E9 203 Compressed Sensing and Sparse SP

24 Mar. 2021.

Last time:

- Properties of local minima of separable cost fns
  - with reweighting based opt.

Today

- Convergence of reweighting based methods
- Some interpretations of the cost fn.
- Time permitting: Sparse Bayesian learning.

<u>Recap</u>: Separable cost fn:

$$\min_{x \in \mathbb{R}^N} \sum_{i=1}^{N} g(x_i) \quad \text{s.t.} \quad y = Ax$$

(or)

$$\min_{x \in \mathbb{R}^N} \|y - Ax\|_2^2 + \lambda \sum_{i=1}^{N} g(x_i)$$

Start with $\boxed{g(x) = (x^2 + \beta)^{p/2}}$, $0 < p < 2$, $\beta > 0$.

Concave in x^2, bound using 1st order Taylor expansion.

$$\Rightarrow \min_{x \in \mathbb{R}^N} \|Ax - y\|_2^2 + \lambda \sum_{\ell=1}^{N} \left(|x_\ell^{(k)}|^2 + \beta\right)^{\left(\frac{p}{2}-1\right)} x_\ell^2$$

(p absorbed into λ.)

$$\Rightarrow w_\ell^{(k)} = \left(|x_\ell^{(k)}|^2 + \beta\right)^{(p/2 - 1)}, \quad W_k^{-2} = \text{diag}\left\{\left(|x_\ell^{(k)}|^2 + \beta\right)^{\left(\frac{p}{2}-1\right)}\right\}.$$

Update: $\boxed{x^{(k+1)} = W_k^2 A^T (\lambda I + A W_k^2 A^T)^{-1} y}$

When $p \to 0$, we get

$$w_\ell^{(k)} = \left(|x_\ell^{(k)}|^2 + \beta\right)^{-1} \qquad \text{[Chartrand-Yin update]}$$

When $\beta \to 0$, we get

$$w_\ell^{(k)} = |x_\ell^{(k)}|^{p-2} \qquad \text{[Focuss]}.$$

As $p \to 0$, $g(x) \to (x^2 + \beta)^0$, so we are minimizing

$$\|Ax - y\|_2^2 + \lambda \underbrace{\sum_{\ell=1}^{N} |x_\ell^2 + \beta|^0}_{\text{as } \beta \to 0, \text{ this } \to \|x\|_0}$$

$\Rightarrow$ An intuitive explanation of why $g(x) = (x^2 + \beta)^{p/2}$ is a good idea.

Alternative viewpoint (connection to $g(x) = \log|x|$)

$$\|x\|_0 = \lim_{p \to 0} \sum_{i=1}^{N} |x_i|^p$$

$$\boxed{\lim_{p \to 0} \frac{1}{p} \sum_{i=1}^{N} \left(|x_i|^p - 1\right) = \sum_{i=1}^{N} \log|x_i|} \qquad \text{[Show this!]}$$

Hence, $\exists$ a 1-1 correspondence betn (P_0) and

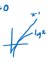
$$\min_{x \in \mathbb{R}^N} \sum_{i=1}^{N} \log|x_i| \quad \text{s.t.} \quad y = Ax.$$

Now, $\log|x_i| \le \frac{1}{2} \log|x_i^2 + \varepsilon| \le \boxed{\frac{x_i^2 + \varepsilon}{2\gamma_i} + \frac{1}{2}\log\gamma_i - \frac{1}{2}}$

where $\varepsilon, \gamma_i \ge 0$ arbitrary.

$\log x \le x - 1$ for $x \ge 0$

$=$ when $\gamma_i = x_i^2 + \varepsilon$.



Consider solving

$$\min_{\substack{x \in \mathbb{R}^N \\ [\gamma_1 \cdots \gamma_N]^T = \gamma \in \mathbb{R}_+^N \\ \varepsilon \in \mathbb{R}_+}} \boxed{\sum_{i=1}^{N} \frac{x_i^2 + \varepsilon}{2\gamma_i} + \frac{1}{2}\log\gamma_i} \quad \text{s.t.} \quad y = Ax.$$

Given γ, ε, quadratic in x w/ a linear constraint

$$x = \Gamma^{\frac{1}{2}} (A\Gamma^{\frac{1}{2}})^{\dagger} y, \quad \text{where } \Gamma = \text{diag}(\gamma).$$

$$(A\Gamma^{\frac{1}{2}})^{\dagger} = \Gamma^{\frac{1}{2}} A^T (A\Gamma A^T)^{-1}.$$

$$x = \Gamma A^T (A\Gamma A^T)^{-1} y.$$

Given x, ε, the optimal γ is

$$\gamma_i = x_i^2 + \varepsilon = |x_i^2 + \varepsilon|$$

The exact schedule for updating ε is not crucial.

Notice: the update is the same as before!

$$x^{(k+1)} = W_k^2 A^T (\lambda I + A W_k^2 A^T)^{-1} y$$

$$W_k^2 = \text{diag}\left\{\left(|x_\ell^{(k)}|^2 + \beta\right)^{1 - \frac{p}{2}}\right\} \xrightarrow[p \to 0]{} \text{diag}\left\{\left(|x_\ell^{(k)}|^2 + \beta\right)\right\}.$$

$$\left(w_\ell^{(k)} = \left(|x_\ell^{(k)}|^2 + \beta\right)^{-1}, \quad W_k^{-2} = \text{diag}\left\{\left(|x_\ell^{(k)}|^2 + \beta\right)^{-1}\right\}. \right)$$

$\Rightarrow$ $\sum_{i=1}^{N} \log|x_i|$ is a "good" cost fn. for solving (P_0) and the above is an iterative approach for solving the surrogate cost minimization problem.

Zangwill's global convergence theorem:

States that, under certain condns, which are satisfied here, the above updates are guaranteed to converge to a local min. or saddle pt. of

$$\sum_{i=1}^{N} \left\{ \left(\frac{x_i^2 + \varepsilon}{\gamma_i}\right) + \log\gamma_i \right\}$$

from any starting point $x^{(0)}$.

Analysis of ℓ_1 reweighting:

Focus on Candés et al's method: $w_i^{(k)} = \left(|x_i^{(k)}| + \varepsilon\right)^{-1}$

and solve $x^{(k+1)} = \arg\min_{x \in \mathbb{R}^N} \|y - Ax\|_2^2 + \lambda \sum_{i=1}^{N} w_i^{(k)} |x_i|$.

$$g(x) = (|x| + \varepsilon)^p, \quad \varepsilon > 0, \quad 0 < p < 1$$

Concave in $|x|$

$$g(x) \le g'(x_0)\, x - (\text{terms that only dep. on } x_0)$$

not imp. for opt.

At the $(k+1)$th update, we solve

$$\boxed{x^{(k+1)} = \arg\min_{x \in \mathbb{R}^N} \|y - Ax\|_2^2 + \lambda \|W_k^{-1} x\|_1}$$

where

$$W_k^{-1} = \text{diag}\left(g'(x^{(k)})\right) = \text{diag}\left(\left(|x_\ell^{(k)}| + \varepsilon\right)^{p-1}\right)$$

So as $p \to 0$, we get the Candés et al (2008) update.

$\Rightarrow$ for small enough ε, we are solving

$$\min_{x \in \mathbb{R}^N} \|y - Ax\|_2^2 + \lambda \|x\|_0, \text{ as desired}$$

In the noiseless case, the weight update is equivalent to solving

$$\boxed{\min_{x \in \mathbb{R}^N} \sum_{i=1}^{N} \log(|x_i| + \varepsilon) \quad \text{s.t.} \quad y = Ax} \quad \text{—} (*).$$

As before, by the Zangwill theory of global convergence, the iterates converge to a local min or saddle pt. of $(*)$.

[HW: Show that

$$\log(|x_i| + \varepsilon) \le \frac{x_i^2}{\gamma_i} + \log\left[\frac{(\varepsilon^2 + 2\gamma_i)^{\frac{1}{2}} + \varepsilon}{2}\right] - \frac{\left[(\varepsilon^2 + 2\gamma_i)^{\frac{1}{2}} - \varepsilon\right]^2}{4\gamma_i}$$

for all $\varepsilon, \gamma_i \ge 0$, with equality iff $\gamma_i = x_i^2 + \varepsilon|x_i|$.

$\Rightarrow$ leads to a different iteratively reweighted ℓ_2 update.

Sparse Bayesian Learning (SBL)

$$y = Ax + w, \quad w \sim \mathcal{N}(0, \sigma^2 I)$$

$$p(y \mid x; \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)^m}} \exp\left(-\frac{\|y - Ax\|_2^2}{2\sigma^2}\right).$$

Directly finding the ML est. of x from $p(y \mid x; \sigma^2)$ amounts to min. $\|y - Ax\|_2^2$, which does not yield sparse solns. $\Rightarrow$ Incorporate a sparsity-promoting prior on x.

FOCUSS, BP etc can be cast into this framework, and can view them as finding the MAP est. of x.

For example, $p(x) \propto \exp\left(-\sum_{i=1}^{N} |x_i|^p\right)$, $p \in [0, 1]$.

Then, the MAP est. of x is

$$x_{MAP} = \arg\max_{x \in \mathbb{R}^N} p(x \mid y) \propto p(y \mid x)\, p(x)$$

$$= \arg\min_{x \in \mathbb{R}^N} \boxed{-\log\left(p(y \mid x)\right) - \log p(x)}$$

$$= \arg\min_{x \in \mathbb{R}^N} \|y - Ax\|_2^2 + \lambda \sum_{i=1}^{N} |x_i|^p \Leftarrow$$

The above is the cost fn. for BP when $p = 1$, and for FOCUSS when $p < 1$.

Our previous algos can be viewed as a MAP est$^{\underline{mn}}$ problem under an appropriately chosen sparsity promoting prior on x.