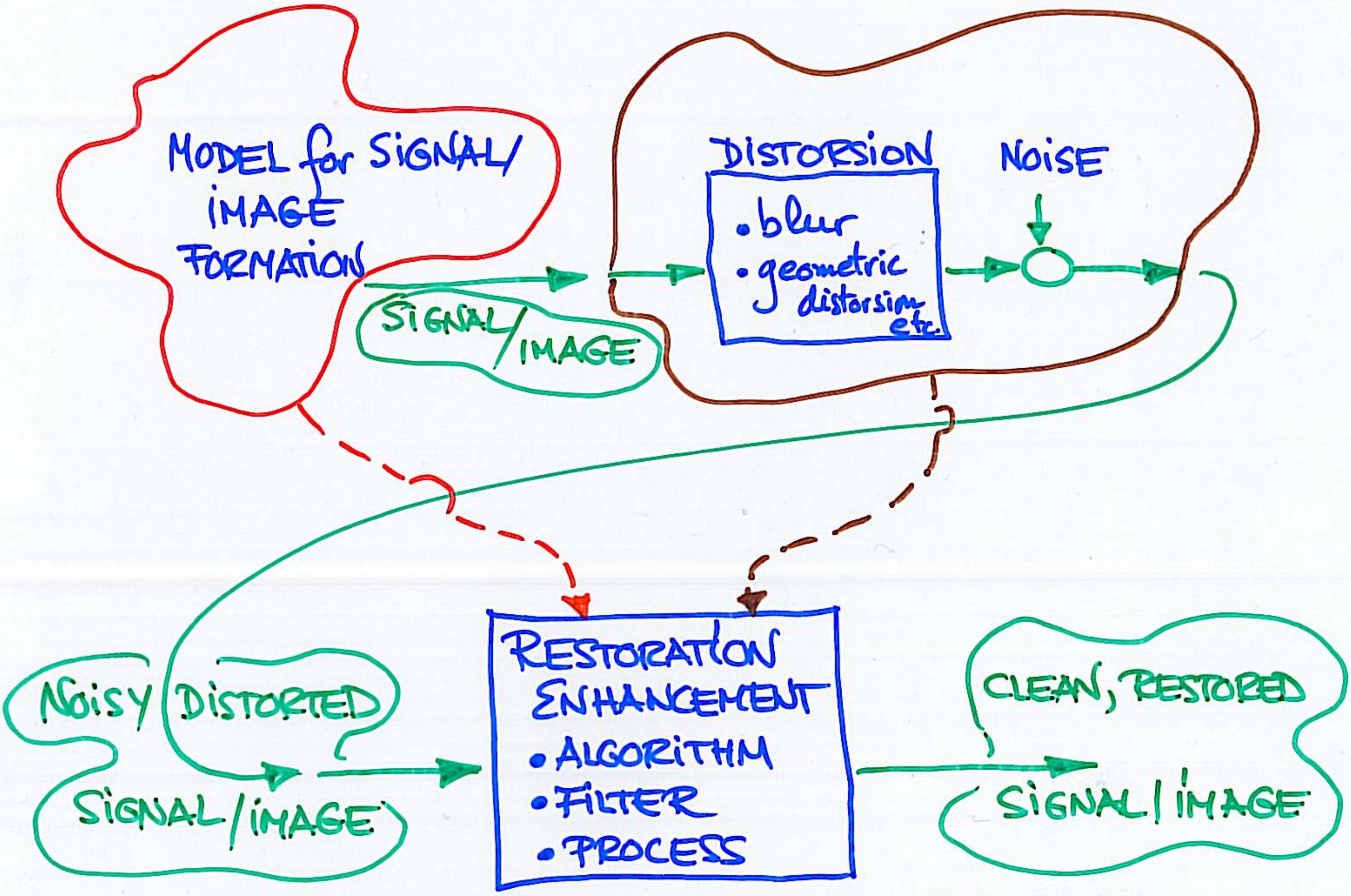
IN SEARCH FOR THE BEST OF THE BEST

or THE CHIMERA OF OPTIMALITY IN SIGNAL/IMAGE PROCESSING

Alfred M. Bruckstein

TECHNION, I.I.T, ISRAEL
and
NANYANG TU., SINGAPORE

SIGNAL / IMAGE PROCESSING



GOAL: Find a RESTORATION ALGORITHM that best exploits DATA (noisy SIGNAL/IMAGE + info about DISTORSION, NOISE + image formation) to produce clean SIGNAL/IMAGE as close as possible to the original.

Starting from the 60's the field of SIGNAL and IMAGE PROCESSING "exploded":

THE DENOISING EXAMPLE:

AVERAGING
- LOCAL/Gaussian SMOOTHING
- KERNEL SMOOTHING
- ADAPTIVE AVERAGING
- ITERATIVE SMOOTHING
- ITERATIVE Adaptive AVERAGING
- NON-LOCAL MEANS
- EXAMPLE BASED
- ⋮

DIFFUSIONS/PDE BASED
- HEAT EQUATION/LINEAR
- NONHOMOGENEOUS HEAT EQ.
- DIRECTIONAL DIFFUSION
- ANISOTROPIC DIFFUSION
- GEOMETRIC DIFFUSIONS
- DIFFUSIONS ON MANIFOLDS
- NON LINEAR PDE'S
- ⋮

FILTERING
- Fourier Transform Based
- LOW PASS FILERING
- ADAPTIVE Transform Domain
- Karhunen Loeve Transform
- Adaptive FILTERING
- WIENER/OPTIMAL Filters
- Coefficient Thresholding
- Multiresolution Filtering
- Wavelets/Edgelets/Curvlets
- X-lets coefficient filtering
- Laplace Beltrami Filters
- Bilateral Filters
- Shrinkage/Adaptive...
- Dictionary Based Sparsity

VARIATIONAL METHODS
- MSE-L_2 FUNCTIONALS
- TV-L_1-FUNCTIONALS
- Dictionary Rep. Based Functionals
- Geometric Functionals
- Overparametrized Rep. Funct.
- Local vs Non-Local Functionals
- Orientation Matching Functionals
- ⋮

$\boxed{\text{LOTS OF METHODS BASED ON A FEW IDEAS}}$

COMBINING & MIXING

THE VARIATIONAL APPROACH

The field of SIGNAL/IMAGE processing learnt optimization & calculus of variations

- 1940's WIENER'S OPTIMAL FILTERING
  signals as stochastic processes
- 1960's KALMAN'S OPTIMAL FILTERS
  state-space models for filtering
- 1980's STRUCTURED MATRIX-BASED O. FILTERS
  lattice filters
- 2000's NEW OPTIMIZATION CRITERIA
  total variation, ℓ_1-functionals
  sparsity methods

The new methods arose due to the availability of more & more computational power! Today we have the means to solve numerically a lot of optimization problems, hence the 'explosion' of new variational methods.

THE GENERAL VARIATIONAL APPROACH

GIVEN: Information on SIGNAL/IMAGE

Information on DISTORSION | NOISE

& THE NOISY, DISTORTED SIGNAL/IMAGE (DATA)

DEFINE A FUNCTIONAL

$$\boxed{\Psi\left[m \mid \text{data}, \text{prior info}\right]}$$

that will measure the quality of m as an estimate of the signal/image

Then solve:

$$\boxed{\hat{m} = \underset{m}{\text{optimize}}\ \Psi\left[m \mid \text{data}\right]}$$

THERE ARE MANY WAYS TO DEFINE "QUALITY MEASURES", HENCE THERE ARE MANY "OPTIMAL SOLUTIONS"

If there are MANY OPTIMAL SOLUTIONS:
WHICH ONE IS THE BEST?

THE ANSWER (TO THIS SILLY QUESTION) IS:

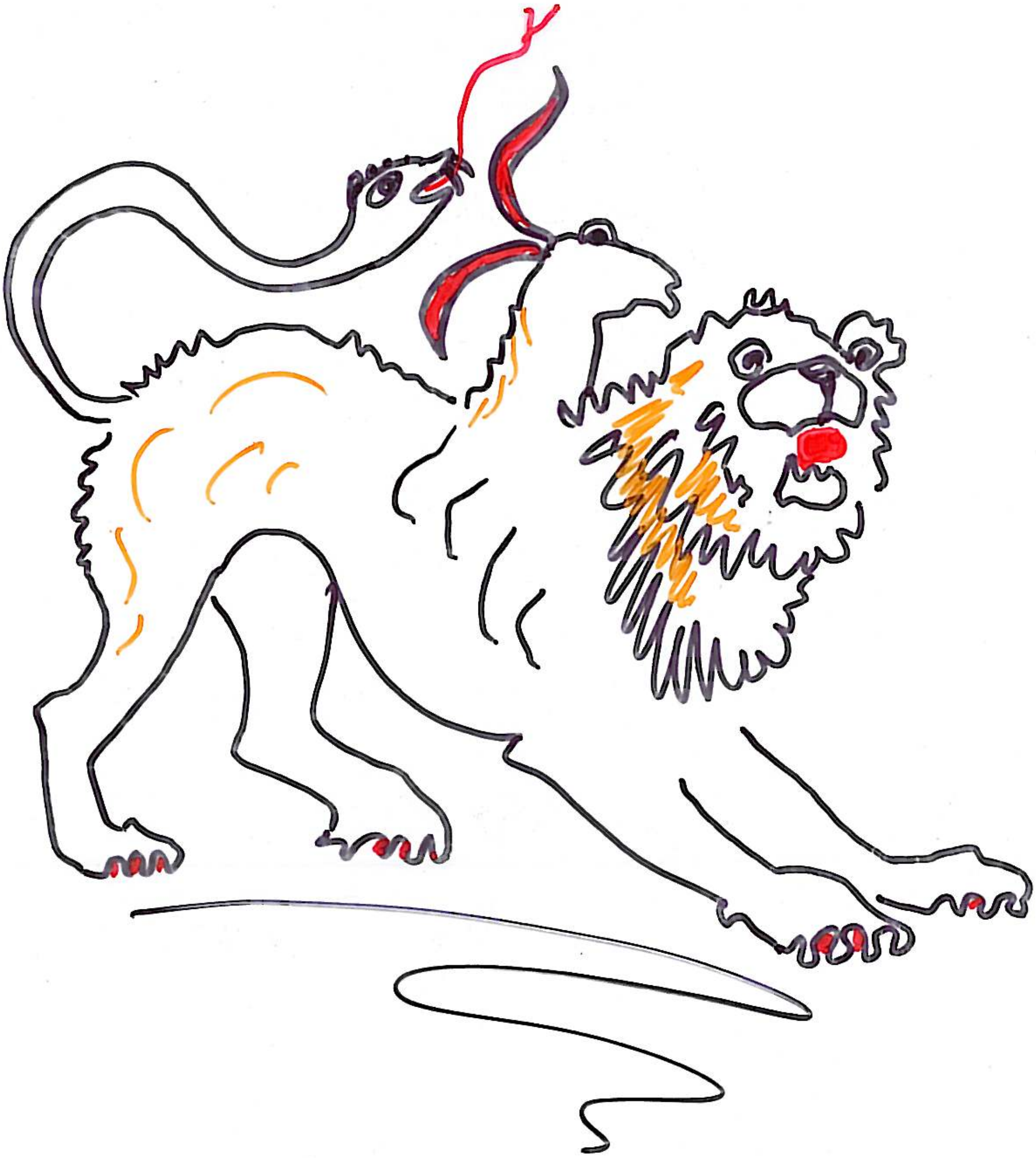
OPTIMALITY "IS LIKE A SEWER:
WHAT YOU GET OUT OF IT DEPENDS
ON WHAT YOU PUT INTO IT"
(paraphrasing Tom Lehrer's view on LIFE)

HENCE THE SUBTITLE

THE CHIMERA OF OPTIMALITY

- an illusion or fabrication of the mind
- an unrealizable dream, fantasy
- a fanciful mental illusion
- a fantastic impracticable plan or desire
- a dream or fantasy vs reality & truth
- a grotesque product of imagination

but also: an imaginary monster made of disparate parts
a mythical monster, composite of a lion, goat & serpent.



THE CHIMAERA: OLD ETRUSCAN STATUE

LET'S CONSIDER SOME EXAMPLES IN THE CONTEXT OF SIGNAL/IMAGE DENOISING

VARIATIONAL DENOISING

ASSUMPTIONS
- the ideal signal $m(x)$ is smooth
- additive noise $n(x)$ is "not smooth"

$$f(x) = m(x) + n(x)$$

Define a functional $\Psi\{\hat{m}(x)\}$ measuring how well suited $\hat{m}(x)$ is as an estimate of $m(x)$

$$\Psi\{\hat{m}(x)\} = \underbrace{\int_\Omega (f(x) - \hat{m}(x))^2 dx}_{\text{if } \hat{m}(x) \cong m(x) \text{ this is the } \mathcal{L}_2\text{-energy of the noise}} + \lambda \underbrace{\int_\Omega (\hat{m}'(x))^2 dx}_{\hat{m}(x) \text{ should be smooth}}$$

If $\hat{m}(x)$ minimizes Ψ we'll beleive it is a good estimate of the ideal $m(x)$. Hence we seek:

$$\hat{m}^{opt}(x) = \underset{\hat{m}(x)}{\text{minimize}}\, \Psi\{\hat{m}(x)\}$$

Note that we arbitrarily defined a $\Psi\{\cdot\}$ -
(here as a measure of the smoothness of $\hat{m}$
combined with the $\mathcal{L}_2$ energy of the noise
estimate!) - this is WHAT WE PUT INTO "OPTIMALITY".

What do we get out:

The Euler-Lagrange equations to be "necessarily"
satisfied by optimal $\hat{m}(x)$ is:

$$\lambda \hat{m}''(x) = \hat{m}(x) - f(x)$$

(or by taking the Fourier Transform)

$$\hat{m}(\omega) = \underbrace{\frac{1}{1+\lambda\omega^2}}_{\substack{\text{a Low Pass Filter with cut-off} \\ \text{controlled by } \lambda.}} f(\omega)$$

a Low Pass Filter with cut-off
controlled by λ. $\lambda \to 0 \quad \hat{m} \to f$
$\lambda \nearrow \infty \quad \hat{m} \to 0$

However we can do it also by an iterative approach. We do gradient descent on the variational functional Ψ

via:

$$\frac{\partial \hat{m}(x;t)}{\partial t} = - \frac{\delta \Psi\{\hat{m}(x,t)\}}{\delta \hat{m}(x,t)} =$$

Diffusion PDE!

$$= \left(\lambda \frac{\partial^2 \hat{m}(x,t)}{\partial x^2} + f(x) - \hat{m}(x,t)\right)$$

a pde in time and space for $\hat{m}(x,t)$ which for $t \to \infty$ should go to a solution of the EL equation.

THIS IS THE NEW WAY TO DO THINGS

as opposed to the old Low Pass Filter.

Why?

Because we want to put into the "OPTIMALITY" criteria ($\Psi\{\ \}$) all kinds of new and interesting things!

MODEL-BASED VARIATIONAL DENOISING

ASSUMPTIONS: • the ideal signal $m(x)$ can be represented as

$$m(x) = \sum_i a_i \varphi_i(x)$$

where $\varphi_1(x), \varphi_2(x) .. \varphi_i(x)$ are a given set of functions and we assume they are normalized and mutually orthogonal, i.e

$$\langle \varphi_i(x), \varphi_j(x) \rangle = \int_\Omega \varphi_i(x)\varphi_j(x)dx = \delta_{ij} \begin{cases} =0 & i \neq j \\ =1 & i=j \end{cases}$$

• and $\sum_i a_i^2 w_i^2$ is small

• the noise is "not smooth"

Here too we'll define a functional

$$\Psi\{\hat{m}(x)\} = \int_\Omega (f(x) - \hat{m}(x))^2 dx + \lambda \sum_i \hat{a}_i^2 w_i^2$$

but we set $\rightarrow$ $\hat{m}(x) = \sum_i \hat{a}_i \varphi_i(x)$

Here we need to solve:

$$\{\hat{a}_1 \hat{a}_2 .. \hat{a}_k ..\}^{opt} = \min_{\{a_i\}} \Psi\{\hat{a}_1, \hat{a}_2 .. \hat{a}_i ..\}$$

For this functional we have a multivariate optimization wrt the coefficients $\hat{a}_i$, hence we do (Lagrange)

$$\frac{\partial}{\partial a_j} \Psi = 0 \quad \text{or}$$

$$-2 \int \left(f(x) - \sum \hat{a}_i \varphi_i(x) \right) \varphi_j + 2 \lambda w_j^2 \hat{a}_j = 0$$

yielding $-\langle f, \varphi_j \rangle + \hat{a}_j (1 + \lambda w_j^2) = 0$

$$\text{or} \quad \hat{a}_j = \frac{1}{1 + \lambda w_j^2} \langle f, \varphi_j \rangle$$

Looks Familiar? Indeed, very similar to the Low Pass Filter result.

If we have no information on the a_i's of the true $m(x)$ we can set $\lambda = 0$ and approximate f by the O.N. sequence $\{\varphi_j(x)\}$ via $a_j = \langle f, \varphi_j \rangle$

(like in Fourier analysis, for example!)

- If w_j^2 are all 1, we get

$$\hat{a}_j = \frac{1}{1+\lambda} \langle f, \varphi_j \rangle$$

- If we know the statistics of the noise and signal coefficients we can recover the entire edifice of classical Wiener filtering this way.

Note that here too we could do gradient descent towards the "optimal" coefficients via

$$\frac{d\hat{a}_j(t)}{dt} = -\frac{\partial \Psi\{\cdot\}}{\partial \hat{a}_j} = \hat{a}_j(t)(1+\lambda w_j^2) - \langle f, \varphi_j \rangle$$

but this is not too interesting!

SO FAR WE SAW TWO EXAMPLES OF DENOISING PROVIDING:

- Low PASS Filtering
- Diffusion PDE / iterative algorithm
- Transform Domain Filtering
- Representation Coefficient "Shrinkage"
- Wiener Filtering

hence ALL THE CLASSICS.

BUT THE FOREFRONT OF RESEARCH is VERY VERY NEAR!

When we write:

$$\Psi\{\hat{m}(t)\} = \int_\Omega \phi(f-\hat{m})\,dx + \int_\Omega \Theta(\hat{m})\,dx$$

or

$$\Psi\{\{\hat{a}_i\}\} = \int \phi\left(f - \sum \hat{a}_i \varphi_i\right) dx + \Lambda\{\hat{a}_i \ldots\}$$

and realize that we can invoke ϕ's Θ's $\{\varphi_i\text{'s}\}$ and Λ's as we desire, we see that

THE "OPTIMAL DENOISING" GAME is REALLY AFOOT (& ENDLESS)

• "TOTAL VARIATION DENOISING" GAME

$$\Psi\{m\} = \int_\Omega |f(x) - m(x)|^p dx + \lambda \int_\Omega |m'(x)| dx$$

• edge preserving

ROF • median-like filtering

(Rudin Osher Fatemi)

$$\Psi\{m\} = \int_\Omega \sqrt{(f(x) - m(x))^2 + \varepsilon_1^2} + \lambda \int \underbrace{\sqrt{(m'(x))^2 + \varepsilon_2^2}}_{\text{like a length}}$$

(Regularized!)

Here one can write the EL Equation

$$\frac{m(x) - f(x)}{\sqrt{(m-f)^2 + \varepsilon_1^2}} - \lambda \frac{d}{dx}\left(\frac{m'(x)}{\sqrt{(m')^2 + \varepsilon_2^2}}\right) = 0$$

and solve this equation via a gradient descent method as discussed before.

(Popular variation $\boxed{p = 2}$ above!)

THE "SHRINKAGE & THRESHOLDING OF REPRESENTATION COEFFICIENTS" GAME

In the representation of the signal $m(x)$ w.r.t. the set $\varphi_1(x), \varphi_2(x) \ldots \varphi_k(x) \ldots$ the denoising game seeks for a set of coefficients $\hat{a}_1, \hat{a}_2 \ldots \hat{a}_k$ so that $\hat{m}(x) = \sum \hat{a}_i \varphi_i(x)$ will minimize

$$\Psi\{\{\hat{a}_i\}\} = \int_\Omega \Big(f(x) - \sum \hat{a}_i \varphi_i(x)\Big)^2 + \Lambda\{\hat{a}_1, \hat{a}_2 \ldots \hat{a}_k\}$$

Here let us select $\Lambda\{\hat{a}_i\} = 2\sum_i \Lambda_i(a_i)$ for some given functions Λ_i (Before we had $2\Lambda_i(a) = \lambda w_i a^2$

Lagrange optimization yields here:

$$-2\int_\Omega \Big(f - \sum \hat{a}_i \varphi_i\Big)\varphi_j \, dx + 2\Lambda_j'(\hat{a}_j) = 0 \quad j = 1, 2, \ldots k.$$

or $$-2\langle f, \varphi_j\rangle + 2\hat{a}_j + 2\Lambda_j'(\hat{a}_j) = 0$$

or $$\hat{a}_j + \Lambda_j'(\hat{a}_j) = \langle f, \varphi_j\rangle$$

The equation

$$\hat{a}_j + \Lambda(\hat{a}_j) = \langle f, \varphi_j \rangle$$

is an implict equation yielding the coefficiant of the φ_j in the representation of the "optimally denoised" signal $\hat{m}(x)$ as a function of the coefficient of φ_j in the representation of the DATA $f(x)$!

Popular choices for $\Lambda(a)$'s are $\Lambda(a) = |a|^p$ for $p = 1, 2, 3, \ldots$.

$$a + (|a|^p)' = A$$ lead to (formaly)

$$a + p|a|^{p-1}(\mathrm{sign}(a))^p = A$$ (if $p = 1$ for ex

we get $a + \mathrm{sign}(a) = A$

or

$$a = \begin{cases} A-1 & \text{if } A > 1 \\ 0 & \text{if } A \in [-1, 1] \\ A-1 & \text{if } A < -1 \end{cases}$$

Hence we have a thresholding of A and shrinkage outside the range $|A| < 1$.

- **A TOTALLY NEW GAME: OVERPARAMETRIZATION**

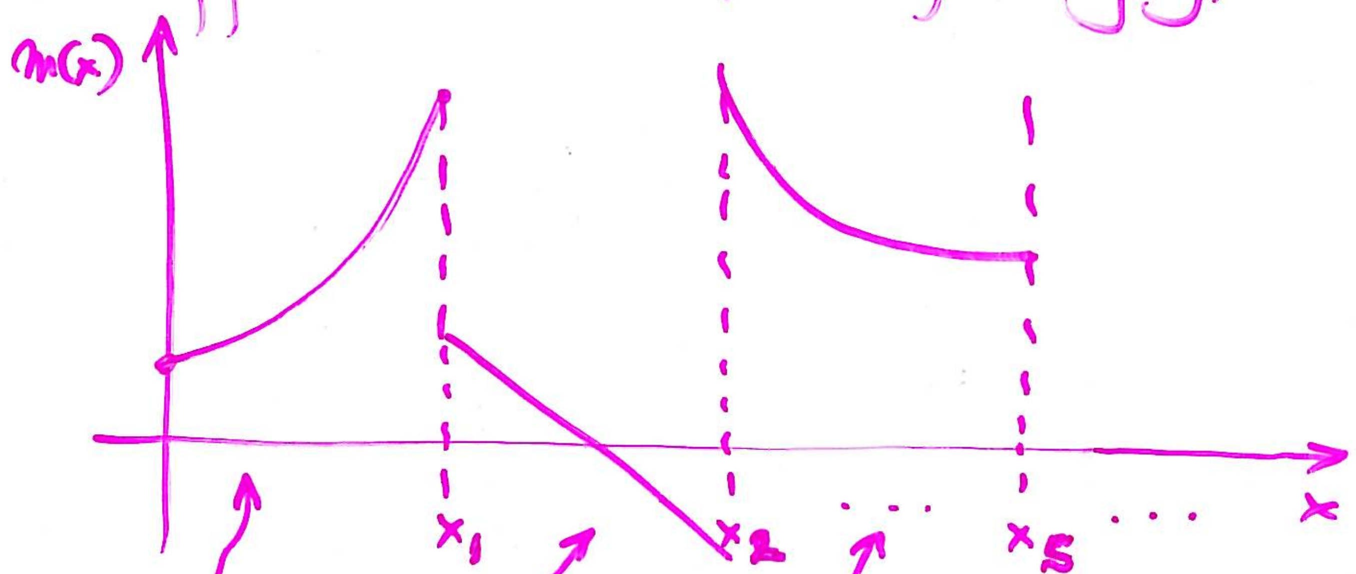
Suppose that there is reason to beleive that the original function $m(x)$ can be represented as follows:

$$m(x) = \sum_{i=0}^{K} a_i(x) \cdot \varphi_i(x)$$

where $\{\varphi_0, \varphi_1 .. \ \varphi_K\}$ are a <u>given set</u> of functions, not necessarily orthonormal. The functions $a_i(x)$ represent "space varying" coefficients (varying continuously in space but slowly, or, perhaps, being piecewise constant functions etc..)

Why would such a model be natural in some cases?

Suppose $m(x)$ is of the following type:



a polynomial of degree at most k in each region but with different coefficients.

Then
$$m(x) = \sum_{i=0}^{k} a_i(x)\, x^i$$

with "basis functions" $1, x, x^2, \ldots x^k$ with varying coefficients will nicely model $m(x)$.
Note that the breakpoints are unknown! but can be readily accomodated by the varying $a_i(x)$'s.

In this context we can search for a denoising algorithm that will be optimal w.r.t. the cost functional

$$\Psi\{a_0(x), a_1(x) \ldots a_K(x)\} = \int \left(f(x) - \sum_{i=0}^{K} a_i(x)\varphi_i(x)\right)^2 + \underbrace{+ \int \sum_{i=0}^{K} |a_i'(x)|^p \, dx}_{\text{requiring smoothness or piecewise smoothness of the } a_i(x)\text{-s}}$$

This method results in even more complicated EL equations, but by now we are not afraid of these, hence we proceed to VALIANTLY optimize ~ and get excellent results:

The Calculus of Variations does it for us, it segments and denoises very well!

THE OPTIC-FLOW ESTIMATION EXAMPLE

From a sequence of images $I(x,y;t)$ determine a time-varying vector field $[u,v] = [u(x,y;t)\ \ v(x,y;t)]$ so that we'll have

$$I(x,y;t) \cong I(x+u(x,y;t)\delta,\ y+v(x,y;t)\delta;\ t+\delta)$$

As stated here the $[u,v]$ field is general so we can apply all the variational tricks that assume no "model" for $[u,v]$.

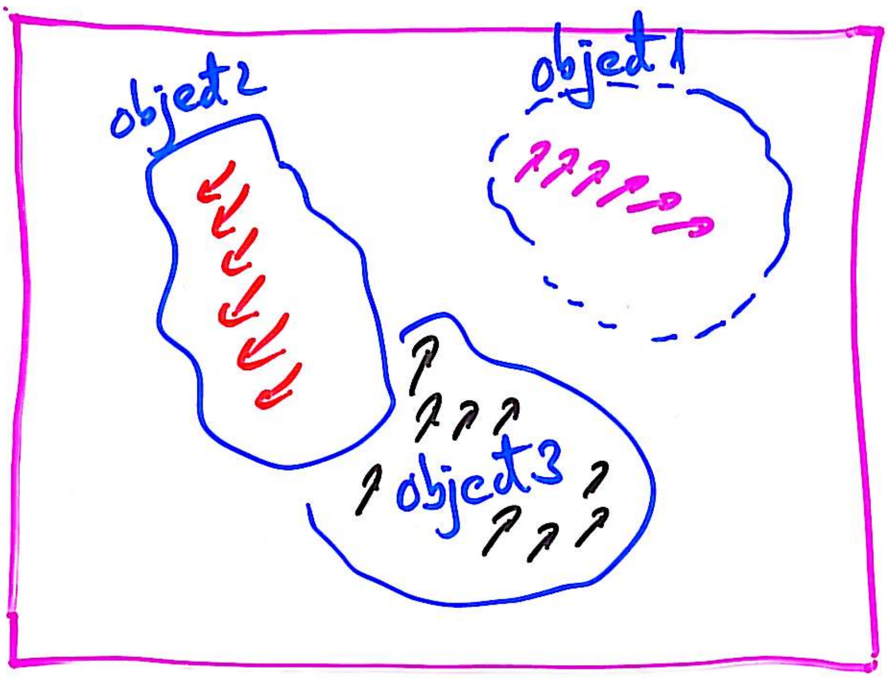
But from the Computer Vision books we know that $u(x,y;t)$ and $v(x,y;t)$ do have structure.

The optic flow of a "rigid" world is of the form:

$$\begin{cases} u(x,y;t) = -\Theta_1(t)\,1 + \Theta_3(t)\,x + \Omega_1(t)\,xy - \Omega_2(1+x^2) + \Omega_3(t)\,y \\ v(x,y;t) = -\Theta_2(t)\cdot 1 + \Theta_3(t)\,y + \Omega_1(t)(1+y^2) - \Omega_2(t)\,xy + \Omega_3(t)\,x \end{cases}$$

But the world is not really rigid: there are

- various objects in the scene moving differently
- non rigid objects (flapping flags)
- a rigid (distant) background etcetera...



Locally when we refer to a patch of the image:

$$\begin{cases} u(x,y,t) = -\theta_1(t) + \theta_3(t)x + \Omega_1(t)xy - \Omega_2(1+x^2) + \Omega_3(t)y \\ v(x,y,t) = -\theta_2(t) + \theta_3(t)y + \Omega_1(t)(1+y^2) - \\ \qquad - \Omega_2(t)xy - \Omega_3(t)x \end{cases}$$

These expression show that

$$\begin{cases} u(x,y,t) = \sum_i a_i(x,y,t)\,\varphi_i(x,y) \\ v(x,y,t) = \sum_i b_i(x,y,t)\,\eta_i(x,y) \end{cases}$$

i.e. exactly the overparametrized model considered!

The optic flow estimation problem can therefore be cast as:

$$E(u,v) = \int \Psi\big(I(x+u, y+v, t+\delta) - I(x,y,t)\big) + \\ + \int \Lambda\Big(\sum_i a_i^2, \sum b_i^2\Big)$$

So we can either

1) do our BEST (OPTIMAL? ☺) recovery of a general $[u,v]$-field and then fit to "various regions" the model above and from the Θ's & Ω's do shape-depth etc from motion

or

2) use an "overparametrized" model

$$\begin{cases} u(x,y;t) = \sum_i a_i(x,y;t)\,\varphi_i(x,y) \\ v(x,y;t) = \sum_i b_i(x,y;t)\,\eta_i(x,y) \end{cases}$$

known

and CRANK the VARIATIONAL wheel to recover OPTIMALLY (☺) piecewise constant (in x,y) a_i and b_i's

Solving the SEGMENTATION & PARAMETER (Θ, Ω) recovery problem while searching for the 'best' optic flow!

CONCLUSIONS

- THERE ARE LOTS OF VARIATIONS ON THE OPTIMAL DENOISING GAME
- THE SELECTION OF AN OPTIMIZATION CRITERION (FUNCTIONAL) IS CRUCIAL
    - THE FUNCTIONAL MUST EXPLOIT (AS MUCH AS POSSIBLE, AS WELL AS POSSIBLE!) THE PRIOR INFORMATION ON SIGNAL & NOISE

because

- WHAT WE GET OUT OF THEM DEPENDS A LOT ON WHAT WE PUT INTO THEM...
- THE SOLUTION OF THE VARIATIONAL PROBLEM IS PROVIDED BY THE E-L THEORY + NUMERICS

Selection of the Functionals remains an ART!