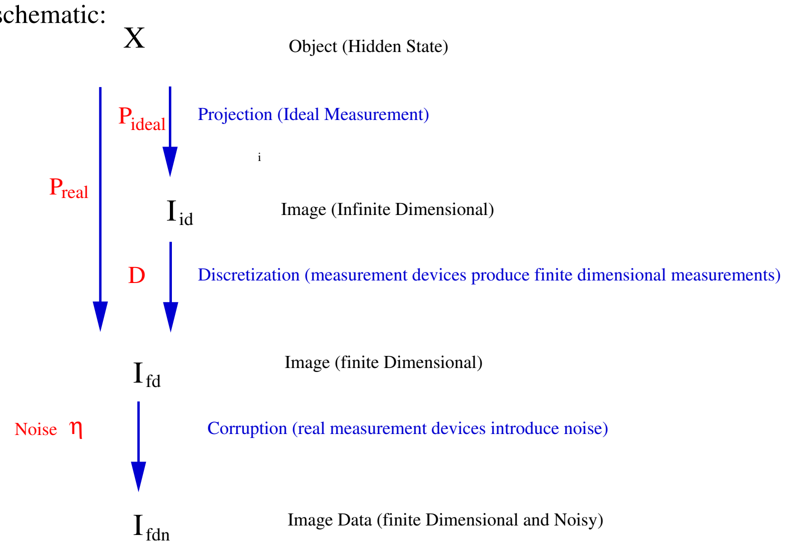
Here we summarize an invited plenary talk for the *Third International Workshop on Contemporary Problems in Mathematical Physics* in which we familiarized the community with problems of interest within image analysis and introduced researchers to some of the important tools and cutting edge developments in the analysis of image data. We explain our high level view and perspectives, show a few examples of data and techniques, and provide references for rurther study. Our goal is to give a feeling for the way a mathematician thinks about and uses abstract structures and geometric/analytic tools to understand and solve problems in the analysis of image data.

*Theoretical Division
†Applied Physics Division

1. Image Analysis: high level description of PDE/Variational approach

1.1. *Inverse Problems*

A schematic:



We will focus on the noise removal and the reconstruction of hidden information by the use of prior information about the object and/or the infinitely resolved image. Therefore, we will be interested in the inversion of one of the following three equations.

$$
\begin{aligned}
I_{fdn} &= P_{real}(x) + \eta \\
I_{fdn} &= D(I_{id}) + \eta \\
I_{fdn} &= I_{fd} + \eta
\end{aligned}
$$

We encounter several difficulties with this problem:

- **Nontrivial null space:** The operator D or P_{real} loses information i.e. it has nontrivial level sets: $D^{-1}(I_{fd})$ and $P_{real}^{-1}(I_{fd})$ are sets, not points.
- **Unknown noise distribution:** Often, the stochastic perturbations η are not well characterized.
- **Systematic Effects** Systematic effects which we have included in D and P_{real} above, can be quite difficult to characterize.

For the rest of the paper, we simplify the above notation as follows.

- **d** is the measured image data (I_{fdn} on the previous page)
- **u** is the object, reconstructed image or denoised image ... (x, I_{id}, I_{fd} previously).

- **P** is any measurement or projection operator

With this notation, we will be mostly concerned with equations of the type:

$$\min_u F(u) = \min_u ||u - d|| + R(u)$$
$$\min_u F(u) = \min_u ||Pu - d|| + R(u)$$

where $R(u)$ is a regularization term in which we implicitly insert prior information about the object and image. (Detail: we often start by considering both d and u to be elements in the inifinite dimensional image space).

We emphasize that $R(u)$ is where the prior knowledge is incorporated. Note that by differentiating (Gateux or Frechet) F, we can turn a minimization problem into a PDE:

$$\min_u F(u) \Rightarrow u_t = -DF(u)$$

where DF is the derivative of F. (For the original variational equation, this is the corresponding Euler-Lagrange equation.)

1.2. *Review*

What makes these minimization problems mathematically interesting? Our answer would then be that u lives in an infinite dimensional space! The structures and concepts relating to Banach spaces multiply and split when one moves from the finite dimensional case to the infinite dimensional case, much like the spectral lines of an atom when moved into a magnetic field. Our intuition is built on experience with $\mathbb{R}^2$ and $\mathbb{R}^3$, which are finite dimensional hilbert spaces (with the typical Euclidean inner product). It therefore takes a fair bit of effort to develop an intuition for infinte dimensional, non-hilbert spaces.

In the following we review ideas we need to understand the variational problems in their natural setting. These ideas include such items as:

- Infinite dimensional $\Rightarrow$ convergence and topologies "bifurcate" – geometry of Banach spaces is highly nontrivial.
- Convexity gives us existence; Strict convexity yeilds uniqueness.
- Lower semicontinuity plus coercivity plus the weak topology plus reflexivity yeilds existence of minimizers.
- Derivatives still work in Banach spaces.

1) A Banach space B is a vector space with a norm that makes it into a complete metric space. The dual space is simply the space of all bounded linear maps l: $l : B \to \mathbb{R}$. Hilbert spaces are special Banach spaces: Angles are defined, all

seperable Hilbert spaces are isometrically isomorphic, L^2 and l^2 are both Hilbert spaces, Hilbert spaces are reflexive, etc.

2) For our purposes we will think of compact sets as those sets in which any infinite subset is *forced* to have an accumulation point. In finite dimensions, compact is equivalent to closed and bounded. However, in infinite dimensions a closed and bounded set can be non-compact.

3) A function is *convex* if for all x and y and every $0 < \alpha < 1$,

$$\alpha f(x) + (1-\alpha) f(y) \leq f(\alpha x + (1-\alpha) y). \tag{1}$$

$f(x)$ is said to be *strictly convex* if the above inequality is strict. (The graph of the function is nowhere flat). Note that convex functions are to optimization what linear systems are to differential equations.

4) A function f is continuous if

$$f(x_0) = \lim_{i\to\infty} f(x_i) \tag{2}$$

for all sequences $x_i \to x_0$. (Detail: in some spaces we need to use *nets* or *moore-smith sequences* and so the above is called *sequential continuity*.) A function f is *Lower Semi-Continuous* (l.s.c.) if

$$f(x_0) \leq \underline{\lim}_{i\to\infty} f(x_i) \tag{3}$$

for all sequences $x_i \to x_0$. (Detail: in some spaces we need to use *nets* or *moore-smith sequences* and so the above is called *sequential lower semi-continuity*.)

5) Now we will briefly explain with an example three types of convergence:

- **Strong Convergence:** $x_i \to x_0$ if $||x_i - x_0|| \to 0$ as $i \to \infty$.
- **Weak Convergence:** $x_i \to_w x_0$ if $l(x_i) \to l(x_0)$ as $i \to \infty$ for all $l \in X^*$.
- **Weak* Convergence:** $l_i \to_{w^*} l_0$ if $l_i(x) \to l_0(x)$ as $i \to \infty$ for all $x \in X$.

Example: L^2 functions on a measure space (Ω, μ)

$$||f|| = \left(\int_\Omega |f(x)|^2 d\mu(x) \right)^{\frac{1}{2}} \tag{4}$$

Due to the Riesz Representation theorem, we can express the dual space X^* as all linear functionals of the form:

$$l_g(f) = \left(\int_\Omega g(x) f(x) d\mu(x) \right) \tag{5}$$

Strong Convergence

$f_i \to f_0$ in the strong topology if

$$\lim_{i\to\infty} ||f_i - f_0|| = \lim_{i\to\infty} \left(\int_\Omega |f_i(x) - f_0(x)|^2 d\mu(x) \right)^{\frac{1}{2}} = 0 \tag{6}$$

Weak Convergence
$f_i \to f_0$ in the weak topology if

$$\lim_{i\to\infty} l_g(f_i) - l_g(f_0) = \lim_{i\to\infty} \int_\Omega g(x)(f_i(x) - f_0(x))d\mu(x) = 0 \qquad (7)$$

for all $l_g \in X^*$ (or equivalently, for all $l_g, g \in X$).
Weak* Convergence
Finally, $l_{g_i} \to l_{g_0}$ in the weak* if

$$\lim_{i\to\infty} l_{g_i}(f) - l_{g_0}(f) = \lim_{i\to\infty} \int_\Omega (g_i(x) - g_0(x))f(x)d\mu(x) = 0 \qquad (8)$$

for all $f \in X$. The primary practical reason for using other forms of convergence is that though one may lose the nice properties of a metric space (the strong topology is always a metric topology), one regains compactness for closed and bounded sets (there are details!).

6) A function $F : B \to \mathbb{R}$ is *coercive* if the level sets are bounded or equivalently if $F(u) \to \infty$ as $||u|| \to \infty$

7) Derivatives seen corectly: $DF(u)$ is simply the linear approximation of F at u. $||F(w) - F(u) - DF(u)(w-u)||/||w-u|| \to 0$ as $||w-u|| \to 0$. Furthermore, if $F : B \to \mathbb{R}$ then DF is in the dual space B^*.

Finally, below is a table summarizing the differences between the finite and infinite dimensional case.

property	**finite dimensional**	**infinite dimensional**
continuity	all linear maps continuous	bounded linear maps continuous
compactness	the unit ball is compact	the unit ball is NOT compact
norms	All norms are equivalent	many different norms on the same space
topologies	Strong = Weak = Weak*	Strong $\neq$ Weak $\neq$ Weak*
sequences	sequences completely characterize topology	need *nets*
linear subspaces	always closed	not necessarily closed

1.3. *Existence, Uniqueness, Stability, Computation*

The above review prepares us to consider the existence of a solution to the problem:

$$F(u^*) = F^* \equiv \inf_u F(u).$$

Outline of establishment of exstence: Given

A1 B is reflexive,
A2 F is coercive, and
A3 F is weakly sequentially lower semi-continuous,

we may follow the steps:

1 Note that there is a sequence u_i such that $F(u_i) \to F^*$.
2 Since F is coercive, the level sets of F are bounded.
3 Under the weak topology, bounded sets are sequentially compact SO there is a subsequence of u_i call it u_{i_n} and a point u^* such that $u_{i_n} \to_w u^*$ as $n \to \infty$.
4 We use the weak sequential lower semi-continuity to obtain that

$$F(u^*) \leq \varliminf_{n\to\infty} F(u_{i_n})$$

5 This implies that $F(u^*) = F^*$.

Consequently, if any particular F satisfies A1, A2, and A3, we automatically have the existence of a minimizer u^*.

If we know that F is strictly convex we can immediatly conclude that any minimizer is unique since if not then $u_1^* \neq u_2^*$ and $F^* \leq F(\lambda u_1^* + (1-\lambda)u_2^*) < \lambda F(u_1^*) + (1-\lambda)F(u_2^*) \leq F^*$ or $F^* < F^*$.

Practical Uniqueness: We can, more practically ask questions that go beyond the existence question above and whose answers have real impact when considering applied problems.

1 How big is the bounding ball in which the minimal solutions live?
2 How big does this bounding ball get as we look for all solutions with F vaues within ε of F^*?
3 How big is the set of u satisfying $F_{\hat{d}}(u) \leq F^* + \varepsilon$ for some $\hat{d}$ within δ of the measured data d.

Very briefly then, practical uniquesness asks, "How big are the *level sets* of F near F^*?"

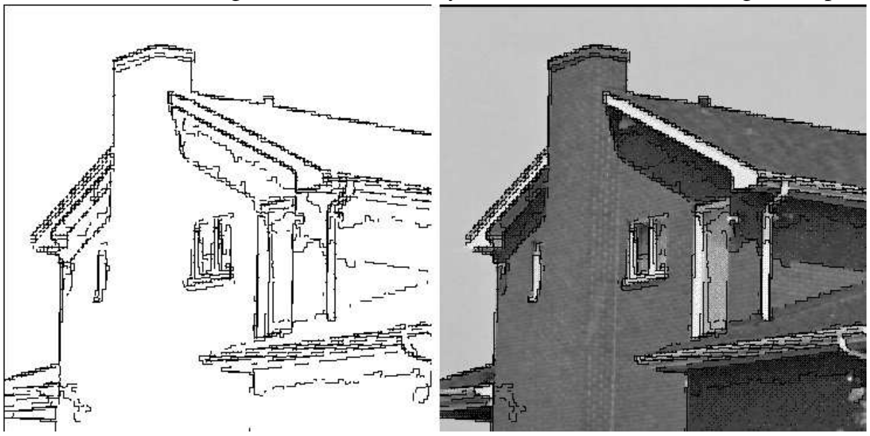
Stability: Although the formal definition of well posed-ness requires the continuous dependence of solutions on the initial data (in our case the measured image), I suggest that in fact what we are interested in *Stability*. (This idea is not new of course.) In particular cases we can have something concrete, for example: *If $F(u)$ is convex and the subdifferential at a minimizing u^* contains an* ε *ball around the origin of the dual space B^* then we know that at least our instability is no worse than* $1/\varepsilon$. This is intimately related to *practical uniqueness* as discussed above.

Computation: This is a huge industry; suffice it to say that one wants a correspondence between computed quantities and theoretical solutions. The verification that computations converge to the theoretical solutions *in the case of finite computational resources is nontrivial and is often more a point of faith than one of science*.

2. BV Functions and Total Variation Minimization

2.1. *Images: the Importance of Edges*

Edges or more generally, regions with high gradient, contain much of the information in an image. This can easily be seen in the following example:



It is therefore natural to consider measures of images $u : \Omega \subset \mathbb{R}^2 \to \mathbb{R}$ of the form:

$$F(u) \equiv \int_\Omega \phi(|\nabla u|)d\mu$$

We will look at $\phi(x) = |x|^p \ p \in (0,2]$.

(In images $F(u) = \int_\Omega |\nabla u|^p d\mu = \sum |\text{pixel differences}|^p$.)

Exersize: Convince yourself using the function (with variable parameter c) $\{f = 0 \ \ (x < 0), f = cx \ \ (0 < x < 1/c), f = 1 \ \ (1/c \leq x)\}$ that the case of $p < 1$

is biased towards edges, $p > 1$ is biased against edges and $p = 1$ is biased niether for nor against edges.

We now look at the case of $p = 1$ in more detail.

2.2. *BV Functions*

We now define a space of functions whose norm is based on the measure we introduced to look at edges, $F(f) \equiv \int_\Omega |\nabla f| d\mu$. The space will be those $f : \mathbb{R}^N \to \mathbb{R}$ such that

$$f \in L^1(\Omega) \text{ and } \int_\Omega |Df| < \infty$$

where we define $\int_\Omega |Df|$ when $f \notin C^1(\Omega)$ by

$$\int_\Omega |Df| \equiv \sup \left\{ \int_\Omega f \operatorname{div}(g) dx : g \in C_c^1(\Omega, \mathbb{R}^N), |g(x)| \leq 1 \ \ \forall x \in \Omega \right\}$$

In this case we define the BV norm by:

$$||f||_{BV} \equiv ||f||_{L^1} + \int_\Omega |Df|$$

The space of functions f such that $||f||_{BV} < \infty$ together with the BV norm is a Banach space of great usefulness and versatility.

$TV(f) \equiv \int_\Omega |Df|$ ($= \int_\Omega |\nabla f|$ when $f \in C^1$) is the *Total Variation* of f. $TV(u)$ can be also computed as an integral over level sets:

$$TV(u) = \int_\Omega |Du| dxdy = \int_\mathbb{R} L(u,r) dr$$

where

$$L(u,r) = \text{length of the boundary of the r-level set.}$$

2.3. *Total Variation Regularization*

Now we consider the image recovery problem and the role that optimization *and* total variation *play in regularized reconstruction from projections.*

A common regularization of the radiographic image reconstruction problem is the use of the L^2 norm to regularize the inverse problem.

$$u_{\text{optimal}} \equiv \arg\min_u F(u) = \arg\min_u \left(||Pu - d||^2 + ||u||_2 \right)$$

where P is the radiographic projection operator and d is the radiographic data. Another regularization (which leads to the usual heat equation approach to smoothing) is given by the minimization:

$$u_{\text{optimal}} \equiv \arg\min_u F(u) = \arg\min_u \left(||Pu - d||^2 + \int_\Omega |\nabla u|^2 \right)$$

But, as noted above, the $|\nabla u|^2$ is biased against edges, while $|\nabla u|$ is biased neither for or against edges. This leads us to consideration of:

$$u_{\text{optimal}} \equiv \arg\min_u F(u) = \arg\min_u \left(||Pu - d||^2 + \int_\Omega |\nabla u| \right)$$

Looking at the last regularization more carefully, we modify it so that we can constrain the solution to *not* match the data but rather differ by an amount predicted from knowledge of the magnitude σ^2 of the noise: $||Pu - d||^2 = \sigma^2$. We also develop the corresponding discrete analog since measured digital images are discrete. We end up with the following continuous functional and it's discrete counterpart:

$$F(u) = \lambda ||Pu - d||^2 + \int_\Omega |\nabla u|$$

$$\hat{F}(u) = \lambda ||Pu - d||^2 + \sum_{i,j} |\nabla_{i,j} u|$$

where $\nabla_{i,j}$ is the discretized gradient. For specific choices of λ we implicitly seek solutions constrained by $||Pu - d||^2 = \sigma_\lambda^2$).

Technically speaking, we need the discretized functional to be strictly convex and coercive for guaranteed existence and uniqueness. It also permits us to prove linear convergence.

2.4. *A Fixed Point Method*

If we compute the derivative of the functional

$$F(u) = \alpha \sum_{i,j} |\nabla_{i,j} u| + \frac{1}{2} ||Pu - d||^2$$

(where $\alpha = \frac{2}{\lambda}$) and set it to zero, we get

$$\alpha \sum_{i,j} \nabla_{i,j}^T \left(\frac{\nabla_{i,j} u}{|\nabla_{i,j} u|} \right) + P^T Pu - P^T d = 0$$

We turn this into an iterative method that (we can prove) converges to a unique fixed point. The iterative method is given by

$$\alpha \sum_{i,j} \nabla_{i,j}^T \left(\frac{\nabla_{i,j} u_{k+1}}{|\nabla_{i,j} u_k|} \right) + P^T P u_{k+1} - P^T d = 0$$

At each step we solve for u_{k+1} using a conjugate gradient method. A last modification to remove the singularity in derivative of the TV term is done by noticing that

$$|\nabla u| = \sqrt{|\nabla u|^2} \approx \sqrt{|\nabla u|^2 + \beta}$$

for small β.

2.5. *Other Methods*

- **Time Marching** This is the method used by Rudin, Osher, and Fatemi. Essentially one simply evolves u by treating the gradient of F as the time right hand side of a differential equation. In other words one solves

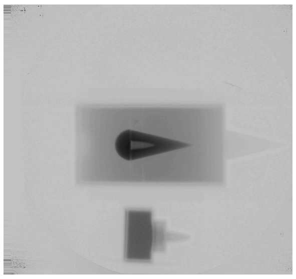
$$u_t = -DF(u).$$

- **Primal-Dual** In this method one uses the Fenchel Transform together with results from variational theory to reformulate the problem. One can obtain much better convergence results.

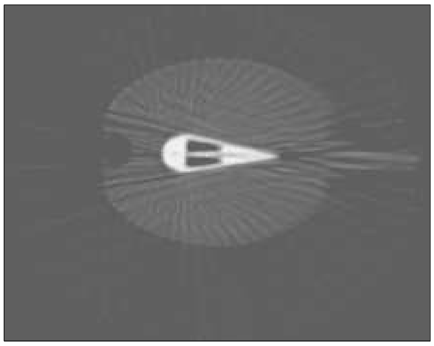
2.6. *Results*

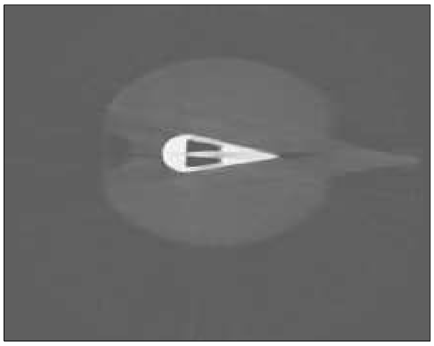
2.6.1. *BC04*

When applied to the tomographic inversion problem, TV minimization can give significantly improved results over the usual SVD regularized approach (whichis itself an improvement on the typical filtered backprojection method). The object (a calibaration object made of lead aluminum and plastic) was radiographed at the proton radiography facility of Los Alamos National Laboratory (Line C of LANSCE). Data was taken from 30 equally spaced angles. one such radiograph is shown below.



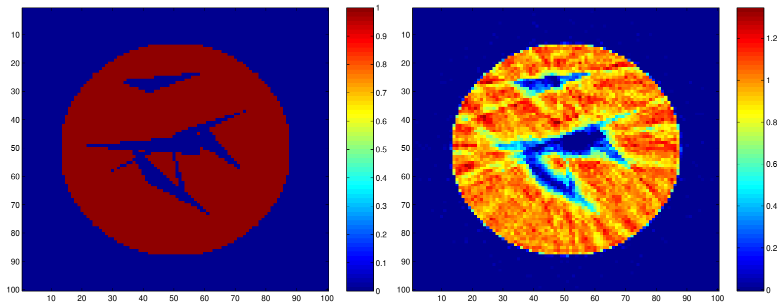
Below we show reconstructions of a slice through the object above. The image on the left shows the positively constrained SVD regularized reconstruction. On the right we show the positively constrained TV regularized reconstruction. (The results were obtained using our own hacks of Pep Mulet's ImageTool, a matlab package.)



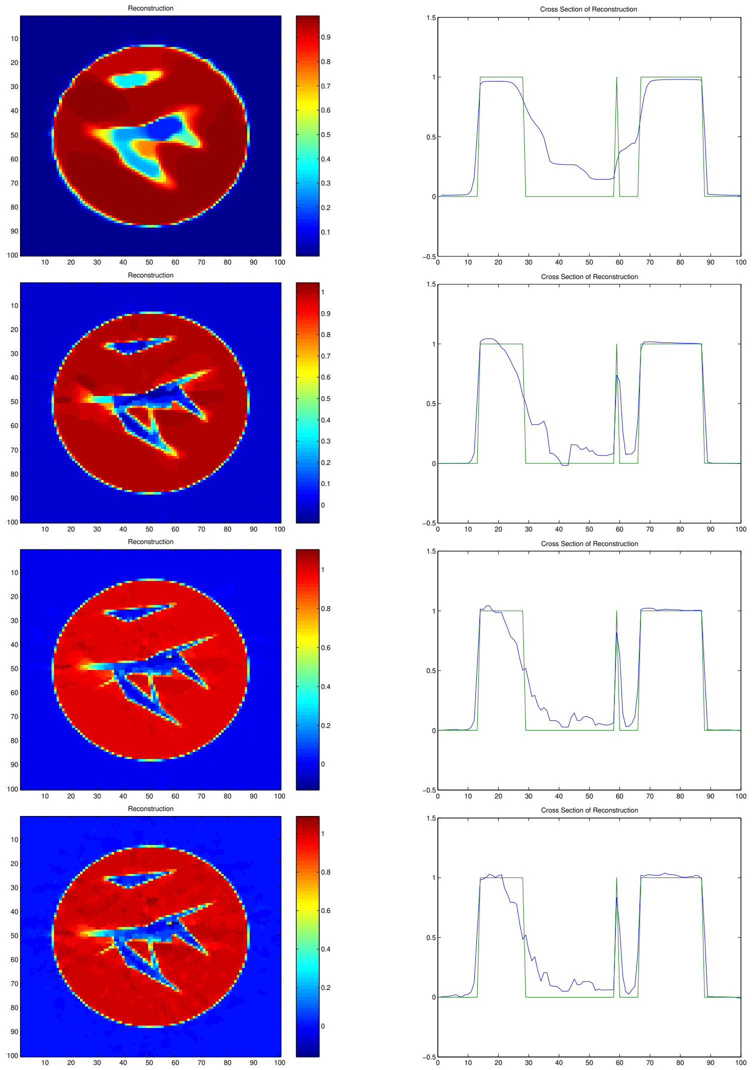


2.6.2. *A Closer Look: Simulated Radiographs*

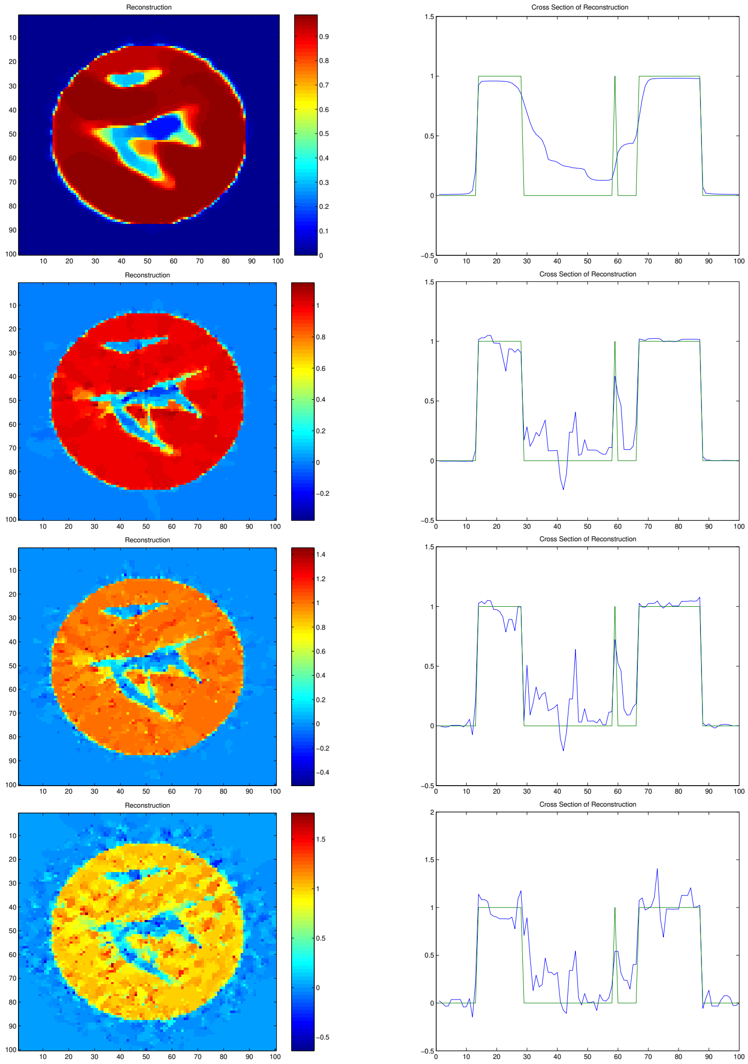
In these experiments we use a modified version of the fixed point code of Curt Vogel's to generate TV regularized reconstructions of a synthetic object. This synthetic object and it's truncated SVD reconstruction from 12 simulated radiographs are shown below:


10
20
30
40
50
60
70
80
90
100
10 20 30 40 50 60 70 80 90 100
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
1.2
1
0.8
0.6
0.4
0.2
0

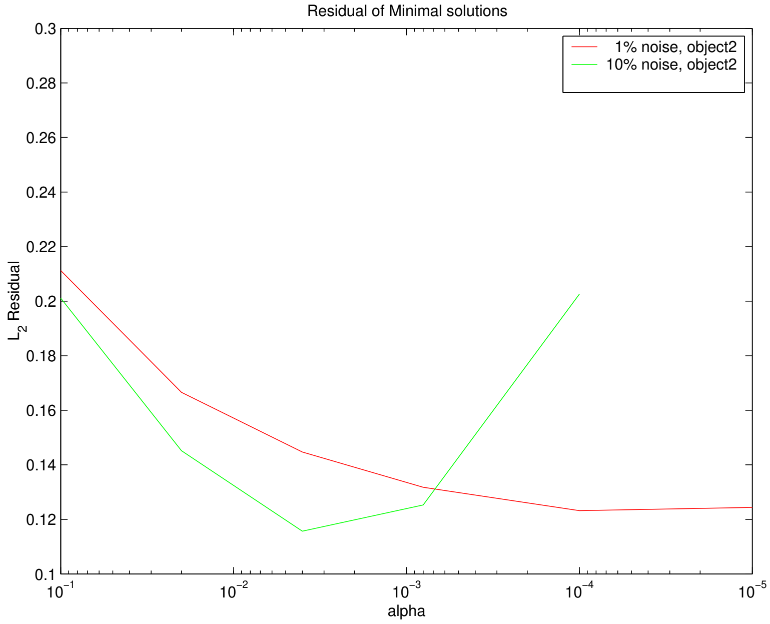
Reconstruction results: 1% noise (shown below for 4 decreasing values of α). Also shown is a line out of the imageat about line 50 which was more challenging than most to reconstruct due to the narrow "valley" the line goes through on the left of the object.



Reconstruction results: 10% noise, same description as above.



In the case of the 10% noise, we se that the last two choices for α require the reconstruction to be *too* close to the data. We know that the truth will be some distance from the data, not coincident with the data. This is confirmed in the following graph of residuals between the reconstruction and the truth (the original synthetic object).



We now move on to methods which have the TV method as a special case.

3. Nonlinear Diffusion, Robust Statistics, and More

In this section we will follow closely the presentation in chapter 4 of Sapiro's book, "Geometric Partial Differential Equations and Image Analysis". The motivation for the consideration of nonlinear diffusion as a method of image reconstruction or recovery can be understodd in the light of a few facts:

- "Anisotropic" or Non-linear diffusion is actually a generalization of TV minimization described above.
- Nonlinear diffusion is closely related to robust diffusion
- A close connection exists between non-linear diffusion and Markov Random Field approaches to image analysis, though we will not go into this in this lecture.

3.1. *Diffusion: Isotropic and "Anisotropic"*

One approach is to let the measured image d be the intial data (i.e. $u(x,0)=d$) and "run"

$$\frac{\partial u}{\partial t} = \nabla \cdot (\nabla u)$$

from $t=0$ to $t=\tau$. Doing so results in:

1 $u(x,\tau)$ is a smoothed version of the image with length scales suppresed smaller than $\sqrt{\tau}$
2 An undesirable loss of edge information.

Another approach is that of Perona-Malik (1990) who use the above with a modified PDE,

$$\frac{\partial u}{\partial t} = \nabla \cdot (g(||\nabla u||)\nabla u)$$

where g, known as the *stopping function*, is chosen so as to supress the diffusion at locations with high gradients.

3.2. *Discretized Perona-Malik*

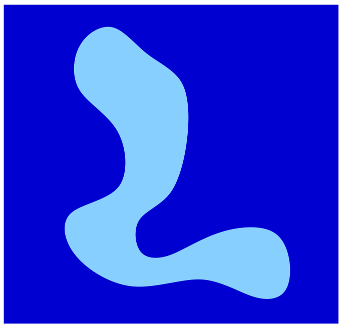
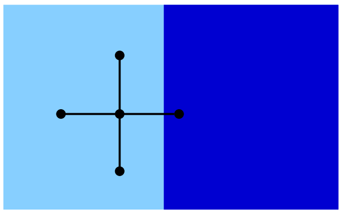
Perona and Malik discretitized their equation as:

$$u_i^{t+1} = u_i^t + \frac{\lambda}{|N|} \sum_{j \in N} g(|u_j^t - u_i^t|)(u_j^t - u_i^t)$$

- N is the neighborhood of the ith pixel
- If we set $g \equiv 1$ then the above is just the usual discrete diffusion equation.
- It is in this discrete form we will consider the relation to robust estimation.

3.3. *Robust estimation*

Suppose now we consider a piecewise constant image corrupted by noise. What is shown below is a portion of the noiseless image and a pixel in the discretization of that image and it's 4 neighboring pixel. The set of neighboring pixels will be denoted by N or N_i (i for the ith pixel).

Our Goal, when denoising:

- We want to average pixels within each region to remove noise
- But some averages include points from more than one region

In the case illustrated above, we want to estimate a piecewise constant function from a noisy version of that function.

Robust estimation approaches this problem by minimizing

$$\min_u \sum_i \sum_{j \in N_i} \rho(|u_j - u_i|, \sigma)$$

where ρ is a *robust error norm* (think of a negative log liklihood) with intrinsic scale σ. Example: in the Guassian case the likelihood function is $C\exp(-x^2/(2\sigma))$ so the negative log liklihood is simply $Cx^2/(2\sigma)$. This is NOT a robust error norm.

Idea Choose a norm such that differences above a scale σ are given less weight (or even much less weight).

We would at least expect the case where there is a separation in scales between the noise and the discontiniuties to respond well to this strategy.

So the question that now arises is: What is the relation between robust estimation and nonlinear diffusion? Recall that in the case of nonlinear diffusion we are flow the image according to:

$$\frac{\partial u}{\partial t} = \nabla \cdot (g(||\nabla u||)\nabla u)$$

and, noting that the continuous form of the robust estimation problem involves the minimization:

$$\min_u \int_\Omega \rho(||\nabla u||) d\Omega$$

which can be rewritten (via gradient descent) as:

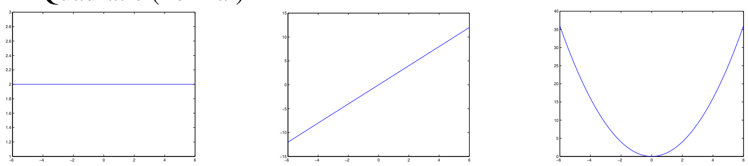
$$\frac{\partial u}{\partial t} = \nabla \cdot \left(\rho'(||\nabla u||) \frac{\nabla u}{||\nabla||} \right)$$

so if we let $g(s) = \rho'(s)/s$ we see that *robust estimation is nothing else than non-linear diffusion and vice versa*.

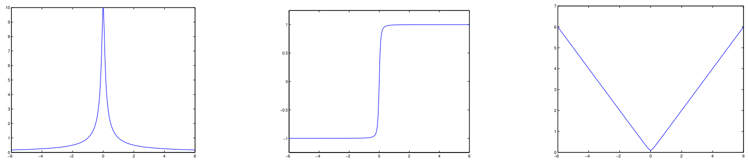
3.4. *Examples of stopping functions*

Each set of three plots include g, ψ, and ρ respectively:

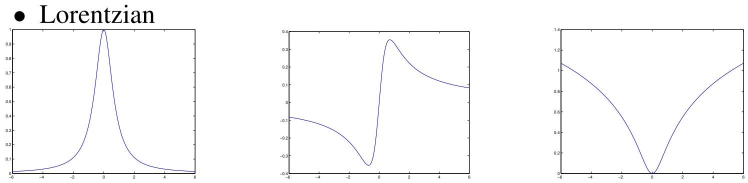
- Quadratic (normal)



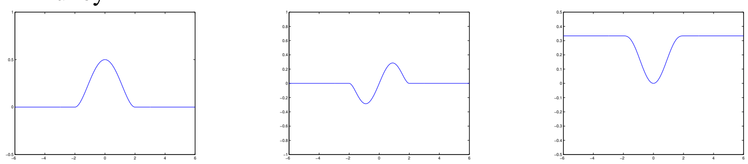
- TV min



- Lorentzian



- Tukey



3.5. *Comparison of results for different stopping functions*

An illustration of the use of nonlinear diffusion with the Lorentzian and Tukey functions is shown below. (The image is taken from from Black, Saprio, Marimont, and Heeger (1998) and can be found also in Sapiro's book .) Note that the Tukey stopping function preserves edges better due to the fact that it *ignores* gradients bigger than a certain magnitude. The original image is shown at the top, and results from the Lorentzian (left) and Tukey (right) stopping functions are shown below. The last two pictures are enlargements of the pictures above them.

4. Conclusion

In this paper we have introduced mathematical concepts important for careful consideration of the variational and PDE apporach to image analysis. We also introduced the TV minimization method with examples and then showed that a nonlinear diffusion approach is a generalization of the TV minimization approach. We briefly explored the relation to robust estimation.

Though the paper is more expository or tutorial in nature, there are new results given showing the efficacy of TV regularization for the radiographic inver problem. Interested readers can pursue the references given to learn more on the PDE/Variational approach to image analysis and can access the DDMA website at Los Alamos National Laboratory (http://ddma.lanl.gov) for the latest in projects involving analysis of image or image-like data.

References

1. Papers to be found as CAM reports at UCLA: Chan, Osher, Mulet, Strong, Vese, Strong, ...
2. Sapiro, “Geometric Partial Diferential Equations and Image Analysis”
3. Aubert and Kornprobst, “Mathematical Problems in Image Processing”
4. Osher and Fedkiw, “Level Set Methods and Dynamic Implicit Surfaces”
5. Ambrosio, Fusco, and Pallara, “Functions of Bounded Variation and Free Discontinuity Problems”
6. Evans and Gariepy, “Measure Theory and Fine Properties of Functions”.
7. Krantz and Parks, “The Geometry of Domains in Space”
8. Ekeland and Temam, “Convex Analysis and Variational Problems”
9. Bertsekas, “Nonlinear Programming”
10. Zeidler, “Nonlinear Functional Analysis and it’s Applications III: Variational Methods and optimization”
11. Vogel, “Computational Methods for Inverse Problems”