# Block-Coordinatewise Methods for Sparse Optimization with Nonsmooth Regularization

Paul Tseng
Mathematics, University of Washington
Seattle

Hong Kong Polytechnic University
February 20, 2008
(Joint works with Sylvain Sardy (Univ. Geneva) and Sangwoon Yun (NUS))

# Talk Outline

- Block-Coordinate Minimization
  - ⋆ Properties
  - ⋆ Application I: Basis Pursuit/Lasso
- Block-Coordinate Gradient Descent
  - ⋆ Properties
  - ⋆ Application II: Group Lasso for Logistic Regression
  - ⋆ Application III: Sparse Inverse Covariance Estimation
- Conclusions & Future Work

# Block-Coordinate Minimization

$$\min_{x=(x_1,\ldots,x_n)} f(x)$$

$f : \Re^n \to \Re$ is cont. diff.

Given $x \in \Re^n$, choose $i \in \{1, ..., n\}$. Update

$$\boxed{x^{\text{new}} = \mathop{\arg\min}_{u \mid u_j = x_j \ \forall j \neq i} f(u).}$$

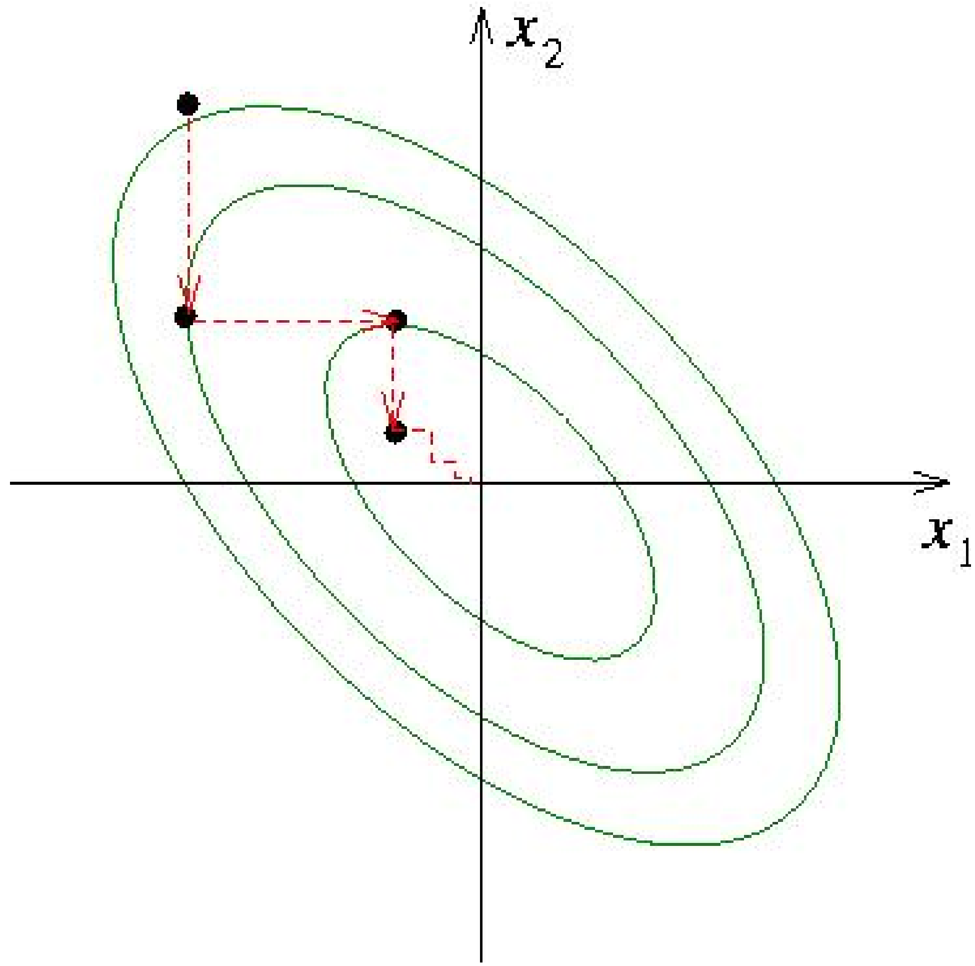
Repeat until “convergence”.

Gauss-Seidel: Choose $i$ cyclically, 1, 2,..., $n$, 1, 2,...

Gauss-Southwell: Choose $i$ with $|\frac{\partial f}{\partial x_i}(x)|$ maximum.

Example:

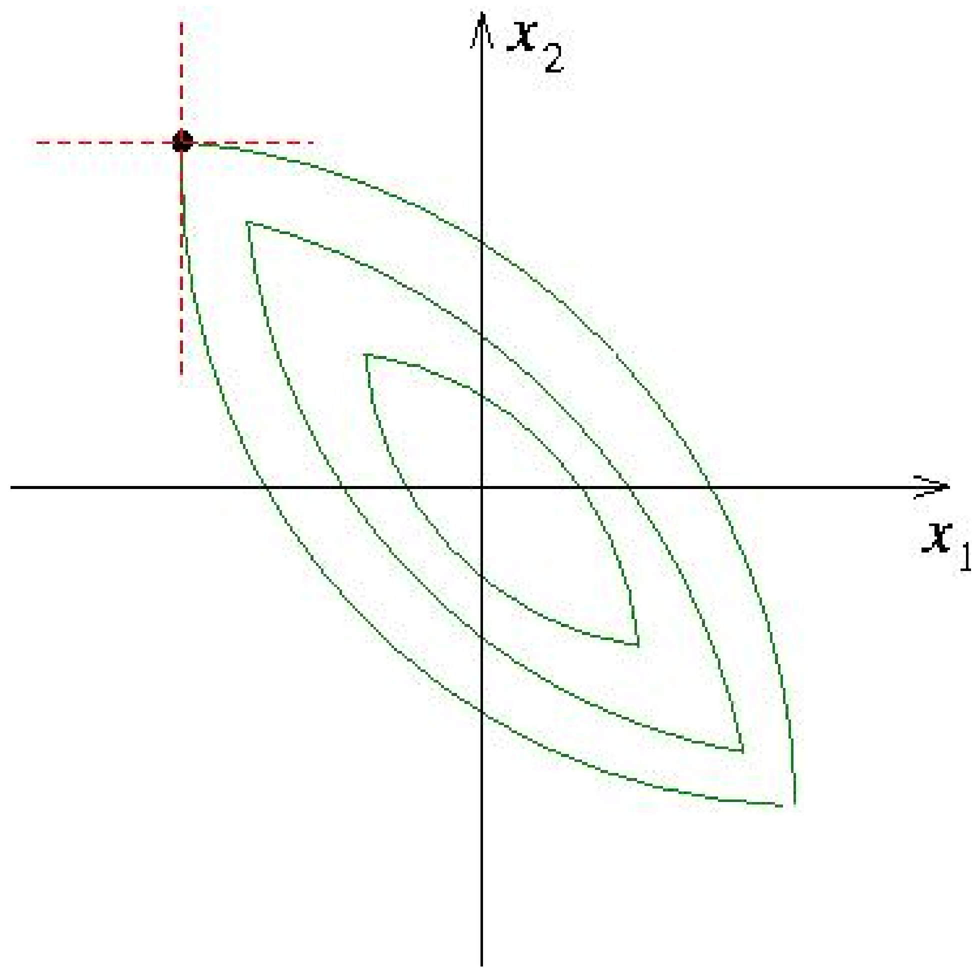
$$\min_{x=(x_1,x_2)} (x_1+x_2)^2 + \frac{1}{4}(x_1-x_2)^2$$



## Properties:

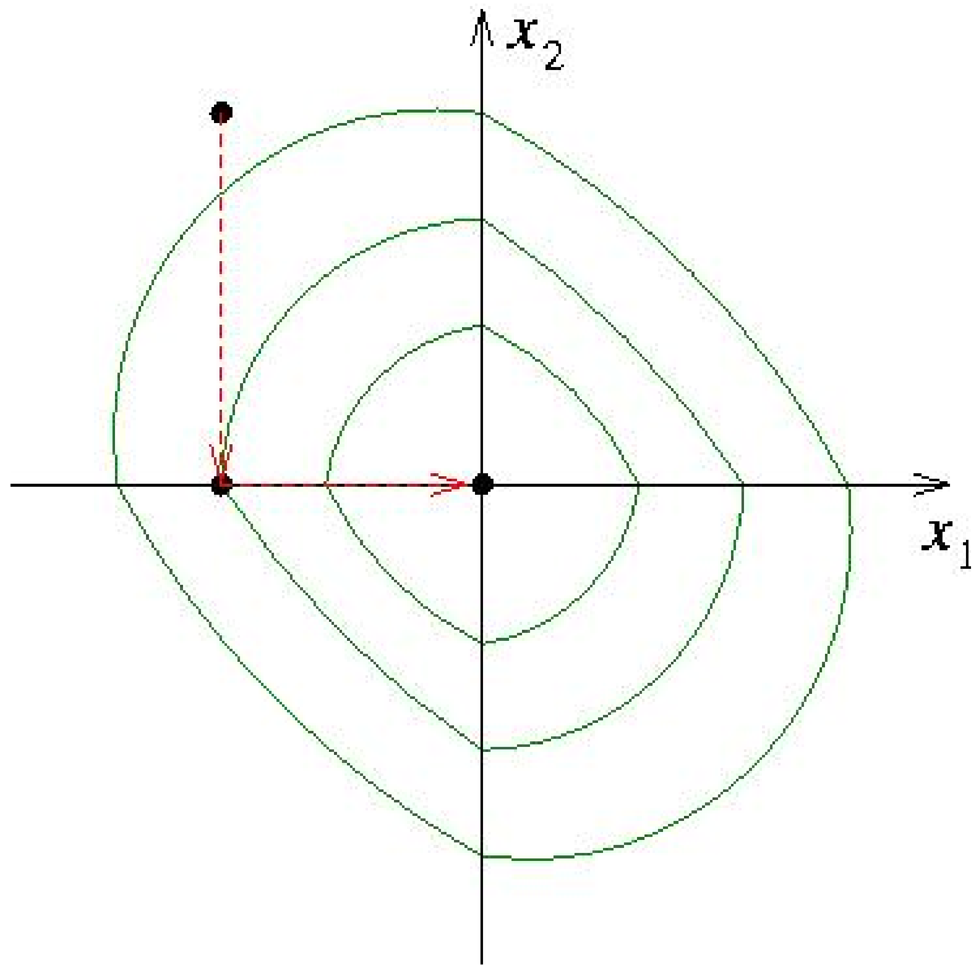
- This method extends to update a block of coordinates at each iteration.
- It is simple, and efficient for large “weakly coupled” problems (off-block-diagonals of $\nabla^2 f(x)$ not too large).
- If $f$ is convex, then very cluster point of the $x$-sequence is a minimizer. Zadeh ’70
- If $f$ is nonconvex, then G-Seidel can cycle Powell ’73 but G-Southwell still converges.
- Can get stuck if $f$ is nondifferentiable.

Example: $$\min_{x=(x_1,x_2)} (x_1+x_2)^2 + \frac{1}{4}(x_1-x_2)^2 + |x_1+x_2|$$



But, if the nondifferentiable part is *separable*, then convergence is possible.

Example:
$$\min_{x=(x_1,x_2)} (x_1+x_2)^2 + \frac{1}{4}(x_1-x_2)^2 + |x_1| + |x_2|$$



# Application I: Basis Pursuit/Lasso

$$\boxed{\min_x F_c(x) := \|Ax - b\|_2^2 + c\|x\|_1}$$

Tibshirani '96, Fu '98

Osborne et al. '98

Chen, Donoho, Saunders '99

...

$A \in \Re^{m \times n}$, $b \in \Re^m$, $c \geq 0$.

- Typically $m \geq 1000$, $n \geq 8000$, and $A$ is dense. $\|\cdot\|_1$ is nonsmooth.
- Can reformulate this as a convex QP and solve using an IP method. Chen, Donoho, Saunders '99

Assume the columns of $A$ come from an overcomplete set of basis functions associated with a fast transform (e.g., wavelet packets).

## BCM for BP:

Given $x$, choose $\mathcal{I} \subseteq \{1, ..., n\}$ with $|\mathcal{I}| = m$ and $\{A_i\}_{i \in \mathcal{I}}$ orthog. Update

$$\boxed{x^{\text{new}} = \arg\min_{u_i = x_i \ \forall i \notin \mathcal{I}} F_c(u)}$$

has closed

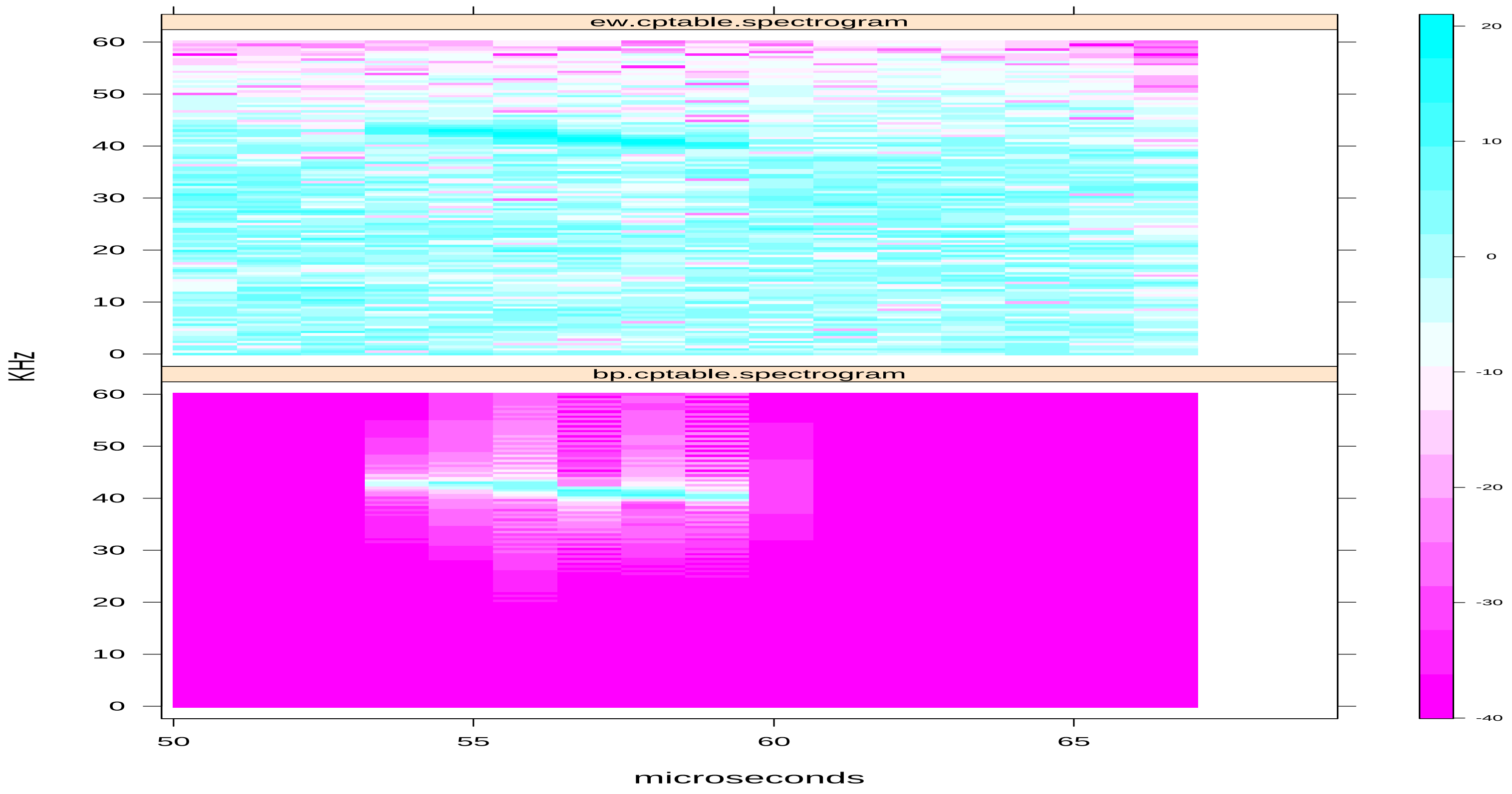
$\longleftarrow$ form soln

Repeat until "convergence".

Gauss-Southwell: Choose $\mathcal{I}$ to maximize $\min_{v \in \partial_{x_\mathcal{I}} F_c(x)} \|v\|_2$.

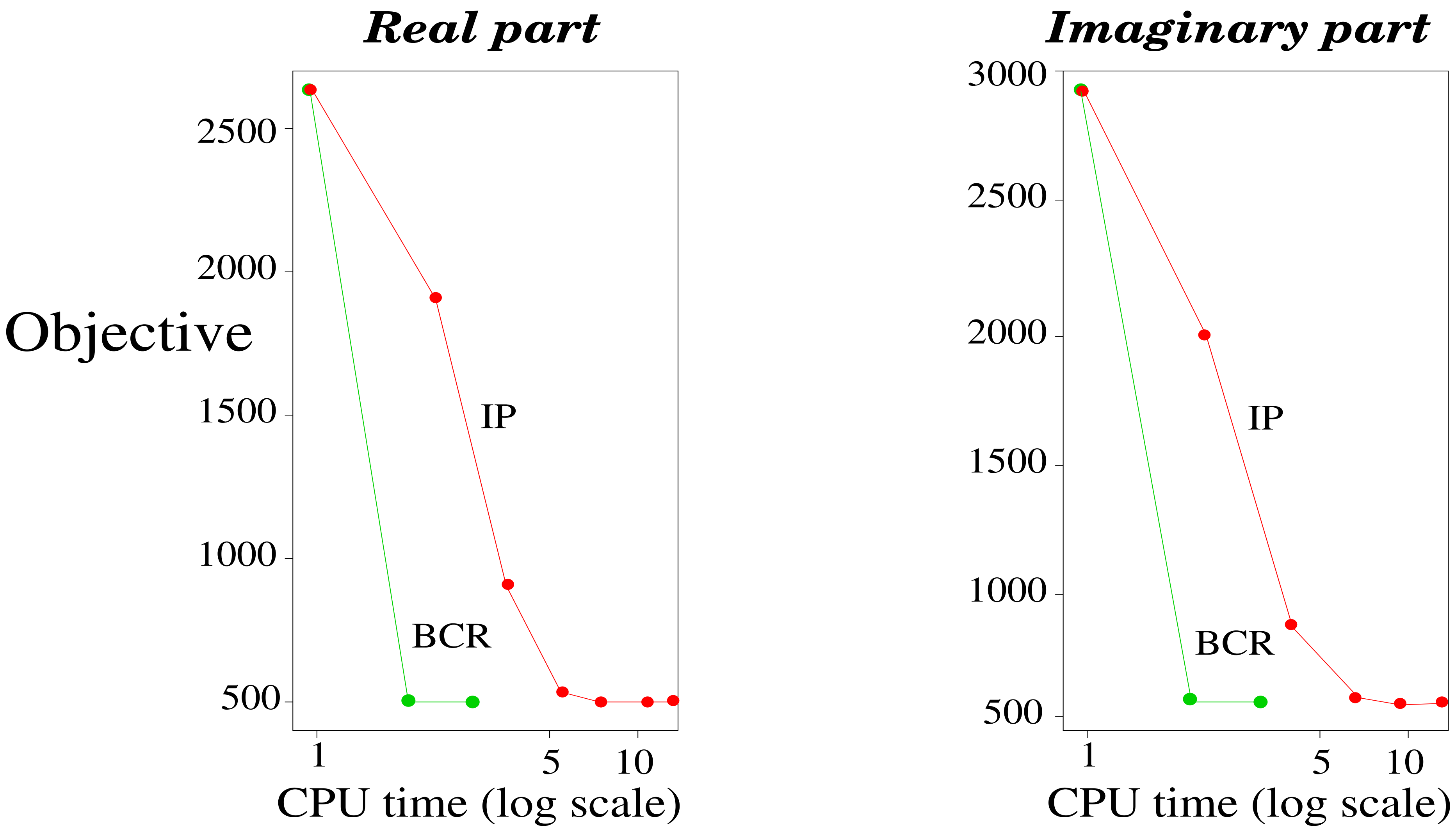
- Finds $\mathcal{I}$ in $O(n + m \log m)$ opers. by algorithm of Coifman & Wickerhauser.
- $x$-sequence is bounded & each cluster point minimizes $F_c$. Sardy, Bruce, T '00

# Electronic surveillance:




$m = 2^{11} = 2048$, $c = 4$, $b$: top image, $A$: local cosine transform, all but 4 levels

# Method efficiency:




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

- IP method requires fewer iterations, but each iteration is expensive (many CG steps per iteration). No good preconditioner for CG is known.

- BCM method requires more iterations, but each iteration is cheap.

- Convergence of BCM depends crucially on

  - differentiability of $\|\cdot\|_2^2$

  - separability of $\|\cdot\|_1$

  - convexity of $F_c$ (stationarity $\Rightarrow$ global minimum)

# Generalization to ML Estimation with $\ell_1$-Regularization?

$$\boxed{\min_x -\ell(Ax; b) + c \sum_{i \in \mathcal{J}} |x_i|}$$

$\ell$ is log likelihood, $\{A_i\}_{i \notin \mathcal{J}}$ are lin. indep "coarse-scale Wavelets", $c \geq 0$

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

Can solve this problem by adapting IP method. But IP method is slow (many CG steps per IP iteration) Antoniadis, Sardy, T '04.

Adapt BCM method?

# Optimization with Nonsmooth Regularization

$$\boxed{\min_x F_c(x) := f(x) + cP(x)}$$

$f : \Re^n \to \Re$ is cont. diff. $\quad c \geq 0$.

$P : \Re^n \to (-\infty, \infty]$ is proper, convex, lsc, and block-separable, i.e.,
$P(x) = \sum_{\mathcal{I} \in \mathcal{C}} P_{\mathcal{I}}(x_{\mathcal{I}})$ $\qquad$ ($\mathcal{I} \in \mathcal{C}$ partition $\{1, ..., n\}$).

- $P(x) = \|x\|_1$ — Basis Pursuit/Lasso
- $P(x) = \sum_{\mathcal{I} \in \mathcal{C}} \|x_{\mathcal{I}}\|_2$ — group Lasso
- $P(x) = \begin{cases} 0 & \text{if } l \leq x \leq u \\ \infty & \text{else} \end{cases}$ — bound constrained NLP

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

# Block-Coord. Gradient Descent Method

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

$$\boxed{\min_{d \mid d_i = 0 \ \forall i \notin \mathcal{I}} \{\nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) - cP(x)\}}$$

direc. subprob

Facts:

- $d_H(x; \{1, ..., n\}) = 0 \ \Leftrightarrow F_c'(x; d) \geq 0 \ \forall d \in \Re^n$. stationarity
- $H$ is diagonal, $P$ is separable $\Rightarrow d_H(x; \mathcal{I}) = \sum_{i \in \mathcal{I}} d_H(x; i)$, $q_H(x; \mathcal{I}) = \sum_{i \in \mathcal{I}} q_H(x; i)$. separab.
- $H$ is diagonal, $P$ is “simple” $\Rightarrow d_H(x; \mathcal{I})$ has “closed form”.

Given $x \in \mathrm{dom}P$, choose $\mathcal{I} \subseteq \{1, ..., n\}$, $H \succ 0$.

Update
$$\boxed{x^{\mathrm{new}} = x + \alpha d_H(x; \mathcal{I}) \qquad (\alpha > 0)}$$

until "convergence."

Gauss-Seidel: Choose $\mathcal{I} \in \mathcal{C}$ cyclically.

Gauss-Southwell: Choose $\mathcal{I}$ with

$$q_D(x; \mathcal{I}) \leq v\, q_D(x; \{1, ..., n\})$$

($0 < v \leq 1$, $D \succ 0$ is diagonal, e.g., $D = I$ or $D = \mathrm{diag}(H)$).

Inexact Armijo LS: $\alpha$ = largest element of $\{1, \beta, \beta^2, ...\}$ satisfying

$$F_c(x + \alpha d) - F_c(x) \;\leq\; 0.1\, \alpha\, q_H(x; \mathcal{I}) \qquad (0 < \beta < 1)$$

## Convergence properties T, Yun '06:

(a) If $\underline{\lambda} I \preceq D, H \preceq \bar{\lambda} I$ ($0 < \underline{\lambda} \leq \bar{\lambda}$), then every cluster point of the $x$-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 $x$-sequence converges linearly in the root sense.

**A1** $f$ is strongly convex, $\nabla f$ is Lipschitz cont. on $\mathrm{dom} P$.

**A2** $f$ is (nonconvex) quadratic. $P$ is polyhedral.

**A3** $f(x) = g(Ax) + q^T x$, where $A \in \Re^{m \times n}$, $q \in \Re^n$, $g$ is strongly convex, $\nabla g$ is Lipschitz cont. on $\Re^m$. $P$ is polyhedral.

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

# Application II: Group Lasso for Logistic Regression

$$\boxed{\min_x f(x) + c \sum_{\mathcal{I} \in \mathcal{C}} \omega_{\mathcal{I}} \|x_{\mathcal{I}}\|_2}$$

Yuan, Lin '06

Kim[3] '06

Meier, van de Geer, Bühlmann '06

...

$c > 0$, $\omega_{\mathcal{I}} > 0$.

$f(x) = \sum_{j=1}^N \log\left(1 + e^{a_j^T x}\right) - y_j a_j^T x \quad (a_j \in \Re^n$, $y_j \in \{0, 1\})$

- $f$ is convex, cont. diff. $\|\cdot\|_2$ is convex, nonsmooth. In prediction of short DNA motifs, $n > 1000$, $N > 11,000$.

- BCM-GSeidel has been used Yuan, Lin '06, but each iteration is expensive. Every cluster point of the $x$-sequence is a minimizer T '01.

- BCGD-GSeidel is significantly more efficient Meier et al '06. Every cluster point of the $x$-sequence is a minimizer T, Yun '06. Linear convergence?

# Application III: Sparse Inverse Covariance Estimation

$$\boxed{\min_{X \in \mathcal{S}_+^n} f(X) + c\|X\|_1}$$

Meinshausen, Bühlmann '06

Yuan, Lin '07

Banerjee, El Ghaoui, d'Aspremont '07

Friedman, Hastie, Tibshirani '07

$c > 0$, $\|X\|_1 = \sum_{ij} |X_{ij}|$,
$f(X) = -\log \det X + \mathrm{tr}(XS)$ ($S \in \mathcal{S}_+^n$ is empirical covariance matrix)

- $f$ is strictly convex, cont. diff. on its domain, $O(n^3)$ ops to evaluate. $\|\cdot\|_1$ is convex, nonsmooth. In applications, $n$ can exceed $6000$.

The Fenchel dual problem Rockafellar '70 is a bound-constrained convex program:

$$\boxed{\min_{W \in \mathcal{S}_+^n, \|W-S\|_\infty \le c} -\log\det(W)}$$

$\|Y\|_\infty = \max_{ij} |Y_{ij}|$.

• IP method requires $O(n^6 \log(1/\epsilon))$ ops to find $\epsilon$-optimal soln. Impractical! Nesterov's first-order smoothing method requires $O(n^{4.5}/\epsilon)$ ops Banerjee et al '07.

• Use BCM-GSeidel to solve the dual problem, cycling thru columns $i = 1, ..., n$ of $W$. Each iteration reduces (via determinant property & duality) to

$$\min_{\xi \in \Re^{n-1}} \frac{1}{2}\xi^T W_{i\neg i \neg}\xi - S_{i\neg i}^T \xi + c\|\xi\|_1.$$

Solve this using IP method ($O(n^3)$ ops) Banerjee et al '07 or BCM-GSeidel Friedman et al '07.

• Can apply BCGD-GSeidel to either primal or dual problem. More efficient? Applied to the primal, each iteration entails

$$\min_{u \in \Re^n} \left\{ \mathrm{tr}((-X^{-1} + S)D) + \frac{1}{2}u^T H u + c\|X + D\|_1 \right\}_{D = u^T e_i + e_i u^T}.$$

For diagonal $H$, the minimizing $D$ has closed form! For each trial $\alpha$ in the Armijo LS, $\det(X + \alpha D)$ can be evaluated from $\det X$ and $X^{-1}$ in $O(n^2)$ ops. Update $X^{-1}$ in $O(n^2)$ ops. Similar application to the dual. Global convergence, asymptotic linear convergence, complexity analysis... Toh, T, Yun, forthcoming.

# Conclusions & Future Work

1. Nonsmooth regularization induces sparsity in the solution, avoids oversmoothing signals, and is useful for variable selection.

2. The regularized problem can be solved effectively by BCM or BCGD, exploiting the problem structure.

3. Extension of BCM, BCGD to handle linear constraints $Ax = b$ is possible, including Support Vector Machine training T, Yun, '07, '08. Some open questions on efficient implementation and convergence analysis remain.

4. Many other applications, including stochastic volatility models Neto, Sardy, T, forthcoming.

5. Extension of BCGD to nonconvex nonsmooth regularization is possible (e.g. $\ell_p$-regularization, $0 < p < 1$) Sardy, T, forthcoming.