# Inexact Coordinate Descent Method for Nonsmooth Separable Minimization

Sangwoon Yun
Department of Mathematics
University of Washington
Seattle, WA 98195, U.S.A.

Paul Tseng
Department of Mathematics
University of Washington
Seattle, WA 98195, U.S.A.

May 16, 2005

# Outline:

1. A Problem Overview

2. Inexact (block) Coordinate Descent Method

3. Convergence Analysis

4. Numerical Experience

5. Conclusion/Future work

## A Problem Overview:

$\diamond$ **An old problem in optimization**

$$Ax \approx b,$$

where $A \in \Re^{m \times n}$ and $b \in \Re^m$ are given.

$\diamond$ **Linear least square problem**

$$\min_{x} \quad \|Ax - b\|_2^2,$$

$\diamond$ **Box-constrained convex QP**

$$\min_{l \leq x \leq u} \quad \|Ax - b\|_2^2,$$

## $\diamond$ $\ell_1$- **regularization**

$$\min_x \quad \|Ax - b\|_2^2 + c\|x\|_1,$$

where $c > 0$ is a user chosen regularization parameter.

$\circ$ Recent interests have focussed on finding solutions that are parsimonious/sparse.

$\circ$ Ex: the "Basis Pursuit" model for signal denoising.

## $\diamond$ Structured Nonsmooth Optimization

$\circ$ Objective function is *sum of smooth func and nonsmooth separable (convex polyhedral) func.*

$$\min_x f(x) + cP(x), \quad P(x) = \sum_j P_j(x_j)$$

where $c > 0$, $f$ smooth, $P$ nonsmooth, (epi$P = \{(x, \zeta) \mid P(x) \le \zeta\}$ is a polyhedral set).

$\circ$ Ex1: box-constrained QP

$$f(x) = \|Ax - b\|_2^2,$$

$$P(x) = \begin{cases} 0 & \text{if } l \le x \le u \\ \infty & \text{else.} \end{cases}$$

$\circ$ Ex2: $\ell_1$- regularization

$$f(x) = \|Ax - b\|_2^2, \qquad P(x) = \|x\|_1$$

## Inexact(block)Coordinate Descent Method

### ◇ Decent Direction

∘ Choose $J(\neq \emptyset) \subset \{1, ..., n\}$,
sym pd $H \in \Re^{n\times n}$

∘ Solve:

$$\begin{array}{ll} \min\limits_{d} & \nabla f(x)^T d + \frac{1}{2} d^T H d + cP(x+d) \\ \text{s.t.} & d_j = 0 \;\forall j \notin J, \end{array}$$

Using the convexity of $P$, it can be seen that

$$(f+cP)(x+\alpha d) \leq (f+cP)(x) - \alpha \frac{1}{2} d^T H d + o(\alpha),$$

for $0 < \alpha < 1$, whenever $d \neq 0$.

- if $J = \{1, ..., n\}$, $P(x) = \begin{cases} 0 & \text{if } l \leq x \leq u \\ \infty & \text{else.} \end{cases}$, then $d$ is a scaled gradient-projection direction for box-constrained minimization;
- if $f$ is quadratic, $H = \nabla^2 f(x)$, then $d$ is a (block) coordinate minimization direction.

## $\diamond$ Stepsize

$\circ$ Choose a stepsize $\alpha$ so that $x^{\text{new}} = x + \alpha d$ achieves sufficient descent.

Armijo rule:

Choose $\alpha$ to be the largest element of $\{\alpha_{\text{init}}\beta^k\}_{k=0,1,\ldots}$ satisfying

$$(f + cP)(x + \alpha d) \leq (f + cP)(x) - \alpha\sigma d^T H d,$$

where $0 < \beta < 1$, $0 < \sigma < \frac{1}{2}$, and $\alpha_{\text{init}} > 0$. This rule, like that for SQP, requires only function evaluations.

$\circ$ By choosing $\alpha_{\text{init}}$ based on previous stepsizes, the number of evaluations can be kept small.

## $\diamond$ Choose J

$\circ$ Gauss-Seidel

$J$ cycles through $\{1\}, \{2\}, ..., \{n\}$ or, more generally, $J$ collectively covers $1, 2, ..., n$ for every fixed number of consecutive iterations.

$\circ$ Gauss-Southwell

Owing to the convex separable nature of $P$, "natural" residual:

$R(x) = (R(x)_j)_{j=1}^n$,

$$R(x)_j = \arg\min_{d_j} \; g_j^T d_j + \frac{1}{2} d_j^T H_{jj} d_j + c P_j(x_j + d_j)$$

where $g = (g_j)_{j=1}^n$, $g = \nabla f(x)$.

Choose $j$ to satisfy

$$\|R(x)_j\|_\infty \geq \omega \|R(x)\|_\infty, \; 0 < \omega \leq 1$$

- Ex: $(H = I)$

  - $P \equiv 0$, $R(x)_j = -g_j$.

  - $P(x) = \begin{cases} 0 & \text{if } l \le x \le u \\ \infty & \text{else} \end{cases}$,
    $R(x)_j = \text{median}\{l_j - x_j, -g_j, u_j - x_j\}$.

  - $P(x) = \|x\|_1$,
    $R(x)_j = -\text{median}\{g_j - c, x_j, g_j + c\}$.

## Convergence Analysis

### ◇ Global Convergence

**○ Proposition**:

Let $\{x^k\}$ be generated by InexactCD-*Gauss-Southwell* $x^{k+1} = x^k + \alpha^k d^k$.

Assume that $P$ is lsc, $\{d^k\}$ is bounded, and $\alpha^k$ is chosen by the Armijo rule.

Then every cluster point of $\{x^k\}$ is a stationary point.

## $\diamond$ **Convergence Rate**

$\circ$ Error Bound

$\mathrm{dist}(x, S) \leq \kappa_1 \|R(x)\|_\infty$ whenever $\|R(x)\|_\infty \leq \epsilon_1$,

for some $\kappa_1 > 0$, $\epsilon_1 > 0$, where $S$ denotes the set of stationary points and $\mathrm{dist}(x, S) = \min_{s \in S} \|x - s\|_2$.

Corollary :

For the case of smooth problems with polyhedral constraints(ref $*$)

$f(x) - v \leq \kappa_2 \|R(x)\|_\infty^2$ whenever $\|R(x)\|_\infty \leq \epsilon_2$,

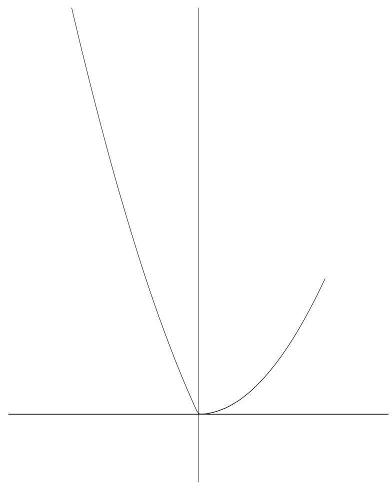
for some $\kappa_2 > 0$, $\epsilon_2 > 0$, where $v = \lim_{k \to \infty} f(x^k)$.

But this key bound used previously for convergence rate analysis fails for the general case.

$*$ Luo, Z.-Q. and Tseng, P., "Error bounds and convergence analysis of feasible descent methods: a general approach".

◦ Example:

$$\min_{x \in \Re} x^2 - x + |x|$$



◦ New key bound:

$(f + cP)(x) - v \leq \kappa_3 \|R(x)\|^2 + h(R(x))$

whenever $\quad \|R(x)\|_\infty \leq \epsilon_3$,

for some $\kappa_3 > 0$, $\epsilon_3 > 0$,

where $v = \lim_{k \to \infty} (f + cP)(x^k)$ and $h(x)$ is a nonnegative linear function.

∘ **Theorem 1**:

Assume $f(x) = g(Ex)$ where $E \in \Re^{m \times n}$, $g$ is strongly convex on $\Re^m$ with $\|\nabla g(x) - \nabla g(y)\| \leq L\|x - y\|$.

Let $\{x^k\}$ be generated by InexactCD-*Gauss-Seidel* (Armijo rule) with $\gamma\|z\|^2 \leq z^T H^k z$, $\limsup_{k,j} \alpha_j^k \leq \frac{\gamma}{L}$ and $\{d^k\}$ bdd.

Then $\{(f+cP)(x^k)\}$ converges at least Q-linearly and $\{x^k\}$ converges at least R-linearly.

Idea for Proof:

For sufficiently large $k$,

$$(f + cP)(x^{k+1}) - \upsilon \leq \kappa \|x^{k+1} - x^k\|^2 +$$
$$\sum_{j=1}^{n} \frac{1-\alpha_j^k}{\alpha_j^k}(\alpha_j^k < A_j^T \mu_j^k, d_j^k > +c(P_j(x_j^k) - P_j(x_j^{k+1})))$$

$$(f + cP)(x^{k+1}) - (f + cP)(x^k) \leq -\tfrac{L}{2}\|x^{k+1} - x^k\|^2 -$$
$$\sum_{j=1}^{n}(\alpha_j^k < A_j^T \mu_j^k, d_j^k > +c(P_j(x_j^k) - P_j(x_j^{k+1})))$$

where $(x_j, \xi_j) \in \text{epi}P_j \Leftrightarrow A_j x_j + a_j \xi_j \leq b_j$, $\kappa > 0$, and $\mu^k$ is some multiplier vector.

∘ **Theorem 2**:

Theorem 1 still holds if, in addition, $f$ is separable and *Gauss-Seidel* is replaced by *Gauss-Southwell*.

∘ **Conjecture**: Theorem 2 still holds without the separability of $f$

## Numerical Experience

∘ Coded in Matlab(running Matlab6.5)

∘ Time is on a Windows Laptop

∘ $H = diag(max(\nabla^2 f(x)_{jj}, 1))$

∘ Stop when $\|R(x)_j\|_\infty < 10^{-4}$

∘ test funcs(except 5) from the More-Garbow-Hillstrom collection

1. Brown almost-linear func(nonconvex)

$$f(x) = \sum_{i=1}^{n}(x_i + \sum_{j=1}^{n} x_j - (n+1))^2 + ((\prod_{j=1}^{n} x_j) - 1)^2$$

with $n = 100$ and $x^0 = (1, ..., 1)$.

2. Extended Rosenbrock func(nonconvex)

$$f(x) = \sum_{i=1}^{n/2}(100 * (x_{2i} - x_{2i-1}^2) + (1 - x_{2i-1})^2$$

with $n = 100$ and $x^0 = (\zeta_j)$ where $\zeta_{2j-1} = -1.2, \zeta_{2j} = 1$.

3. Extended Powell singular func(convex)

$$f(x) = \sum_{i=1}^{n/4}((x_{4i-3} + 10x_{4i-2})^2 + 5(x_{4i-1} - x_{4i} - 1)^2 + (x_{4i-2} - 2x_{4i-1})^4 + 10(x_{4i-3} - x_{4i})^4)$$

with $n = 1000$ and $x^0 = (\zeta_j)$ where $\zeta_{4j-3} = 3, \zeta_{4j-2} = -1, \zeta_{4j-1} = 0, \zeta_{4j} = 1$.

4. Variably dimensioned func(convex)

$$f(x) = \sum_{i=1}^{n}(x_i - 1)^2 + (\sum_{i=1}^{n} i(x_i - 1))^2 + (\sum_{i=1}^{n} i(x_i - 1))^4$$

with $n = 100$ and $x^0 = (1 - (j/n))$.

5. Quadratic func(satisfy assumption)

$$f(x) = (\sum_{i=1}^{n} x_i - n)^2$$

with $n = 1000$ and $x^0 = (1, ..., 1)$.

6. Linear func-full rank(satisfy assumption)

$$f(x) = (\sum_{i=1}^{n}(x_i - \frac{2}{m}(\sum_{j=1}^{n} x_j) - 1)^2 + (\frac{2}{m}(\sum_{j=1}^{n} x_j) + 1)^2$$

with $n = 1000, m = 1001$ and $x^0 = (1, ..., 1)$.

## Conclusion

1. Faster convergence if the smooth func $f$ is (partially) separable.
2. InexactCD-*Gauss-Southwell* is faster than InexactCD-*Gauss-Seidel*, especially, if $f$ is nonseparable.

## Future work

1. Prove the conjecture(Linear rate convergence for InexactCD-*Gauss-Southwell* still holds without the separability of $f$).
2. In our test, $n(J) = 1$. Can it be more efficient if we use block coordinate due to the separability structure of $f$?
3. More test on other functions and applications(e.g., regularized nonlinear least square)
4. Convergence acceleration for nonseparable function $f$?

# Reference

## References


[1] Bertsekas, D. P., Nonlinear Programming, 2nd edition, Athena Scientific, Belmont, 1999.

[2] Chen, S., Donoho, D., and Saunders, M., Atomic decomposition by basis pursuit, SIAM J. Sci. Comput. 20 (1999), 33-61.

[3] Luo, Z.-Q. and Tseng, P., Error bounds and convergence analysis of feasible descent methods: a general approach, Ann. Oper. Res. 46 (1993), 157–178.

[4] Mangasarian, O. L, Sparsity-preserving SOR algorithms for separable quadratic and linear programming, Comput. Oper. Res. 11 (1984), 105–112.

[5] Rockafellar, R. T., Convex Analysis, Princeton Univ. Press, Princeton, 1970.

[6] Sardy, S., Bruce, A., and Tseng, P., Block coordinate relaxation methods for nonparametric wavelet denoising, J. Comput. Graph. Stat. 9 (2000), 361-379.

[7] Tseng, P., Convergence of block coordinate descent method for nondifferentiable minimization, J. Optim. Theory Appl. 109 (2001), 473-492.

| **Problem name** | **c** | **♯ nonzero in sol** | **Inexact CD G-Seidel** | **Inexact CD G-Southwell** |
|---|---|---|---|---|
| | | | **iter/cpu** | **iter/cpu** |
| `Brown` | 10 | 100 | 150518/1958.4 | 1701/26.8 |
| `Lin` | 100 | 100 | 124918/1382.0 | 406/6.4 |
| `dim=100` | 1000 | 99 | /(>5000) | 3589/58.0 |
| `Ext` | 1 | 100 | 9499/289.1 | 9550/287.8 |
| `Ros` | 10 | 0 | 1399/40.9 | 1750/49.9 |
| `dim=100` | 100 | 0 | 399/9.1 | 300/8.8 |
| `Ext` | 1 | 1000 | 32997/4632.7 | 22500/3303.1 |
| `Pow` | 10 | 250 | 7999/988.8 | 4000/575.7 |
| `dim=1000` | 100 | 0 | 3997/266.1 | 1250/182.6 |
| `Var` | .1 | | /(>5000) | /(>5000) |
| `Dim` | 1 | | /(>5000) | /(>5000) |
| `dim=100` | 10 | | /(>5000) | /(>5000) |
| `Quad` | .1 | 610 | /(>3600) | 19962/2928.5 |
| `func` | 1 | 1 | /(>3600) | 1998/292.4 |
| `dim=1000` | 10 | 0 | 100001/945.1 | 1993/290.0 |
| `Lin` | .1 | 1000 | 1000/349.1 | 1000/321.0 |
| `f rank` | 1 | 1000 | 1000/350.3 | 1000/323.1 |
| `dim=1000` | 10 | 0 | 1000/340.0 | 1000/322.2 |

Table 1: test result