A gentle introduction to

Encoding prior knowledge in image- and data analysis

J. Lellmann

Heidelberg Laureate Forum 2015
24 August 2015
A + B = 4

A = ?  B = ?
A natural object (and data in general) is often

- composed of few objects,
- each of which is “simple” (e.g., geometrically)
\( b = T(u) \)
b = T(u) + n
\[ b = T'(u) + n' \]
Missing data
Given measurements $\mathbf{b}$, find image data $\mathbf{u}$ so that

$$\mathbf{b} = \mathbf{T}(\mathbf{u}) + \mathbf{n}$$

$\mathbf{T}$ structural operator, $\mathbf{n}$ random noise

Often the direct reconstruction is not unique, not stable, or not deterministic – we need prior knowledge
Variational methods

We reconstruct the unknown data $u$ from the measurements $b$ by minimizing the energy

$$\min_u \{ D(T(u); b) + R(u) \}$$

Advantages:

• **Intuitive** – we specify what the results should look like
• **Often statistical motivation** – maximum a posteriori estimate
• **Modular, reusable components**
\int_{\Omega} \| u(x) - b(x) \|_2^2 dx + \lambda \int_{\Omega} \| \nabla u(x) \|^2 dx
Trade-offs

Model complexity vs. computability
Local minimizers vs. global minimizers

Top-down approach
- Physically/biologically motivated
- Advantages:
  - Very specific to the problem
  - Model parameters have meaning

Bottom-up approach
- Built from simple, well-understood components
- Advantages:
  - Mathematical analysis
  - Efficient and/or global optimization often possible
Convexity assures that every local minimizer of the energy is also a global minimizer.
Non-smoothness

Non-smoothness often allows perfect recovery.

If $f$ is convex but non-differentiable, then (Fermat):

$$0 \in \partial f(u) \Rightarrow u \text{ is a global minimizer}$$
A simple example

Assume $u$ is scalar and our (perfect!) prior knowledge is that $u$ is zero. We measure $b = u + n$ and try:

$$u^* = \frac{b}{1 + \lambda}$$

$$\min_{u \in \mathbb{R}} \frac{1}{2} (u - b)^2 + \frac{\lambda}{2} u^2$$
A simple example

Assume $u$ is scalar and our (perfect!) prior knowledge is that $u$ is zero. We measure $b = u + n$ and try:

The solution is exact for a wide range of measurements!

$$u^* = \begin{cases} 
0, & |b| \leq \lambda, \\
 b - \lambda \text{sgn}(b), & |b| > \lambda. 
\end{cases}$$

$$\min_{u \in \mathbb{R}} \frac{1}{2}(u - b)^2 + \lambda |u|$$
Smooth:
\[ u = \min_{u \in \mathbb{R}} \frac{1}{2} (u - b)^2 + \lambda u^2 \]

Nonsmooth:
\[ u = \min_{u \in \mathbb{R}} \frac{1}{2} (u - b)^2 + \lambda \left\| u \right\| \]
Basis Pursuit

Given an overcomplete dictionary $A$ and a vector $u$ with few non-zero components, recover $u$ from $b = A u$:

$$\min_{u \in \mathbb{R}^m} \|u\|_0 \text{ subject to } Au = b$$

Applications in compression, data separation, machine learning, approximation theory,…

Compressive Sensing, Dictionary Learning consider how to choose $A$ optimally.

Candes, Romberg, Tao ’06; Donoho ’06
Overview: Foucart, Rauhut ‘10
Phase transition

This gives the exact (sparsest) solution in many cases!
Sparse Shape Decomposition

Lellmann, Breitenreicher, Schnörr ‘11
\[ \int_{\Omega} \|u(x) - b(x)\|^2 dx + \lambda \int_{\Omega} \|\nabla u(x)\|^2 dx \]
\[ \int_\Omega \| u(x) - b(x) \|_2^2 \, dx + \lambda \int_\Omega \, d \| Du \|_2 \]
Back to the roots
Gradient Descent

To minimize $f(u)$, follow the gradient downwards:

$$u_t = -\nabla f(u) \quad \Rightarrow \quad \frac{u^{k+1} - u^k}{t^k} \in -\partial f(u^k)^{-1}$$

This backward step is unique and can be computed explicitly for many “simple” convex functions.

Then apply backward steps to saddle-point form

$$\inf_u \sup_v F(u) + v^\top Au - G(v)$$

This is slow: $O(1/N)$, linear if strongly convex. **BUT:**

Rockafellar ’76; Lions, Mercier ’79; Glowinski, Marocco ’75; Gabay, Mercier ’78; Bertsekas, Tsitsiklis ’89

Alvarez, Attouch ’01; Nesterov ’04; Pock et al. ’09; Combettes, Pesquet ’11; Shefi, Teboulle ‘14
The 80-20 rule

Sometimes quantity (speed) beats quality (accuracy)!
Demo
But what if...
Labeling problems
First approach:

$$\min_{u: \Omega \rightarrow X} f(u) := D(u; I) + R(u)$$

$$X := \{ \text{yellow, green, blue} \}$$

This is a combinatorial problem and generally very hard:

- \(X\) does not have an additive structure - no gradients,
- in particular there is no convexity.
Let’s replace it!
Relaxation

Potts '52; Boykov et al. '98, '01; Kleinberg, Tardos '01; Zach et al. '08; Lellmann, Becker, Schnörr '09; Chambolle, Cremers, Pock '11
Relaxation

Hard decisions are replaced by soft “probabilities”

Potts ‘52; Boykov et al. ‘98, ‘01; Kleinberg, Tardos ‘01; Zach et al. ’08; Lellmann, Becker, Schnörr ‘09; Chambolle, Cremers, Pock ‘11
We would like to extend the problem

\[
\min_{u':\Omega \to X} f'(u')
\]

to the probability measures:

\[
\min_{u:\Omega \to \mathcal{P}(X)} f(u)
\]

How to define \( f \)?

- If \( u \) is integral at every point, \( f \) should agree with \( f' \)
- Otherwise, \( f \) should not create artificial minimizers
Relaxing segmentation

Assume

• assigning label $i$ to point $x$ costs $s_i(x)$.
• boundary between label $i$ and $j$ costs $d(i,j)$.

Then a possible local relaxation is

$$
\min_{u \in BV(\Omega, \mathcal{P}(X))} f(u) := \int_\Omega \langle u(x), s(x) \rangle \, dx + \int_\Omega d\Psi(Du)
$$

$$
\Psi(z) = \sup_{v \in \mathcal{D}} \langle v, z \rangle
$$

$$
\mathcal{D} = \{(v^1, \ldots, v^L) \in \mathbb{R}^{d \times L} \mid \|v^i - v^j\| \leq d(i, j), \sum_k v^k = 0\}
$$

Chambolle, Cremers, Pock ’11
Lellmann, Schnörr, SIAM J. Imaging Sci. ’10
Is it optimal?
Is it optimal?
Proving optimality

For two classes, recovery is exact:

\[ f(\text{round}(u_{\text{relaxed}}^*)) = f(u_{\text{integer}}^*) \]

For \( n > 2 \) classes, the discrete problem is NP-hard. But:

\[ \mathbb{E}_{\gamma} f(\text{round}_\gamma(u_{\text{relaxed}}^*)) \leq 2 \frac{\max_{i \neq j} d(i, j)}{\min_{i \neq j} d(i, j)} f(u_{\text{integer}}^*) \]

Two-class: Strang ‘83; Chan, Esedoglu, Nikolova ‘06; Zach et al. 09, Olsson et al. 09
Finite-dimensional: Dahlhaus et al. ‘94, Kleinberg, Tardos ‘99, Boykov et al. ‘01, Komodakis Tziritas ‘07
Non-convex functions

Manifold-valued data

Cremers, Strekalovskiy ‘12
Lellmann, Strekalovskiy, Kötter, Cremers ‘13
Manifold-valued data

Cremers, Strekalovskiy ’12
Lellmann, Strekalovskiy, Kötter, Cremers ‘13
RGB-Depth Segmentation (Diebold et al., SSVM ‘15)

Kolev et al., Int. J. Comp. Vis. ‘09
Why so complicated?
Combinatorial methods

Markov Random Fields, Graphical Models, Graph Partitioning,…

Solve 2-class case using min-cut/max-flow, n-class case using combinatorial solvers: integer program, branch and bound/cut, move making, commercial solvers
A continuous world

\[
\min_{u: \Omega \to \mathcal{P}(X)} \int \ldots
\]

\[
\min_{u \in X^n} \sum
\]
Variational methods are intuitive

(True) non-smoothness is essential

If it doesn’t fit, think big!

We live in a continuous world (and we actually need the hard math)