Calculus of Variations, Lecture 4, 2/12/2017.

[notes: No lecture 2/20 (Pres Day). On 2/17 we'll start with 2nd half of the "Lecture 3" notes — an example of the "calibration method"]

In Lecture 3 we discussed some dual pairs involving $L^1 + L^\infty$ type norms. —

L^1-L^∞ duality also arises in other settings, some quite current and important; these notes provide a brief introduction to a few such examples.

The heart of the matter is that L^1 optzn often produces sparse solns. (So: if the truth is sparse, L^1 optzn is biased toward it.)

Warmup example #1: in inverse problems we're often trying to "invert" a non-invertible operator (or one whose inverse is badly conditioned); for example, given an "image" $f(x)$ we may know only a blurred version $g = K * f$ (eg K = a Gaussian), or maybe $g = K * f + e$ where e is random ("noise").

In such setting many choices of f are consistent with the data (or nearly so), and the inversion must somehow make a choice. A standard (very old) idea is to favor regularity via Tychonoff regn, eg

$$\min_f \int |K*f - g|^2 + \lambda \int |\nabla f|^2$$

Advantage: pbm is quadratic $\Rightarrow$ EL eqn is linear in f. If we "correct" f is expected to be smooth this is a good idea. But if "correct" f is, say, piecewise constant then this regn will blur it. An obvious fix would be to consider

$$(1) \quad \min_f \int |K*f - g|^2 + \lambda \int |\nabla f|$$

since then discontinuities are permitted. (Here it is ∇f that might be sparse.)

This idea lies at heart of some recent work on inverse pbms (eg Daubechies, Defrise, De Mol, CPAM 57 (11), 2004, 1413-1457), and also at heart of "total variation

regn for image processing"

$$(2) \quad \min_u \int |u-g|^2 + \lambda \int |\nabla u|$$

(I huge lit on this now; 1st paper was probably Rudin, Osher, Fatemi, Physica D 60 (1992) 259-268).

Note: numerical soln of pbms like (1) + (2) pose challenges, due to non-smooth character of $\int |\nabla u|$. A "primal-dual" approach to (2) is discussed in Chan, Golub, Mulet, SIAM J Sci Comp 20(6), 1964-1977.

Warmup example #2: Let's consider a problem similar to (1) but

- let's work in a basis
  $f = x_1 \varphi_1 + x_2 \varphi_2 + \cdots$
  so practically speaking only $x = (x_1, x_2, \cdots)$ matters
- let's suppose K is diagonal in this basis; since it should be smoothing, suppose $(Kx)_j = k_j x_j$ with $k_j \to 0$ monotonically as $j \to \infty$.

- lets use a weighted l_1 norm for regularization

$$\sum_{j=1}^{\infty} \gamma_j |x_j| \qquad \text{with } \gamma_j \to \infty \text{ as } j \to \infty, \text{ sometimes}$$

(rather than trying to write $\int |\nabla f|$ in this basis)

Then analogue of (1) is

$$(3) \quad \min_{x_j} \sum_j (k_j x_j - y_j)^2 + \gamma_j |x_j|$$

where y_j is given ($\vec{y}$ = basis expansion of g). Obviously we can do the optimization separately at each j. Dropping the index, we find

$$\min_x (kx-y)^2 + \gamma |x|$$

is achieved when

$$kx = y - \frac{\gamma}{2k} \quad \text{if} \quad y > \frac{\gamma}{2k}$$
$$x = 0 \quad \text{if} \quad |y| < \gamma/2k.$$
$$kx = y + \frac{\gamma}{2k} \quad \text{if} \quad y < -\gamma/2k.$$

So: if $|y_j| \to 0$ as $j \to \infty$, only a few initial j's will have nonzero x_j's (recall that $\gamma_j / k_j \to \infty$ as $j \to \infty$). This captures, in simple form, why ℓ^1 regn favors sparseness.

Now a brief intro to compressed sensing

Typical scheme for remote sensing:

- collect info that is a linear transform of the true image (ie if reality is f, you measure some linear operator such as $K * f$)
- use image processing software (eg JPEG) to store the image; it typically uses a basis (eg a wavelet basis) + relies on fact that real images are well-approx by sparse reprs in this basis (just as real nos are well-approx by a few digits of a decimal expansion).

Obvious idea: since 2nd bullet throws away a lot of information, let's look for a scheme for characterizing the image using relatively low-dim'l lin transforms, but also using hypoth of sparseness.

Concrete proposal: working in a basis (as in warmup #2) + assuming truth is sparse (so: $y = Ax_*$ where x_* has just a few nonzero entries and $A: \mathbb{R}^n \to \mathbb{R}^n$ a low-rank lin transf), try to recover x_* from y by solving

$$(4) \quad \min_{Ax=y} \sum_j |x_j|$$

(or maybe $\min_{\|Ax-y\|_{L^2} \le \delta} \sum_j |x_j|$ or $\min_x \|Ax-y\|_{L^2}^2 + \lambda \sum_j |x_j|$)

Hope is that soln to (4) will produce exact "truth" x_*. (Note: (4) is not intractable numerically, since it's equiv to the linear program $\min_{Ax=y} \sum_j t_j$.)

$$-t_j \le x_j \le t_j$$

Compressed sensing is mostly about choosing A in a suitably random way (eg randomly sampling the Fourier transform of x_0) and proving that this scheme works with high probability.

Here, however, let us stick to the following methodological question: how can we ever hope to prove, for a specific A and x_0, that this method works?

The answer lies in the dual to (4), which we easily derive as follows:

$$\min_{Ax=y} \sum |x_j| = \min_{Ax=y} \max_{|\sigma_j|\le 1} \sum \sigma_j x_j$$

$$= \max_{|\sigma_j|\le 1} \min_{Ax=y} \sum \sigma_j x_j$$

The thing $Ax=y$ constrains the linear form $\sum \sigma_j x_j$ only if the latter is a linear combo of the former (this is a basic lemma from linear programming) so the only σ's of interest have the form $\sigma_j = \sum_i z_i a_{ij}$ i.e. $\sigma = A^T z$,

and then the value is $\sum_i \sigma_i x_i = \langle z, Ax \rangle$. So the dual problem is $= \langle z, y \rangle$.

$$\max_{z} \sum_i z_i y_i$$

$$-1 \le (A^T z)_j \le 1 \quad \text{for each } j$$

Evidently we get equality when

$$\sigma_j x_j = |x_j| \quad \text{for each } j,$$

i.e. when $\sigma_j = +1$ whenever $x_j > 0$
$\sigma_j = -1$ whenever $x_j < 0$,
(no condition if $x_j = 0$).

But a little more is true: suppose, for a given x^*, we can find σ s.t.

$|\sigma_j| \le 1$ for all j + $\sigma = A^T z$ for some z (so σ is admissible in the dual) and whenever

$\sigma_j = 1$ whenever $x^*_j > 0$
$\sigma_j = -1$ whenever $x^*_j < 0$
$|\sigma_j| < 1$ whenever $x^*_j = 0$.
↑ strict!

Then any soln $\hat{x}$ of our primal plem (*) has "the same sparsity structure as x^*" (i.e. the same list of nonzero coeffts).
[In particular, if soln of $Ax=y$ is unique in the low-dim'l space of x's with this sparsity structure, then x^* is the unique soln of (4).]

Proof of this claim is elementary:

- if σ has the properties listed above then σ solves the dual + x^* solves the primal, since complementary slackness condn $\langle \sigma, x^* \rangle = \sum_j |x_j^*|$ is satisfied

- if $\hat{x}$ is any other soln of primal then it too must satisfy
$$\langle \sigma, \hat{x} \rangle = \sum_j |\hat{x}_j|$$
Since $|\sigma_j| \leq 1$ for all j we see that this implies
$$\sigma_j \hat{x}_j = |\hat{x}_j| \quad \text{for each } j$$
Since $|\sigma_j| < 1$ when $x_j^* = 0$ we see that $x_j^* = 0 \Rightarrow \hat{x}_j = 0$. ∎ (This is a more careful statt of what I meant above by "$\hat{x}$ has same sparsity structure

as x^*.)

How to show such σ exists? A typical approach is to solve an L^2-type pblm in z:

(5) $$\min \sum z_i^2$$
$$(A^T z)_j = 1 \text{ for each index } j \text{ st } x_j^* > 0$$
$$(A^T z)_j = -1 \text{ for each index } j \text{ st } x_j^* < 0$$

and show that the assoc $\sigma = A^T z$ has the desired property that $|\sigma_j| < 1$ at the remaining indices.

This has a chance of working since

* soln of (5) is the least-norm soln of a system of linear eqns — we understand soln very well (by linear algebra)

* soln of primal pblm (4) is typically unique, but soln of its dual (the σ pblm) is typically far from unique, so we may hope by this

method to pick out a convenient soln.

Typical thm: when A has been chosen in suitable (random) way, this approach succeeds (with high probability).

In compressed sensing literature σ is called a "dual certificate of optimality".

For good introductions (including proofs that this works, for some simple classes of linear transfs A; also plenty of refs to literature) see eg

- Carlos Fernandez-Granda's notes from his Spring 2016 class "Optimization-based Data Analysis" (still avail on his website), section on "Random projections + compressed sensing"

- Afonso Bandeira's notes "Ten Lectures + Forty-two Open Problems in the Mathematics of Data Science" (still avail on his website), section on "Compressed sensing + sparse recovery"