Solving ℓ_1 Regularized Least Square Problems with Hierarchical Decomposition

M. Zhong[1]

mzhong1@umd.edu

[1]AMSC and CSCAMM
University of Maryland College Park

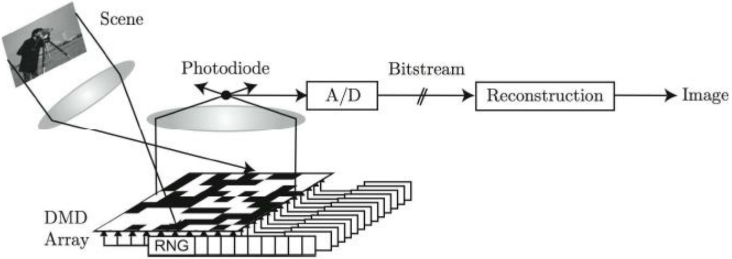
Project Proposal for AMSC 663
October 2nd, 2012

Outline

Outline

Example (A Sparse Signal in Compress Sensing)

How to encode a large sparse signal using a relatively small number of linear measurements?



[1]

[1] http://dsp.rice.edu/cscamera

A Formulation

An Intuitive Approach

Problem (The Original One)

$$\min_{x \in \mathbf{R}^n} \{ ||x||_0 | Ax = b \} \tag{1}$$

- Where the $|| \cdot ||_0$ norm means the number of non-zero elements in a vector, $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$, and $m \ll n$ (under-determined system).
- The $|| \cdot ||_0$ problem is solved mainly using combinatorial optimization, and thus it is NP hard[2].
- The $|| \cdot ||_0$ is not convex, one can convexify the problem by using either ℓ_1 or ℓ_2 norm, and then (1) can be solved using convex optimization techniques (namely linear programming).

[2] B.K.Natarajan, 95

A Better Formulation
In $||\cdot||_1$

We'll use $||\cdot||_p$ in instead of $||\cdot||_{\ell_p}$.

Problem (The better one)

$$\min_{x\in\mathbf{R}^n}\{||x||_p|Ax=b\} \tag{2}$$

- $Ax = b$ has infinitely many solutions.
- When $p = 2$, $x = A^*(AA^*)^{-1}b$ is the global minimizer[3], however it is not sparse, due to the limitation of ℓ_2 norm.
- When $p = 1$, it can induce sparsity; and under the Robust Uncertainty Principles[4], one can recover the minimizer of $p = 0$ problem from solving (2) with $p = 1$.

[3] proof in my report
[4] E.J.Candes & T.Tao, 05

Outline

Unconstrained Optimization

With parameter λ

With $p = 1$ and a regularization parameter, we have:

Problem (Unconstrained with parameter λ)

$$\min_{x \in \mathbf{R}^n} \{ ||x||_1 + \frac{\lambda}{2} ||b - Ax||_2^2 \} \tag{3}$$

- $\lambda > \frac{1}{||A^* b||_\infty}$ in order to have non-zero optimizer [5].
- As λ increases, x tends to be sparser [6].
- λ is used to reduce overfitting, and it also adds bias to the problem (more emphasis on the least square part).
- Similar formulation appears in Signal Processing (BDN), Statistics (LASSO), Geophysics, etc.

---

[5] J.J.Fuchs, 2004

[6] R. Tibshirani, 1996

Solving (3) is also equivalent to solve this following problem:

Problem (Signum Equation)

$$sgn(x) + \lambda(A^* b - A^* Ax) = 0 \quad (4)$$

- $sgn(a) = \begin{cases} 1, & a > 0 \\ -1, & a < 0 \end{cases}$ for scalars, component wise for vectors.
- It's derived using calculus of variation.
- It's a nonlinear equation, and no closed form solution is found so far.
- It can be solved busing "A Fixed-Point Continuation Method"[7].

[7] Elaine T. Hale, et al, 07

Outline

Disadvantages

We try to address them all

(2) can be defined more generally as:

$$\min_{x \in \mathbf{R}^n} \{J(x) | Ax = b\} \tag{5}$$

Where $J(x)$ is continuous and convex.

- When $J(x)$ is coercive, the set of solutions of (5) is nonempty and convex.
- When $J(x)$ is strictly or strongly convex, then solution of (5) is unique.
- However $J(x) = ||x||_1$ is not strictly convex.
- The solutions of (3) and (4) depend on some wise choice of the parameter λ.

Outline

History
On Image Processing

A Hierarchical Decomposition is used to solve the following problem:

$$\min_{x \in \mathbf{R}^n} \{||x||_U + \frac{\lambda}{2}||b - Tx||_W^p\} \tag{6}$$

Where λ just represents different scales.

- The first 2 papers were published with $U = BV$, $W = \ell_2$, $p = 2$, and $T = I$.[8]
- 3 more papers were published with T being a differential operator.[9]
- 2 more papers afterward with emphasis on $T = \nabla$.[10]
- We'll do it with $U = \ell_1$, $W = \ell_2$, $p = 2$, and $T = A$.

---

[8] E. Tadmor, S. Nezzar & L. Vese, 04 and 08
[9] E. Tadmor & P. Athavale, 09, 10, and 11
[10] E. Tadmor & C. Tan, 10 and preprint

Theorem (Validation Principles)

$$
\begin{aligned}
< x, T^*(b - Tx) > &= ||x||_U \cdot ||T^*(b - Tx)||_{U^*} \\
||T^*(b - Tx)||_{U^*} &= \frac{1}{\lambda}
\end{aligned}
\tag{7}
$$

iff x solves (6)

- λ is the regularization parameter in (3).
- $\langle \cdot, \cdot \rangle$ is an inner product, and U^* is the dual space of U.
- An initial λ_0 is also found as long as it satisfies:

$$
\frac{1}{2} < \lambda_0 ||T^* b||_{U^*} \leq 1 \tag{8}
$$

- An optimal stopping λ_J is also found.[11]

[11] S.Osher, et al. 05

Outline

Data: A and b, pick $\lambda_0(\textit{from}(8))$

Result: $x = \sum_{j=0}^{J} x_j$

Initialize: $r_0 = b$, and $j = 0$;

while *A certain* λ_J *is not found* **do**

$$x_j(\lambda_j, r_j) := arg \min_{x \in \mathbf{R}^n} \{||x||_1 + \frac{\lambda_j}{2}||r_j - Ax||_2^2\};$$

$$r_{j+1} = r_j - Ax_j;$$

$$\lambda_{j+1} = 2 * \lambda_j;$$

$$j = j + 1;$$

end

- $b = A(\sum_{j=0}^{J} x_j) + r_{J+1}$.
- By (7), $||A^* r_{J+1}||_\infty = \frac{1}{\lambda_{J+1}}$.

Examples

Numerical Results[12]

Figure: Increase in the details with successive increment of scales

[12] E. Tadmor, S. Nezzar & L. Vese, 08

Outline

We'll solve $\min_{x \in \mathbf{R}^n} \{\tau ||x||_1 + \frac{1}{2}||b - Ax||_2^2\}$[13] with the following transformation:

- $(a)_+ = \begin{cases} a, & a \geq 0 \\ 0, & \text{otherwise} \end{cases}$
- $u = (x)_+$, $v = (-x)_+$, and $z = \begin{bmatrix} u \\ v \end{bmatrix}$.
- $y = A^* b$ and $c = \tau \mathbf{1}_{2n} + \begin{bmatrix} -y \\ y \end{bmatrix}$.
- $B = \begin{bmatrix} A^*A & -A^*A \\ -A^*A & A^*A \end{bmatrix}$.

Then it becomes: $\min_{z \in \mathbf{R}^{2n}} \{c^* z + \frac{1}{2} z^* B z \equiv F(z) | z \geq 0\}$.

[13] $\tau = \frac{1}{\lambda}$

Data: A, b, τ, and $z^{(0)}$, pick $\beta \in (0,1)$[a] and $\mu \in (0,1/2)$[b]

Result: $z^{(K)}(\tau) := \min_{z \in \mathbf{R}^{2n}} \{c^* z + \frac{1}{2} z^* B z | z \geq 0\}$

Initialize: $k = 0$;

while *A convergence test is not satisfied* **do**

Compute $\alpha_0 = arg\min_{\alpha} F(z^{(k)} - \alpha g^{(k)})$[c];

Let $\alpha^{(k)}$ be the first in the sequence $\alpha_0, \beta\alpha_0, \beta^2\alpha_0, \ldots,$ such that $F((z^{(k)} - \alpha^{(k)} \nabla F(z^{(k)}))_+) \leq F(z^{(k)}) - \mu \nabla F(z^{(k)})^* (z^{(k)} - (z^{(k)} - \alpha^{(k)} \nabla F(z^{(k)}))_+)$ and set $z^{(k+1)} = (z^{(k)} - \alpha^{(k)} \nabla F(z^{(k)}))_+$;

$k = k + 1$;

end

---

[a] β is controlling the step length α_0.
[b] μ is making sure $F(\cdot)$ is decreased sufficiently from Armijo Rule.
[c] $g^{(k)}$ is a projected gradient.

GPSR

Convergence And Computational Efficiency

- The convergence of the algorithm is already shown.[14]
- It is more robust than IST, $\ell_1\_\ell_s$, ℓ_1-magic toolbox, and the homotopy method.
- The Description of the algorithms is clear and easy to implement.
- $Bz = \begin{bmatrix} A^*A(u-v) \\ -A^*A(u-v) \end{bmatrix}$ and $c^*z = \tau \mathbf{1}_n^*(u+v) - y^*(u-v)$
- $z^*Bz = ||A(u-v)||_2$ and $\nabla F(z) = c + Bz$
- Hence the $2n \times 2n$ system can be computed as a $n \times n$ system.
- Ax and A^*x can be defined as function calls instead of direct matrix-vector multiplication in order to save memory allocation.

[14] M.A.T. Figueiredo, et al, 2007

Milestones

Things that I'm working on and will do

- Implement and Validate the GPSR by the end of October of 2012.
- Implement A Fixed-Point Continuation Method by early December of 2012.
- Validate A Fixed-Point Continuation Method by the end of January of 2013.
- Implement the whole Hierarchical Decomposition algorithm by the end of February of 2013.
- Validate the whole Hierarchical Decomposition code by the end of March of 2013.
- Codes will be implemented in Matlab, and validations are provided by (7).
- Deliverables: Matlab codes, presentation slides, complete project,

For Further Reading I

Eitan Tadmor, Suzanne Nezzar, and Luminita Vese
A Multiscale Image Representation Using Hierarchical (BV, $\mathcal{L}^2$) Decompositions.
2004.

Eitan Tadmor, Suzanne Nezzar, and Luminita Vese
Multiscale Hierarchical Decomposition of Images with Applications to Deblurring, Denoising and Segmentation.
2008.

For Further Reading II

- Mario A. T. Figueiredo, Roberg D. Nowak, Stephen J. Wright
  Gradient Projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems.
  2007.

- Elaine T. Hal, Wotao Yin, and Yin Zhang
  A Fixed-Point Continuation Method for ℓ_1-Regularized Minimization with Applications to Compressed Sensing.
  2007.

For Further Reading III

Seung-Jean Kim, K. Koh, M. Lustig, Stephen Boyd, and Dimitry Gorinevsky
An Interior-Point Method for Large Scale ℓ_1-Regularized Least Squares.
2007.

Stanley Osher, Yu Mao, Bin Dong, Wotao Yin
Fast Linearized Bregman Iteration for Compressive Sensing and Sparse Denoising.
2008.