Optimization Methods with Signal Denoising Applications

Paul Tseng
Mathematics, University of Washington
Seattle

Taiwan Normal University
March 2, 2006

(Joint works with Sylvain Sardy (EPFL), Andrew Bruce (MathSoft), and Sangwoon Yun (UW))

Talk Outline

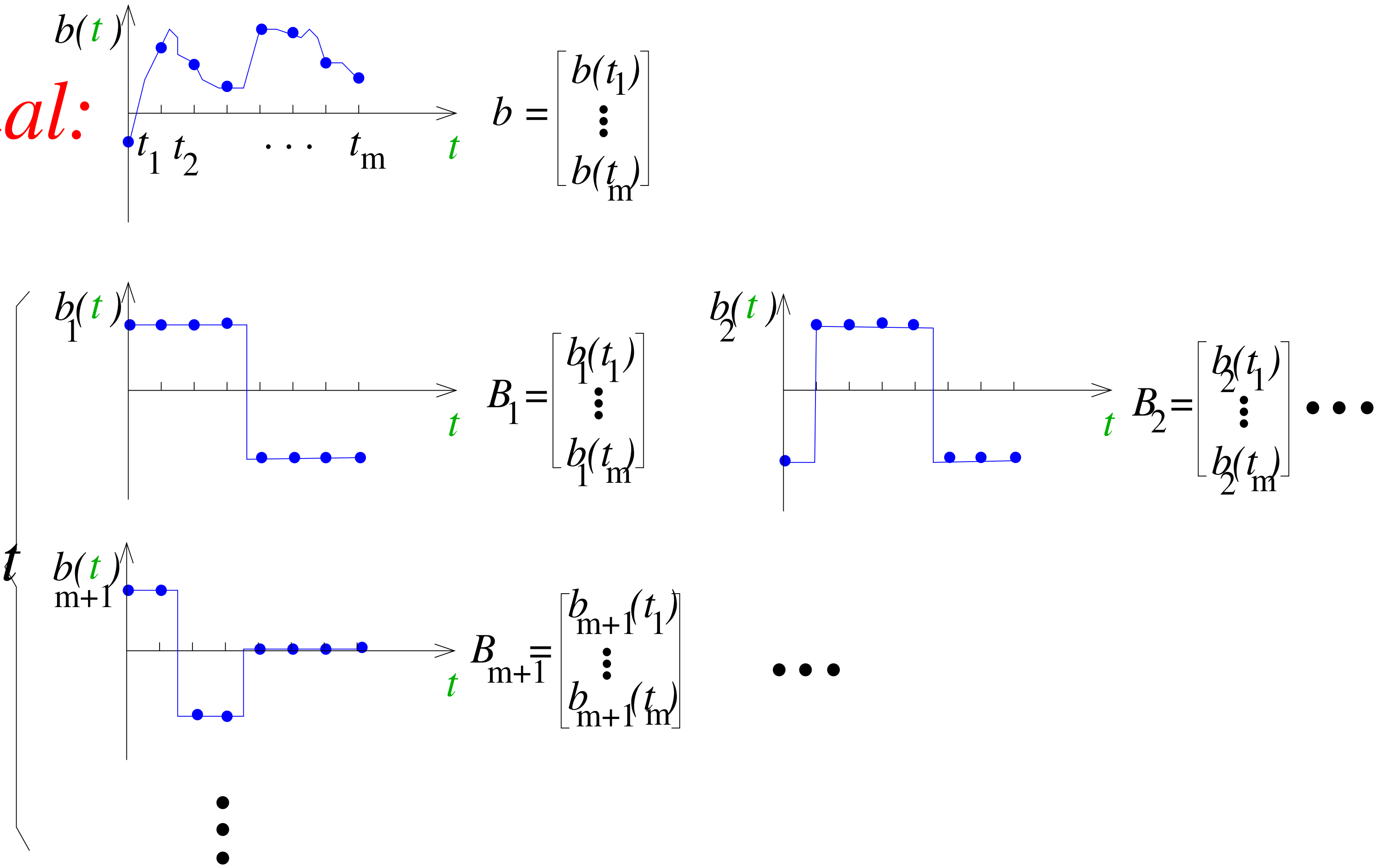
- Basic Problem Model
  - ⋆ Primal-Dual Interior Point Method
  - ⋆ Block Coordinate Minimization Method
  - ⋆ Applications
- General Problem Model
  - ⋆ Block Coordinate Gradient Descent Method
  - ⋆ Convergence
  - ⋆ Numerical Testing (ongoing)
- Conclusions & Future Work

Basic Problem Model

Observed Noisy Signal:

Denoised Signal:

Wavelet basis & its transl.



$$B_1 w_1 + \cdots + B_n w_n = [B_1 \ \cdots \ B_n] \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix} = Bw \qquad (n \geq m)$$

Find w so that $Bw - b \approx 0$ and w has "few" nonzeros.

Formulate this as an unconstrained convex optimization problem:

P1

$$\min_{w \in \Re^n} \|Bw - b\|_2^2 + c\|w\|_1 \qquad (c > 0)$$

"Basis Pursuit"

Chen, Donoho, Saunders

Difficulty: Typically $m \geq 1000$, $n \geq 8000$, and B is dense. $\|\cdot\|_1$ is nonsmooth.

Primal-Dual Interior Point Method for *P1*

Idea: Reformulate *P1* as a convex QP, and apply primal-dual IP method.

QP Reformulation of *P1*:

P1
$$\min_{w \in \Re^n} \|Bw - b\|_2^2 + c\|w\|_1$$

Substitute $w = w^+ - w^-$ with $w^+ \geq 0,\ w^- \geq 0,\quad \|w\|_1 = e^T(w^+ + w^-)$:

$$\min_{\substack{w^+ \geq 0 \\ w^- \geq 0}} \|\underbrace{Bw^+ - Bw^- - b}_{y}\|_2^2 + ce^T(w^+ + w^-)$$

$$\begin{array}{ll} \min & \|y\|_2^2 + ce^T \begin{bmatrix} w^+ \\ w^- \end{bmatrix} \\ {\scriptstyle w^+ \geq 0 \atop w^- \geq 0} & \underbrace{[B\ \ -B]}_{A} \underbrace{\begin{bmatrix} w^+ \\ w^- \end{bmatrix}}_{x} + y = b \end{array}$$

QP Reformulation of *P1*:

$$\begin{array}{ll} \min & \|y\|_2^2 + ce^T x \\ x \geq 0 & Ax + y = b \end{array}$$

KKT Optimality Condition for QP:

$$\begin{aligned} Ax + y &= b, \qquad x \geq 0 \\ A^T y + z &= ce, \qquad z \geq 0 \\ Xz &= 0 \end{aligned} \qquad (X = \mathrm{diag}[x_1, ..., x_{2n}])$$

Perturbed KKT Optimality Condition:

$$\begin{aligned} Ax + y &= b, \qquad x > 0 \\ A^T y + z &= ce, \qquad z > 0 \\ Xz &= \mu e \qquad (\mu > 0) \end{aligned}$$

Primal-Dual IP method: Apply damped Newton method to solve inexactly the perturbed KKT equations while maintaining $x > 0, z > 0$. Decrease μ after each iteration. Fiacco-McCormick '68, Karmarkar '84,...

Method description:

Given $\mu > 0$, $x > 0$, y, $z > 0$, solve

$$\begin{aligned} A\Delta x + \Delta y &= b - Ax - y, \\ A^T\Delta y + \Delta z &= ce - A^T y - z, \\ Z\Delta x + X\Delta z &= \mu e - Xz \end{aligned}$$

Newton Eqs.

Update

$$\begin{aligned} x^{\text{new}} &= x + .99\beta_p \Delta x, \\ y^{\text{new}} &= y + .99\beta_d \Delta y, \\ z^{\text{new}} &= z + .99\beta_d \Delta z, \\ \mu^{\text{new}} &= (1 - \min\{.99, \beta_p, \beta_d\})\mu, \end{aligned}$$

where $\beta_p = \min_{i:\Delta x_i < 0} \left\{ \frac{x_i}{-\Delta x_i} \right\}, \quad \beta_d = \min_{i:\Delta z_i < 0} \left\{ \frac{z_i}{-\Delta z_i} \right\}$

Implementation & Initialization:

- Newton Eqs. reduce to

$$(I + AZ^{-1}XA^T)\Delta y \ = \ r \ .$$

  Solve by Conjugate Gradient (CG) method.

  Multiplication by $\underbrace{A}_{m \times 2n}$ & A^T require $O(m \log m)$ & $O(m(\log m)^2)$ opers.

- Initialization as in Chen-Donoho-Saunders '96
- Theoretical convergence?

  CG preconditioning?

Block Coord. Minimization Method for *P1*

Method description:

Given w, choose $\mathcal{I} \subseteq \{1, ..., n\}$ with $|\mathcal{I}| = m$, $\{B_i\}_{i \in \mathcal{I}}$ is orthog.

Update

$$w^{\text{new}} = \arg \min_{u_i = w_i \ \forall i \notin \mathcal{I}} \|Bu - b\|_2^2 + c\|u\|_1$$

closed

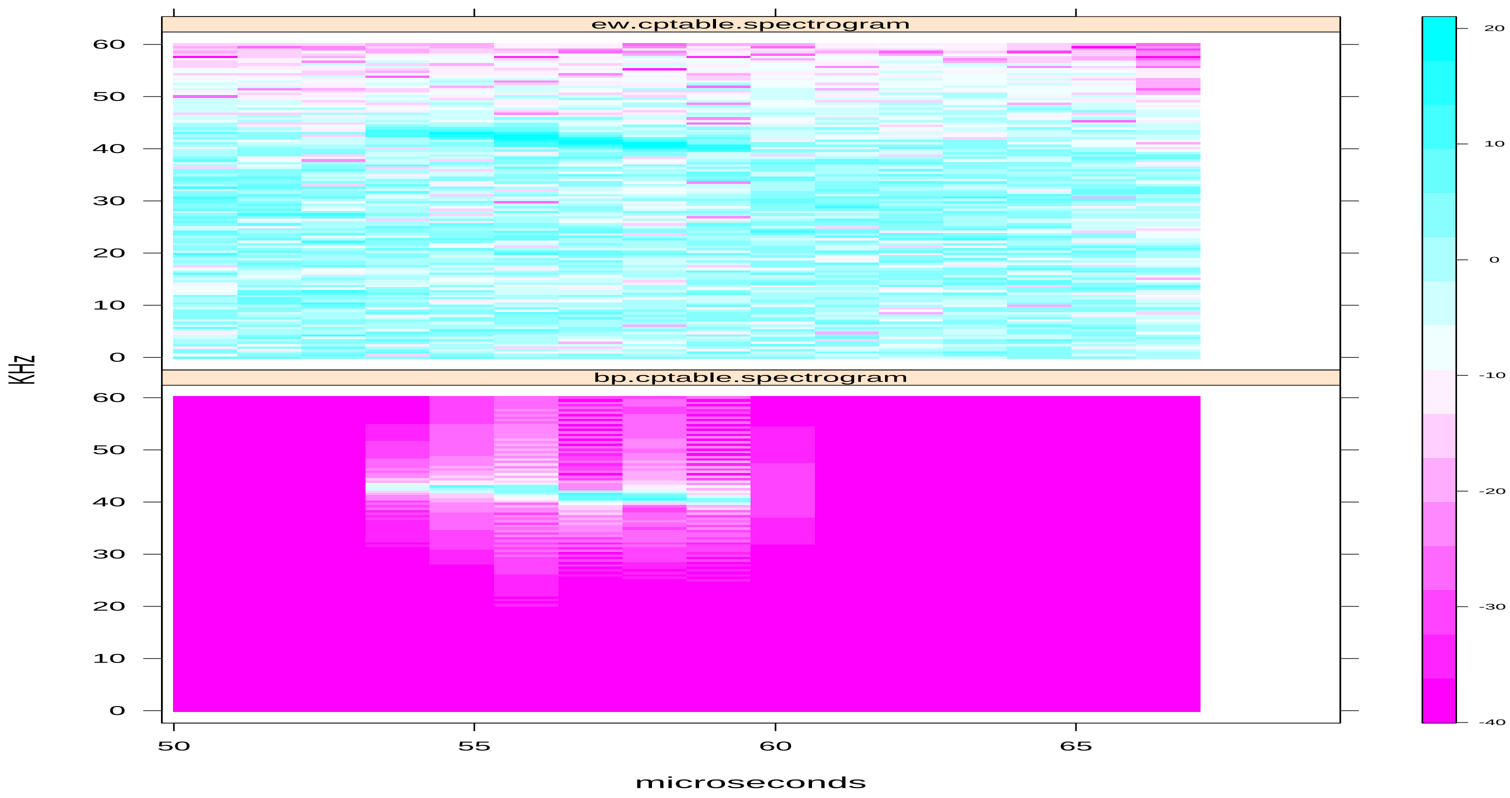
$\longleftarrow$ form soln

- Choose $\mathcal{I}$ to maximize $\displaystyle\min_{v \in \partial_{u_{\mathcal{I}}}(\|Bu-b\|_2^2 + c\|u\|_1)|_{u=w}} \|v\|_2$.
  Requires $O(m \log m)$ opers. by algorithm of Coifman & Wickerhauser.
- Theoretical convergence: w-sequence is bounded & each cluster point solves *P1*.

Convergence of BCM method depends crucially on

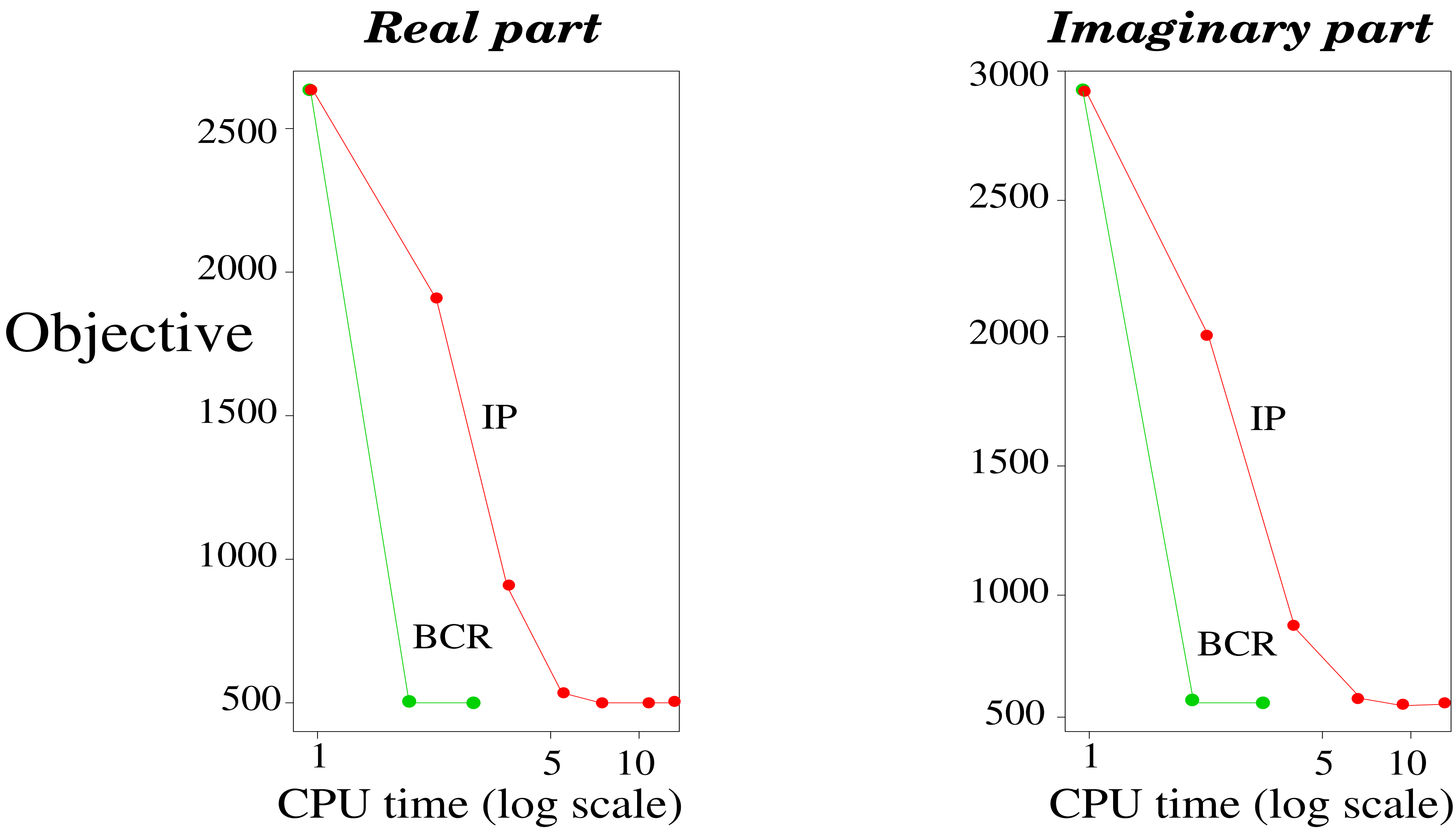
- differentiability of $\|\cdot\|_2^2$
- separability of $\|\cdot\|_1$
- convexity $\Rightarrow$ global minimum

Application 1: Electronic surveillance



$m = 2^{11} = 2048$, $c = 4$, local cosine transform, all but 4 levels

Method efficiency:



Comparing CPU times of IP and BCM methods (S-Plus, Sun Ultra 1).

ML Estimation

P2:

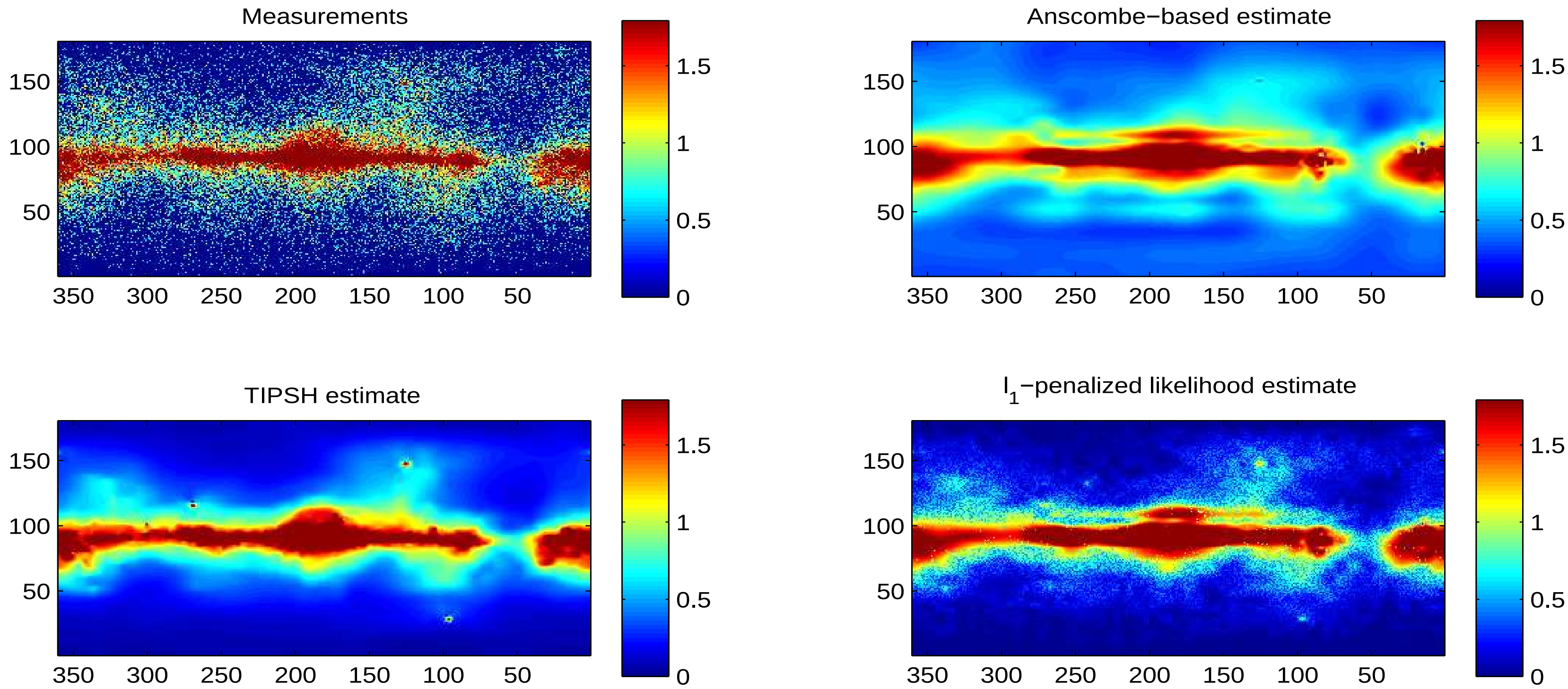
$$\min_{w} -\ell(Bw; b) + c \sum_{i \in \mathcal{J}} |w_i| \qquad (c > 0)$$

ℓ is log likelihood, $\{B_i\}_{i \notin \mathcal{J}}$ are lin. indep "coarse-scale Wavelets"

- $-\ell(y; b) = \frac{1}{2}\|y - b\|_2^2$ Gaussian noise
- $-\ell(y; b) = \sum_{i=1}^{m} (y_i - b_i \ln y_i) \quad (y_i \geq 0)$ Poisson noise

Solve *P2* by adapting IP method. BCM?

Application 2: Γ ray astronomy



$m = 720 \cdot 360$, c chosen by CV, Symmlets of order 8 (levels 3-8). Spatially inhomogeneous Poisson noise.

But IP method is slow (many CG steps). ☹

Adapt BCM method?

General Problem Model

P3

$$\min_{w} F_c(w) := f(w) + cP(w) \qquad (c \geq 0)$$

$f : \Re^N \rightarrow \Re$ is smooth.

$P : \Re^N \rightarrow (-\infty, \infty]$ is proper, convex, lsc, and $P(w) = \sum_{j=1}^{n} P_j(w_j)$ ($w = (w_1, ..., w_n)$).

- $P(w) = \|w\|_1$
- $P(w) = \begin{cases} 0 & \text{if } l \leq w \leq u \\ \infty & \text{else} \end{cases}$

Block Coord. Gradient Descent Method for *P3*

Idea: Do BCM on a quadratic approx. of f.

For $w \in \mathrm{dom}P$, $\mathcal{I} \subseteq \{1, ..., n\}$, and $H \succ 0_n$, let $d_H(w; \mathcal{I})$ and $q_H(w; \mathcal{I})$ be the optimal soln and obj. value of

$$\boxed{\min_{d|d_i=0\ \forall i \notin \mathcal{I}} \{g^T d + \frac{1}{2} d^T H d + cP(w+d) - cP(w)\}}$$

direc. subprob

with $g = \nabla f(w)$.

Facts:

- $d_H(w; \{1, ..., n\}) = 0 \ \Leftrightarrow F_c'(w; d) \geq 0 \ \forall d \in \Re^N$. stationarity
- H is diagonal $\Rightarrow d_H(w; \mathcal{I}) = \sum_{i \in \mathcal{I}} d_H(w; i)$, $q_H(w; \mathcal{I}) = \sum_{i \in \mathcal{I}} q_H(w; i)$. separab.

Method description:

Given $w \in \mathrm{dom} P$, choose $\mathcal{I} \subseteq \{1, ..., n\}$, $H \succ 0_n$. Let $d = d_H(w; \mathcal{I})$.

Update

$$w^{\text{new}} = w + \alpha d \qquad (\alpha > 0)$$

- α = largest element of $\{1, \beta, \beta^2, ...\}$ satisfying
  $F_c(w + \alpha d) - F_c(w) \leq \sigma \alpha q_H(w; \mathcal{I})$ ($0 < \beta < 1$, $0 < \sigma < 1$) Armijo

- $\mathcal{I} = \{1\}, \{2\}, ..., \{n\}, \{1\}, \{2\}, ...$ Gauss-Seidel

- $\|d_D(w; \mathcal{I})\|_\infty \geq \upsilon \|d_D(w; \{1, ..., n\})\|_\infty$ ($0 < \upsilon \leq 1$, $D \succ 0_n$ is diagonal, e.g., $D = I$ or $D = \mathrm{diag}(H)$). Gauss-Southwell-d

- $q_D(w; \mathcal{I}) \leq \upsilon \, q_D(w; \{1, ..., n\})$. Gauss-Southwell-$q$

Convergence Results: (a) If

- $0 < \underline{\lambda} \leq \lambda_i(D), \lambda_i(H) \leq \bar{\lambda} \; \forall i$,
- α is chosen by Armijo rule,
- $\mathcal{I}$ is chosen by G-Seidel or G-Southwell-d or G-Southwell-q,

then every cluster point of the w-sequence generated by BCGD method is a stationary point of F_c.

(b) If in addition P and f satisfy any of the following assumptions, then the w-sequence converges at R-linear rate (excepting G-Southwell-d).

C1 f is strongly convex, ∇f is Lipschitz cont. on $\mathrm{dom} P$.

C2 f is (nonconvex) quadratic. P is polyhedral.

C3 $f(w) = g(Ew) + q^T w$, where $E \in \Re^{m \times N}$, $q \in \Re^N$, g is strongly convex, ∇g is Lipschitz cont. on $\Re^m$. P is polyhedral.

C4 $f(w) = \max_{y \in Y}\{(Ew)^T y - g(y)\} + q^T w$, where $Y \subseteq \Re^m$ is polyhedral, $E \in \Re^{m \times N}$, $q \in \Re^N$, g is strongly convex, ∇g is Lipschitz cont. on $\Re^m$. P is polyhedral.

Notes:

- BCGD has stronger global convergence property (and cheaper iteration) than BCM.

- Proof of (b) uses a local error bound on $\mathrm{dist}\,(w, \{\mathrm{stat.\ pts.\ of}\ F_c\})$.

Numerical Testing (ongoing):

- Implement BCGD method in Matlab.

- Numerical tests with f from Moré-Garbow-Hillstrom set and CUTEr set (Gould, Orban, Toint '05), $P(w) = \|w\|_1$, and different c (e.g., $c = .1, 1, 10$).

- Comparison with MINOS 5.5.1 (Murtagh, Saunders '05), a Fortran implementation of an active-set method, applied to a reformulation of *P3* with $P(w) = \|w\|_1$ as

$$\min_{\substack{w^+ \geq 0 \\ w^- \geq 0}} f(w^+ - w^-) + c\, e^T(w^+ + w^-).$$

- Preliminary results are "promising".

$f(w)$	n	Description
BAL	1000	Brown almost-linear func, nonconvex, dense Hessian.
BT	1000	Broyden tridiagonal func, nonconvex, sparse Hessian.
DBV	1000	Discrete boundary value func, nonconvex, sparse Hessian.
EPS	1000	Extended Powell singular func, convex, 4-block diag. Hessian.
ER	1000	Extended Rosenbrook func, nonconvex, 2-block diag. Hessian.
QD1	1000	$f(w) = \left(\sum_{i=1}^{n} w_i - 1\right)^2$, convex, quad., rank-1 Hessian.
QD2	1000	$f(w) = \sum_{i=1}^{n} \left(w_i - \frac{2}{n+1}\sum_{j=1}^{n} w_j - 1\right)^2 + \left(\frac{2}{n+1}\sum_{j=1}^{n} w_j + 1\right)^2$, strongly convex, quad., dense Hessian.
VD	1000	$f(w) = \sum_{i=1}^{n} (w_i - 1)^2 + \left(\sum_{j=1}^{n} j(w_j - 1)\right)^2 + \left(\sum_{j=1}^{n} j(w_j - 1)\right)^4$, strongly convex, dense ill-conditioned Hessian.

Table 1: Least square problems from Moré, Garbow, Hillstrom, 1981

		MINOS	BCGD-G-Southwell-d	BCGD-G-Southwell-q
$f(w)$	c	♯nz/objec/cpu	♯nz/objec/cpu	♯nz/objec/cpu
BAL	1	1000/1000/43.9	1000/1000/.1	1000/1000/.1
	10	1000/9999.9/43.9	1000/9999.9/.1	1000/9999.9/.1
	100	1000/99997.5/44.3	1000/99997.5/.1	1000/99997.5/.2
BT	.1	1000/71.725/134.4	999/71.394/4.5	999/71.394/5.0
	1	999/672.41/95.3	21/672.70/292.6	995/991.06/1.3(?)
	10	0/1000/77.7	0/1000/.01	0/1000/.01
DBV	.1	0/0/52.7	0/4.5E-9/.1	0/4.5E-9/.04
	1	0/0/52.9	0/4.5E-9/.1	0/4.5E-9/.04
	10	0/0/53.0	0/4.5E-9/.01	0/4.5E-9/.01
EPS	1	1000/351.14/58.5	500/351.14/.3	500/351.14/.3
	10	243/1250/45.7	250/1250/.05	250/1250/.05
	100	0/1250/50.7	0/1250/.01	0/1250/.02
ER	1	1000/436.25/72.0	1000/436.25/.5	1000/436.25/.4
	10	0/500/51.5	0/500/.1	0/500/.01
	100	0/500/52.4	0/500/.03	0/500/.01
QD1	.1	1000/.0975/29.9	1/.0975/.01	1/.0975/.02
	1	1000/.75/37.8	1/.75/.01	1/.75/.01
	10	0/1/38.6	0/1/.01	0/1/.01
QD2	.1	1000/98.5/74.2	0/98.5/.01	0/98.5/.03
	1	1000/751/75.8	0/751/.01	0/751/.02
	10	0/1001/53.1	0/1001/.01	0/1001/.01
VD	1	1000/937.59/43.9	1000/937.66/856.3	1000/937.66/869.0
	10	413/6726.80/57.1	1000/6746.74/235.7	999/6746.74/246.9
	100	136/55043/57.8	1000/55078/12.6	1000/55078/13.3

Table 2: Performance of MINOS, BCGD-Gauss-Southwell-d/q, with $w^{\text{init}} = (1, 1, \ldots, 1)$

Conclusions & Future Work

1. For ML estimation, ℓ_1-penalty imparts parsimony in the coefficients and avoid oversmoothing the signals.

2. The resulting estimation problem can be solved effectively by IP method or BCM method, exploiting the problem structure, including nondiffer. of ℓ_1-norm. Which to use? Depends on problem.

3. Problem reformulation may be needed.

4. For general problem model, we propose BCGD method. Numerical testing is ongoing.

5. Applications to denoising, regression, SVM?