Block coordinate descent (BCD)

DSA5103 Lecture 6

Yangjing Zhang

16-Feb-2023

NUS

Today's content

1. Coordinate descent
2. Applications
   — linear regression, Lasso, box-contrained linear regression
3. Block coordinate descent

Coordinate descent

Recap

The optimization algorithms we have studied so far are gradient based methods:

- **Gradient descent methods** for $\min_{x \in \mathbb{R}^n} f(x)$

$$f : \mathbb{R}^n \to \mathbb{R} \text{ convex differentiable}$$

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

- **(Accelerated) proximal gradient methods** for $\min_{x \in \mathbb{R}^n} f(x) + g(x)$

$$f : \mathbb{R}^n \to \mathbb{R} \text{ convex differentiable, } \nabla f \; L\text{-Lipschitz continuous}$$

$$g : \mathbb{R}^n \to (-\infty, +\infty] \text{ closed proper convex } (g = \lambda \| \cdot \|_1, \, g = \delta_C)$$

$$x^{(k+1)} = P_{\alpha_k g}\left(x^{(k)} - \alpha_k \nabla f(x^{(k)})\right) \text{ (+ acceleration)}$$

Block coordinate descent methods

- Gradient descent, PG (proximal gradient), and APG (accelerated proximal gradient) involve gradient computation $\nabla f(\cdot)$
- Last time, SMO: block coordinate descent methods for solving the dual form of soft-margin SVM
- Today, we study in detail block coordinate descent methods. It does not need full gradient computation $\nabla f(\cdot)$, but sometimes need $\nabla_i f(\cdot) := \frac{\partial}{\partial x_i} f(\cdot)$

Coordinate-wise minimizer

Definition (Coordinate-wise minimizer)

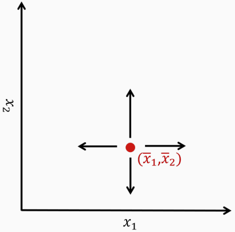
For any $f : \mathbb{R}^n \to (-\infty, +\infty]$, we say $\bar{x}$ is a coordinate-wise minimizer of f if $\bar{x} \in \operatorname{dom} f$ and

$$f(\bar{x} + de_i) \geq f(\bar{x}) \quad \forall\, i \in [n],\, d \in \mathbb{R} \tag{1}$$

where $e_i = (0, \ldots, 1, \ldots, 0)^T \in \mathbb{R}^n$ is the i-th standard basis vector.

- When $n = 2$, (1) is

$$\begin{aligned} f(\bar{x}_1 + d, \bar{x}_2) &\geq f(\bar{x}_1, \bar{x}_2) \\ f(\bar{x}_1, \bar{x}_2 + d) &\geq f(\bar{x}_1, \bar{x}_2) \quad \forall\, d \in \mathbb{R} \end{aligned}$$



- When $n = 3$, (1) is

$$\begin{aligned} f(\bar{x}_1 + d, \bar{x}_2, \bar{x}_3) &\geq f(\bar{x}_1, \bar{x}_2, \bar{x}_3) \\ f(\bar{x}_1, \bar{x}_2 + d, \bar{x}_3) &\geq f(\bar{x}_1, \bar{x}_2, \bar{x}_3) \\ f(\bar{x}_1, \bar{x}_2, \bar{x}_3 + d) &\geq f(\bar{x}_1, \bar{x}_2, \bar{x}_3) \quad \forall\, d \in \mathbb{R} \end{aligned}$$

Coordinate-wise minimizer

Definition (Coordinate-wise minimizer)

For any $f : \mathbb{R}^n \to (-\infty, +\infty]$, we say $\bar{x}$ is a coordinate-wise minimizer of f if $\bar{x} \in \operatorname{dom} f$ and

$$f(\bar{x} + de_i) \geq f(\bar{x}) \quad \forall\, i \in [n],\, d \in \mathbb{R} \tag{1}$$

where $e_i = (0, \ldots, 1, \ldots, 0)^T \in \mathbb{R}^n$ is the i-th standard basis vector.

- (1) is equivalent to

$$\bar{x}_i \in \arg\min_{x_i} f(\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_n) \quad \forall\, i \in [n].$$

Coordinate-wise minimizer

Definition (Coordinate-wise minimizer)

For any $f : \mathbb{R}^n \to (-\infty, +\infty]$, we say $\bar{x}$ is a coordinate-wise minimizer of f if $\bar{x} \in \operatorname{dom} f$ and

$$f(\bar{x} + d e_i) \geq f(\bar{x}) \quad \forall\, i \in [n],\, d \in \mathbb{R} \tag{1}$$

where $e_i = (0, \ldots, 1, \ldots, 0)^T \in \mathbb{R}^n$ is the i-th standard basis vector.

- (1) is equivalent to

$$\bar{x}_i \in \arg\min_{x_i} \; f(\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_n) \quad \forall\, i \in [n].$$

- Question: a coordinate-wise minimizer $\overset{?}{\Longrightarrow}$ a global minimizer

Example

Verify that $(\bar{x}_1; \bar{x}_2) = (-3; -3)$ is a coordinate-wise minimizer of

$$f(x_1, x_2) = x_1^2 + x_2^2 + 20|x_1 - x_2|$$

Solution.

Example

Verify that $(\bar{x}_1; \bar{x}_2) = (-3; -3)$ is a coordinate-wise minimizer of

$$f(x_1, x_2) = x_1^2 + x_2^2 + 20|x_1 - x_2|$$

Solution. We need to verify that

$$\bar{x}_1 \in \arg\min_{x_1} \ f(x_1, \bar{x}_2) = x_1^2 + 20|x_1 + 3| + 9 \qquad (*)$$
$$\bar{x}_2 \in \arg\min_{x_2} \ f(\bar{x}_1, x_2) = x_2^2 + 20|x_2 + 3| + 9$$

Namely, -3 is a global minimizer of $\min_x \ x^2 + 20|x + 3|$

Example

Verify that $(\bar{x}_1; \bar{x}_2) = (-3; -3)$ is a coordinate-wise minimizer of

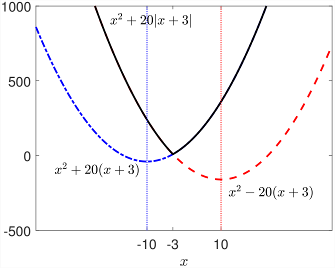
$$f(x_1, x_2) = x_1^2 + x_2^2 + 20|x_1 - x_2|$$

Solution. We need to verify that

$$\bar{x}_1 \in \arg\min_{x_1} \; f(x_1, \bar{x}_2) = x_1^2 + 20|x_1 + 3| + 9 \qquad (*)$$

$$\bar{x}_2 \in \arg\min_{x_2} \; f(\bar{x}_1, x_2) = x_2^2 + 20|x_2 + 3| + 9$$

Namely, -3 is a global minimizer of $\min_x \; x^2 + 20|x + 3|$



Method 1: verify graphically

$$\begin{cases} x^2 + 20(x+3), & \text{if } x \geq -3 \\ x^2 - 20(x+3), & \text{if } x < -3 \end{cases}$$

Example

Verify that $(\bar{x}_1; \bar{x}_2) = (-3; -3)$ is a coordinate-wise minimizer of

$$f(x_1, x_2) = x_1^2 + x_2^2 + 20|x_1 - x_2|$$

<u>Solution</u>. We need to verify that

$$\bar{x}_1 \in \arg\min_{x_1} \; f(x_1, \bar{x}_2) = x_1^2 + 20|x_1 + 3| + 9 \qquad (*)$$

$$\bar{x}_2 \in \arg\min_{x_2} \; f(\bar{x}_1, x_2) = x_2^2 + 20|x_2 + 3| + 9$$

Method 2: Recall[1] that $x^* \in \arg\min \; f(x) \iff 0 \in \partial f(x^*)$; If $g(x) = |x|, x \in \mathbb{R}$, then $\partial g(0) = [-1, 1]$.

(*) holds due to that $0 \in 2\bar{x}_1 + 20\,\partial g(\bar{x}_1 + 3) = -6 + [-20, 20]$.

[1] lecture 4, pages 24-25

Example

Verify that $(\bar{x}_1; \bar{x}_2) = (-3; -3)$ is a coordinate-wise minimizer of

$$f(x_1, x_2) = x_1^2 + x_2^2 + 20|x_1 - x_2|$$

Solution. We need to verify that

$$\bar{x}_1 \in \arg\min_{x_1} \; f(x_1, \bar{x}_2) = x_1^2 + 20|x_1 + 3| + 9 \qquad (*)$$

$$\bar{x}_2 \in \arg\min_{x_2} \; f(\bar{x}_1, x_2) = x_2^2 + 20|x_2 + 3| + 9$$

Method 2: Recall[1] that $x^* \in \arg\min \, f(x) \iff 0 \in \partial f(x^*)$; If $g(x) = |x|, x \in \mathbb{R}$, then $\partial g(0) = [-1, 1]$.

(*) holds due to that $0 \in 2\bar{x}_1 + 20 \, \partial g(\bar{x}_1 + 3) = -6 + [-20, 20]$.

Remark: $(-3; -3)$ is not a global minimizer (a global minimizer is $(0; 0)$).

[1] lecture 4, pages 24-25

Example

Verify that $(\bar{x}_1; \bar{x}_2) = (0; 0)$ is a coordinate-wise minimizer of

$$f(x_1, x_2) = (x_1 - x_2)^2 + |x_1| + |x_2|$$

Solution. We need to verify that

$$\bar{x}_1 \in \arg\min_{x_1} \; f(x_1, \bar{x}_2) = x_1^2 + |x_1|$$
$$\bar{x}_2 \in \arg\min_{x_2} \; f(\bar{x}_1, x_2) = x_2^2 + |x_2|$$

0 is a global minimizer of $\min_{x} \; x^2 + |x|$. ($x^2 + |x| \geq 0$ for any x)

Remark: $(0; 0)$ is indeed a global minimizer.

Coordinate-wise minimizer: differentiable

a coordinate-wise minimizer $\overset{\text{differentiable}}{\Longrightarrow}$ a global minimizer

Claim: A coordinate-wise minimizer $\bar{x}$ of a convex function f is a global minimizer of f whenever f is differentiable at $\bar{x}$.

Proof: Since f is differentiable at $\bar{x}$,

$$\bar{x}_i \in \arg\min_{x_i} f(\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_n)$$

implies that

$$\nabla_i f(\bar{x}) = \frac{\partial}{\partial x_i} f(\bar{x}) = 0.$$

Thus $\nabla f(\bar{x}) = (\nabla_1 f(\bar{x}), \ldots, \nabla_n f(\bar{x})) = 0$, $\bar{x}$ is a global minimizer of f.

Question: same question for non-differentiable function?

Coordinate-wise minimizer: non-differentiable

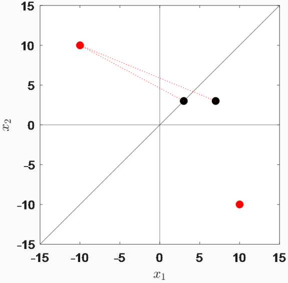
a coordinate-wise minimizer $\overset{\text{non-differentiable}}{\not\Longrightarrow}$ a global minimizer

Claim: A coordinate-wise minimizer $\bar{x}$ of a convex function f is not necessarily a global minimizer of f when f is not differentiable at $\bar{x}$.

Example: $f : \mathbb{R}^2 \to \mathbb{R}$ is convex but not differentiable when $x_1 = x_2$

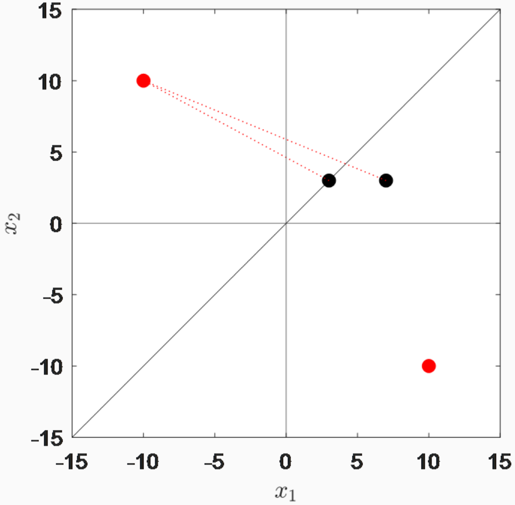
$$f(x_1, x_2)$$
$$= \begin{cases} (x_1+10)^2 + (x_2-10)^2, & \text{if } x_1 \geq x_2 \\ (x_1-10)^2 + (x_2+10)^2, & \text{if } x_1 < x_2 \end{cases}$$
$$= x_1^2 + x_2^2 + 20|x_1 - x_2| + 200$$

The global minimizer of f is $(0, 0)$.



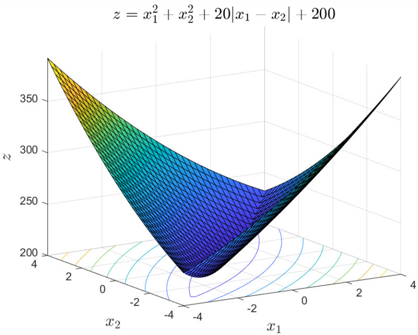
Coordinate-wise minimizer: non-differentiable

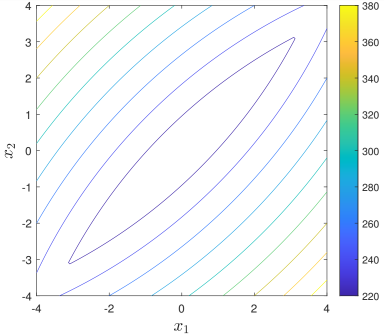
- $(3; 3)$ is a coordinate-wise minimizer of f (In fact, any point (c, c), $|c| \leq 10$ is a coordinate-wise minimizer).



Coordinate-wise minimizer: non-differentiable

Plot and contour plot of $f(x_1, x_2) = x_1^2 + x_2^2 + 20|x_1 - x_2| + 200$





Coordinate-wise minimizer: separable non-differentiable

a coordinate-wise minimizer $\underset{\text{but separable}}{\overset{\text{non-differentiable}}{\Longrightarrow}}$ a global minimizer

Claim: For the following problem where the non-differentiable part can be decomposed into a sum of functions over each coordinate

$$\min_{x \in \mathbb{R}^n} \quad F(x) := f(x) + \underbrace{\sum_{i=1}^{n} r_i(x_i)}_{\text{separable}}$$

$f : \mathbb{R}^n \to \mathbb{R}$ convex differentiable

each $r_i : \mathbb{R}^n \to (-\infty, +\infty]$ closed proper convex

A coordinate-wise minimizer of F is a global minimizer of F.

Proof: Denote $r(x) := \sum_{i=1}^{n} r_i(x_i)$

$$\bar{x}_i \in \arg\min_{x_i} \; f(\bar{x}_1, \ldots, \bar{x}_{i-1}, x_i, \bar{x}_{i+1}, \ldots, \bar{x}_n) + r_i(x_i) \quad \forall\, i$$

$$\iff 0 \in \nabla_i(\bar{x}) + \partial r_i(\bar{x}_i) \quad \forall\, i \iff 0 \in \nabla f(\bar{x}) + \partial r(\bar{x})$$

Coordinate descent method

Target problem

$$\min_{x \in \mathbb{R}^n} \quad F(x) := f(x) + \sum_{i=1}^{n} r_i(x_i)$$

$f : \mathbb{R}^n \to \mathbb{R}$ convex differentiable

each $r_i : \mathbb{R}^n \to (-\infty, +\infty]$ closed proper convex

Why this problem?

- Roughly, a coordinate descent method will search for a coordinate-wise minimizer
- In general, a coordinate-wise minimizer is not necessarily a global minimizer
- For the target problem, a coordinate-wise minimizer is indeed a global minimizer

Coordinate descent method

Algorithm (Coordinate descent method)

Choose $x^{(0)} \in \mathrm{dom} F$. Set $k \leftarrow 0$

repeat until convergence

$$
\begin{aligned}
x_1^{(k+1)} &\leftarrow \arg\min_{x_1} f(x_1, \quad x_2^{(k)}, \quad x_3^{(k)}, \quad \ldots, x_n^{(k)}) + r_1(x_1) \\
x_2^{(k+1)} &\leftarrow \arg\min_{x_2} f(x_1^{(k+1)}, \quad x_2, \quad x_3^{(k)}, \quad \ldots, x_n^{(k)}) + r_2(x_2) \\
x_3^{(k+1)} &\leftarrow \arg\min_{x_3} f(x_1^{(k+1)}, \quad x_2^{(k+1)}, \quad x_3, \quad \ldots, x_n^{(k)}) + r_3(x_3) \\
&\quad \vdots \\
x_n^{(k+1)} &\leftarrow \arg\min_{x_n} f(x_1^{(k+1)}, \quad x_2^{(k+1)}, \quad x_3^{(k+1)}, \quad \ldots, x_n) + r_n(x_n)
\end{aligned}
$$

$k \leftarrow k+1$

end(repeat)

Coordinate descent method

We make the following remarks

- $x^{(k)}$ has a subsequence converging to a global minimizer x^*; function value $F(x^{(k)})$ converges to $F(x^*)$. See [1] for details of convergence properties
- the coordinates can be cycled through in any arbitrary order; the most-often used order is in cyclic order: $x_1 \to x_2 \to \cdots \to x_n$
- after we solve for $x_i^{(k+1)}$, we use its new value from then on! Therefore, the minimizations can not be performed in parallel
- Later, we extend coordinate descent to block coordinate descent: instead of minimizing over individual coordinates, any block of coordinates can be minimized over
- There is no global convergence result for non-convex F

Applications

Toy example

Apply coordinate descent method for

$$\min_{x=(x_1;x_2)\in\mathbb{R}^2} \quad f(x_1,x_2) = (x_1-x_2)^2 + |x_1| + |x_2|$$

with initial point $x^{(0)} = (6;6)$.

<u>Solution</u>. For $k = 0, 1, 2, \ldots,$ iterations are

$$x_1^{(k+1)} \in \arg\min_{x_1} \; f(x_1, x_2^{(k)}) = \arg\min_{x_1} \; \left(x_1 - x_2^{(k)}\right)^2 + |x_1|$$

$$x_2^{(k+1)} \in \arg\min_{x_2} \; f(x_1^{(k+1)}, x_2) = \arg\min_{x_2} \; \left(x_2 - x_1^{(k+1)}\right)^2 + |x_2|$$

[2] lecture 4, pages 32,34,35

Toy example

Apply coordinate descent method for

$$\min_{x=(x_1;x_2)\in\mathbb{R}^2} \quad f(x_1,x_2) = (x_1 - x_2)^2 + |x_1| + |x_2|$$

with initial point $x^{(0)} = (6;6)$.

<u>Solution</u>. For $k = 0, 1, 2, \ldots,$ iterations are

$$x_1^{(k+1)} \in \arg\min_{x_1} \ f(x_1, x_2^{(k)}) = \arg\min_{x_1} \ \left(x_1 - x_2^{(k)}\right)^2 + |x_1|$$

$$x_2^{(k+1)} \in \arg\min_{x_2} \ f(x_1^{(k+1)}, x_2) = \arg\min_{x_2} \ \left(x_2 - x_1^{(k+1)}\right)^2 + |x_2|$$

By the definition of proximal mapping[2]

$$x_1^{(k+1)} = \arg\min_{x_1} \ \frac{1}{2}|x_1| + \frac{1}{2}\left(x_1 - x_2^{(k)}\right)^2 = P_{0.5|\cdot|}\left(x_2^{(k)}\right) = S_{0.5}\left(x_2^{(k)}\right)$$

$$x_2^{(k+1)} = S_{0.5}\left(x_1^{(k+1)}\right)$$

[2]lecture 4, pages 32,34,35

Toy example

Apply coordinate descent method for

$$\min_{x=(x_1;x_2)\in\mathbb{R}^2} \quad f(x_1,x_2) = (x_1 - x_2)^2 + |x_1| + |x_2|$$

with initial point $x^{(0)} = (6;6)$. Compute $x^{(1)}$ and $x^{(2)}$.

<u>Solution</u>.

$$x_1^{(k+1)} = S_{0.5}\left(x_2^{(k)}\right), \qquad x_2^{(k+1)} = S_{0.5}\left(x_1^{(k+1)}\right)$$

$\underline{k=0}$. $x^{(1)} = (5.5;5)$

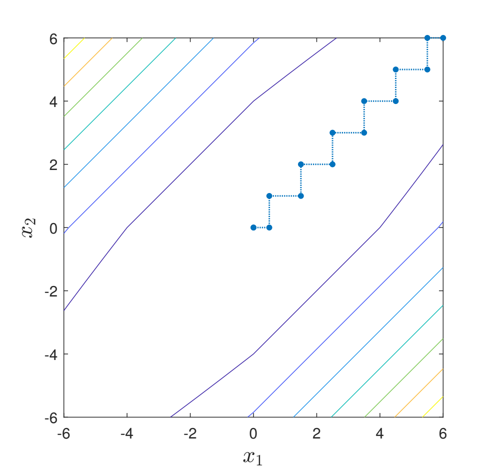
$$x_1^{(1)} = S_{0.5}\left(x_2^{(0)}\right) = S_{0.5}(6) = 5.5,\ x_2^{(1)} = S_{0.5}\left(x_1^{(1)}\right) = S_{0.5}(5.5) = 5$$

$\underline{k=1}$. $x^{(1)} = (4.5;4)$

$$x_1^{(2)} = S_{0.5}\left(x_2^{(1)}\right) = S_{0.5}(5) = 4.5,\ x_2^{(2)} = S_{0.5}\left(x_1^{(2)}\right) = S_{0.5}(4.5) = 4$$

Toy example

Contour plot of $f(x_1, x_2) = (x_1 - x_2)^2 + |x_1| + |x_2|$ and iterates $x^{(k)}$.



Application 1: linear regression

$$\min_{\beta \in \mathbb{R}^p} \quad L(\beta) = \frac{1}{2}\|X\beta - Y\|^2$$

where $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$ (assume intercept term $\beta_0 = 0$)

Solution. Consider minimizing over β_i with all β_j, $j \neq i$ fixed.

Application 1: linear regression

$$\min_{\beta \in \mathbb{R}^p} \quad L(\beta) = \frac{1}{2}\|X\beta - Y\|^2$$

where $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^n$ (assume intercept term $\beta_0 = 0$)

Solution. Consider minimizing over β_i with all β_j, $j \neq i$ fixed. Note that

$$X\beta = X_{\cdot 1}\beta_1 + X_{\cdot 2}\beta_2 + \cdots + X_{\cdot p}\beta_p = X_{\cdot i}\beta_i + X_{-i}\beta_{-i}$$

$$\begin{aligned}
\text{where} \quad X &= [X_{\cdot 1} \cdots X_{\cdot (i-1)} \; X_{\cdot i} \; X_{\cdot (i+1)} \cdots X_{\cdot p}] \\
X_{-i} &= [X_{\cdot 1} \cdots X_{\cdot (i-1)} \qquad X_{\cdot (i+1)} \cdots X_{\cdot p}] \text{ delete } i\text{-th column} \\
\beta_{-i} &= [\beta_1 \cdots \beta_{i-1} \; \beta_{i+1} \cdots \beta_p]^T \text{ delete } i\text{-th entry}
\end{aligned}$$

Therefore, $L(\beta) = \frac{1}{2}\|X_{\cdot i}\beta_i + X_{-i}\beta_{-i} - Y\|^2$ and we set

$$0 = \nabla_i L(\beta) = X_{\cdot i}^T(X_{\cdot i}\beta_i + X_{-i}\beta_{-i} - Y) \Rightarrow \beta_i = \frac{X_{\cdot i}^T(Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}$$

Application 1: linear regression

Algorithm (Coordinate descent method for linear regression)

Initialize β.

repeat until convergence

for $i = 1, \ldots, p$

$$\beta_i \leftarrow \frac{X_{\cdot i}^T (Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}$$

end(for)

end(repeat)

Note: update of β_i can also be written as $\beta_i \leftarrow \beta_i - \dfrac{X_{\cdot i}^T (X\beta - Y)}{\|X_{\cdot i}\|^2}$

Synthetic data for linear regression

- Generate an $n \times p$ feature matrix X, each entry follows a standard normal distribution $X_{ij} \sim N(0,1)$
- Generate a sparse $\beta_{\text{true}} \in \mathbb{R}^p$, e.g.,

```
beta_true = wthresh(randn(p,1),'s',0.5);
```

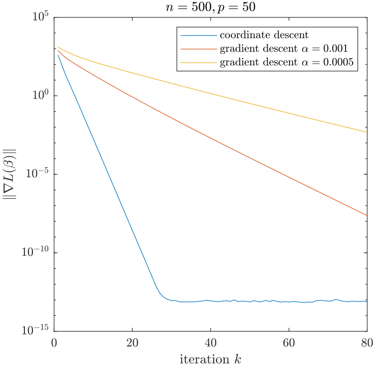
- The response vector

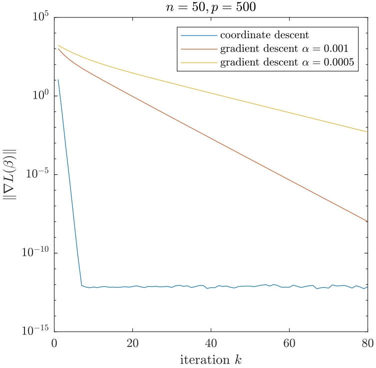
$$Y = X\beta_{\text{true}} + 0.1\epsilon$$

where $\epsilon_i \sim N(0,1)$, $i \in [n]$ is the Gaussian noise.

Coordinate descent: $\beta_i \leftarrow \beta_i - \dfrac{X_{\cdot i}^T(X\beta - Y)}{\|X_{\cdot i}\|^2}$, $i = 1, \ldots, p$

Gradient descent: $\beta_i \leftarrow \beta_i - \alpha X_{\cdot i}^T(X\beta - Y)$, $i = 1, \ldots, p$

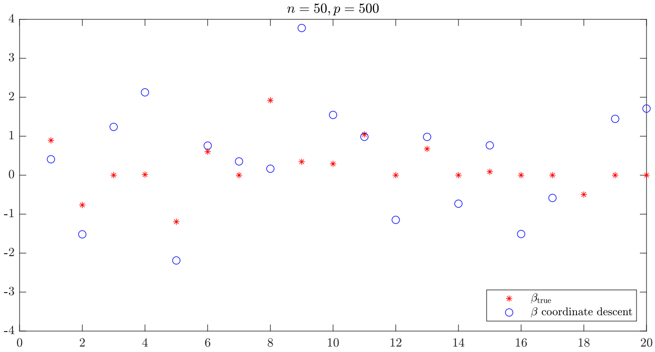




- Coordinate descent is often faster than gradient descent with a constant step size for linear regression.
- Coordinate descent is "parameter-free", but cannot be performed in parallel. Gradient descent needs to choose step size α, it can be performed in parallel.

Plot the first 20 entries of β.

Many entries of β_{true} is zero. However, the estimated β from linear regression may not be sparse.



```
n = 50; p = 500; % n=sample sizes p=#features
X = randn(n,p); % random feature matrix
beta_true = wthresh(randn(p,1),'s',0.5); % sparse true beta
Y = X*beta_true + 0.1*randn(n,1); % response vector
%% coordinate descent
beta = zeros(p,1); % initialization
norm_grad1 = zeros(80,1); % record results
for k = 1:80
    for i = 1:p
        Xi = X(:,i);
        beta(i) = beta(i) - Xi'*(X*beta - Y)/(Xi'*Xi);
    end
    norm_grad1(k) = norm(X'*(X*beta - Y));
end
%% gradient descent
beta = zeros(p,1); % initialization
norm_grad2 = zeros(80,1); % record results
alpha = 0.001; % constant step size
for k = 1:80
    beta = beta - alpha*X'*(X*beta - Y);
    norm_grad2(k) = norm(X'*(X*beta - Y));
end
semilogy(norm_grad1); hold on; semilogy(norm_grad2); % plot
```

Application 2: Lasso

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2}\|X\beta - Y\|^2 + \lambda\|\beta\|_1$$

where $X \in \mathbb{R}^{n\times p}$, $Y \in \mathbb{R}^n$, $\lambda > 0$. The non-differentiable term is separable: $\lambda\|\beta\|_1 = \sum_{i=1}^{p} \lambda|\beta_i|$.

Solution. Consider minimizing over β_i with all β_j, $j \neq i$ fixed. We first write $\|X\beta - Y\|^2$ in a quadratic form of β_i.

$$\begin{aligned}
&\frac{1}{2}\|X\beta - Y\|^2 = \frac{1}{2}\|X_{\cdot i}\beta_i + \overbrace{X_{-i}\beta_{-i} - Y}^{:=\Delta}\|^2 \\
=&\frac{1}{2}\|X_{\cdot i}\beta_i\|^2 + \langle X_{\cdot i}\beta_i, \Delta\rangle + \frac{\|\Delta\|^2}{2} \\
=&\frac{1}{2}\|X_{\cdot i}\|^2\beta_i^2 + \left(X_{\cdot i}^T\Delta\right)\beta_i + \frac{\|\Delta\|^2}{2}
\end{aligned}$$

Application 2: Lasso

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2}\|X\beta - Y\|^2 + \lambda\|\beta\|_1$$

where $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, $\lambda > 0$. The non-differentiable term is separable: $\lambda\|\beta\|_1 = \sum_{i=1}^{p} \lambda|\beta_i|$.

Solution. We solve $\quad \min_{\beta_i} \frac{1}{2}\|X_{\cdot i}\|^2 \beta_i^2 + \left(X_{\cdot i}^T \Delta\right) \beta_i + \lambda|\beta_i|$

Application 2: Lasso

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2}\|X\beta - Y\|^2 + \lambda\|\beta\|_1$$

where $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, $\lambda > 0$. The non-differentiable term is separable: $\lambda\|\beta\|_1 = \sum_{i=1}^{p} \lambda|\beta_i|$.

Solution. We solve $\quad \min_{\beta_i} \frac{1}{2}\|X_{\cdot i}\|^2 \beta_i^2 + \left(X_{\cdot i}^T \Delta\right) \beta_i + \lambda|\beta_i|$

$$\overset{\text{complete}}{\underset{\text{squares}}{\Longleftrightarrow}} \quad \min_{\beta_i} \frac{1}{2}\|X_{\cdot i}\|^2 \left(\beta_i + \frac{X_{\cdot i}^T \Delta}{\|X_{\cdot i}\|^2}\right)^2 + \lambda|\beta_i|$$

$$\Longleftrightarrow \quad \min_{\beta_i} \frac{1}{2}\left(\beta_i + \frac{X_{\cdot i}^T \Delta}{\|X_{\cdot i}\|^2}\right)^2 + \frac{\lambda}{\|X_{\cdot i}\|^2}|\beta_i|$$

Therefore,

$$\beta_i = P_{\frac{\lambda}{\|X_{\cdot i}\|^2}|\cdot|}\left(-\frac{X_{\cdot i}^T \Delta}{\|X_{\cdot i}\|^2}\right) = S_{\lambda/\|X_{\cdot i}\|^2}\left(\frac{X_{\cdot i}^T (Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}\right)$$

Application 2: Lasso

Apply coordinate descent method for

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2}\|X\beta - Y\|^2 + \lambda\|\beta\|_1$$

where $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, $\lambda > 0$.

<u>Alternative solution</u>. "Differentiate" the objective function w.r.t. β_i

Application 2: Lasso

Apply coordinate descent method for

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2}\|X\beta - Y\|^2 + \lambda \|\beta\|_1$$

where $X \in \mathbb{R}^{n \times p}$, $Y \in \mathbb{R}^n$, $\lambda > 0$.

<u>Alternative solution</u>. "Differentiate" the objective function w.r.t. β_i

$$\begin{aligned} 0 &\in X_{\cdot i}^T (X_{\cdot i}\beta_i + X_{-i}\beta_{-i} - Y) + \partial(\lambda|\cdot|)(\beta_i) \\ &= \|X_{\cdot i}\|^2 \beta_i - X_{\cdot i}^T (Y - X_{-i}\beta_{-i}) + \partial(\lambda|\cdot|)(\beta_i) \end{aligned}$$

$$\iff 0 \in \beta_i - \frac{X_{\cdot i}^T (Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2} + \partial\left(\frac{\lambda}{\|X_{\cdot i}\|^2}|\cdot|\right)(\beta_i)$$

$$\stackrel{\bullet}{\iff} \beta_i = P_{\frac{\lambda}{\|X_{\cdot i}\|^2}|\cdot|}\left(-\frac{X_{\cdot i}^T (Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}\right) = S_{\lambda/\|X_{\cdot i}\|^2}\left(\frac{X_{\cdot i}^T (Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}\right)$$

- Due to that

$$y = P_f(x) = \arg\min_y \left\{ f(y) + \frac{1}{2}\|y - x\|^2 \right\} \iff 0 \in y - x + \partial f(y)$$

Application 2: Lasso

Algorithm (Coordinate descent method for Lasso)

Initialize β.

repeat until convergence

for $i = 1, \ldots, p$

$$\beta_i \leftarrow S_{\lambda/\|X_{\cdot i}\|^2}\left(\frac{X_{\cdot i}^T(Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}\right)$$

end(for)

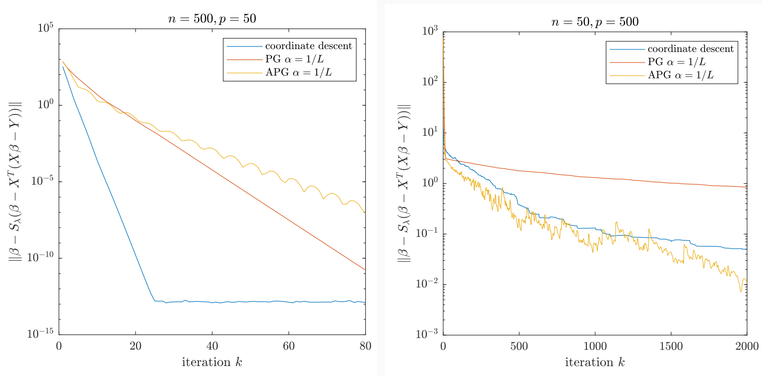
end(repeat)

Note: compare it with the linear regression case in page 21. We here have the additional soft-thresholding operator.

- Use the synthetic data in page 22.
- Set $\lambda = 0.5$, $L = \lambda_{\max}(X^T X)$, step size $\alpha = 1/L$ for PG and APG
- Plot the residual[3]

$$\left\| \beta^{(k)} - S_\lambda \left(\beta^{(k)} - X^T (X\beta^{(k)} - Y) \right) \right\| < \varepsilon.$$

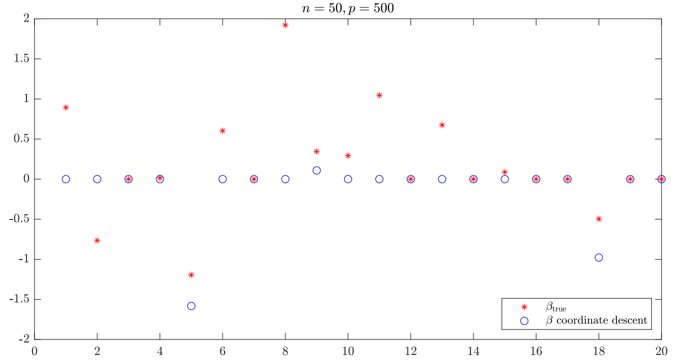
against iteration k



[3] lecture 4, page 50

Plot the first 20 entries of β.

The estimated β from Lasso is indeed sparse.



```
n = 50; p = 500; % n=sample sizes p=#features
X = randn(n,p); % random feature matrix
beta_true = wthresh(randn(p,1),'s',0.5); % sparse true beta
Y = X*beta_true + 0.1*randn(n,1); % response vector
lambda = 0.5; % L1 penalty parameter
%% coordinate descent
maxiter = 2000;
beta = zeros(p,1); % initialization
norm_grad1 = zeros(maxiter,1); % record results
for k = 1:maxiter
    for i = 1:p
        Xi = X(:,i);
        ni = Xi'*Xi;
        beta(i) = wthresh(beta(i) - Xi'*(X*beta - Y)/ni,'s',
            lambda/ni);
    end
    norm_grad1(k) = norm(beta - wthresh(beta - X'*(X*beta -
        Y),'s',lambda));
end
```

```
%% proximal gradient alpha = 1/L
beta = zeros(p,1); % initialization
norm_grad2 = zeros(maxiter,1); % record results
alpha = 1/eigs(X'*X,1); % step size = 1/L
for k = 1:maxiter
    beta = wthresh(beta - alpha*X'*(X*beta - Y),'s',alpha*
        lambda);
    norm_grad2(k) = norm(beta - wthresh(beta - X'*(X*beta -
        Y),'s',lambda));
end

t = zeros(maxiter + 1,1);
t(1) = 1; t(2) = 1;
for k = 3:maxiter+1
    t(k) = (1 + sqrt(1 + 4*t(k-1)^2))/2;
end
```

```matlab
%% Accelerated proximal gradient alpha = 1/L
beta = zeros(p,1); % initialization
norm_grad3 = zeros(maxiter,1); % record results
beta_old = beta;
for k = 1:maxiter
    beta_bar = beta + (t(k) - 1)/(t(k+1))*(beta - beta_old);
    beta_new = wthresh(beta_bar - alpha*X'*(X*beta_bar - Y),
        's',alpha*lambda);
    norm_grad3(k) = norm(beta_new - wthresh(beta_new - X'*(X
        *beta_new - Y),'s',lambda));
    beta_old = beta;
    beta = beta_new;
end
%% plot
semilogy(norm_grad1);
hold on;
semilogy(norm_grad2);
semilogy(norm_grad3);
legend({'coordinate descent','PG','APG'});
```

Application 3: box-constrained regression

Apply coordinate descent method for linear regression under the box constraint

$$\begin{aligned} \min_{\beta \in \mathbb{R}^p} \quad & \frac{1}{2}\|X\beta - Y\|^2 \\ \text{s.t.} \quad & l \leq \beta \leq u \end{aligned}$$

Solution. The inequality in the constraint $l \leq \beta \leq u$ is component-wise:

$$l_i \leq \beta_i \leq u_i,\ i \in [p].$$

With an indicator function of the "box" $C = \{\beta \mid l \leq \beta \leq u\}$, the problem can be written as

$$\min_{\beta \in \mathbb{R}^p} \quad \frac{1}{2}\|X\beta - Y\|^2 + \delta_C(\beta)$$

Separable:

$$\delta_C(\beta) = \sum_{i=1}^{p} \delta_{C_i}(\beta_i), \quad C_i = \{\beta_i \mid l_i \leq \beta_i \leq u_i\}$$

Application 3: box-constrained regression

Solution. (Repeat derivations in page 28)

$$0 \in \|X_{\cdot i}\|^2\beta_i - X_{\cdot i}^T(Y - X_{-i}\beta_{-i}) + \partial(\delta_{C_i})(\beta_i)$$

$$\Longleftrightarrow 0 \in \beta_i - \frac{X_{\cdot i}^T(Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2} + \partial(\frac{1}{\|X_{\cdot i}\|^2}\delta_{C_i})(\beta_i)$$

$$\Longleftrightarrow \beta_i = P_{\delta_{C_i}}\left(-\frac{X_{\cdot i}^T(Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}\right) = \Pi_{C_i}\left(\frac{X_{\cdot i}^T(Y - X_{-i}\beta_{-i})}{\|X_{\cdot i}\|^2}\right)$$

Besides,

$$\Pi_{C_i}(\beta_i) = \begin{cases} u_i, & \text{if } \beta_i > u_i \\ \beta_i, & \text{if } l_i \le \beta_i \le u_i \\ l_i, & \text{if } \beta_i < l_i \end{cases}$$

Block coordinate descent

Block coordinate descent method

- Up to now, we update one coordinate and the solve a univariate problem. It can extended to block case, where we update a block of coordinates.
- Target problem (x can be partitioned into m blocks)

$$\min_{x \in \mathcal{X}} \quad F(x) := f(x) + \sum_{i=1}^{m} r_i(x_i)$$

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_m$$

$f : \mathcal{X} \to \mathbb{R}$ convex differentiable

each $r_i : \mathcal{X}_i \to (-\infty, +\infty]$ closed proper convex

- When the minimization over each block is easy, we apply block coordinate descent method

Block coordinate

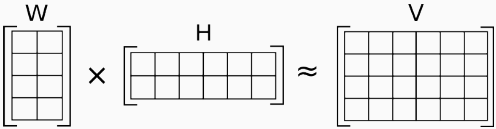
- For problem with vector variable x, a block of coordinate can be
  - ▷ a single element x_i
  - ▷ a subvector containing a block of elements $x_{i_i}, x_{i_2}, \ldots, x_{i_k}$
- For problem with matrix variable $X \in \mathbb{R}^{m \times n}$, a block of coordinate can be
  - ▷ a single element X_{ij}
  - ▷ a row $X_{i\cdot}$ or a column $X_{\cdot j}$
  - ▷ a submatrix

Application 4: NMF

Consider nonnegative matrix factorization (NMF) problem (we postpone the introduction of NMF model to next lecture)

$$\begin{aligned} \min_{W,H} \quad & \frac{1}{2}\|V - WH\|^2 \\ \text{s.t.} \quad & W \geq 0,\ H \geq 0 \end{aligned}$$

where $V \in \mathbb{R}^{m\times n}$, $W \in \mathbb{R}^{m\times r}$, $H \in \mathbb{R}^{r\times n}$, $W \geq 0$ means $W_{ij} \geq 0$ (same for $H \geq 0$)



- The objective function is non-convex w.r.t. (W, H), but convex in W (with H fixed) and convex in H (with W fixed)
- Treat a column of W or a row of H as a block ($2r$ blocks in total)
- The constraints $W \geq 0,\ H \geq 0$ are separable

Application 4: NMF

Notation:

- $W_{\cdot i}$ is i-th column of W, $W_{\cdot(-i)}$ is constructed from W after deleting the i-th column of W
- $H_{i\cdot}$ is i-th row of H, $H_{(-i)\cdot}$ is constructed from H after deleting the i-th row of H
- $\langle x, y\rangle = \mathrm{Tr}(x^T y)$, $\|x\| = \sqrt{\langle x, x\rangle}$ for vectors x, y or matrices x, y. In particular,

$$\begin{cases} \|X\| = \|X\|_F = \sqrt{\mathrm{Tr}(X^T X)}, & \text{for a matrix } X \\ \|x\| = \|x\|_2 = \sqrt{x^T x}, & \text{for a vector } x \end{cases}$$

Therefore, we have

$$WH = \sum_{i=1}^{r} W_{\cdot i} H_{i\cdot} = W_{\cdot i} H_{i\cdot} + W_{\cdot(-i)} H_{(-i)\cdot}$$

Application 4: NMF

We first try to write the objective function as a function of $W_{\cdot i}$.

Application 4: NMF

We first try to write the objective function as a function of $W_{\cdot i}$.

$$\frac{1}{2}\|V - WH\|^2 = \frac{1}{2}\|\overbrace{V - W_{\cdot(-i)}H_{(-i)\cdot}}^{:=\Delta} - W_{\cdot i}H_{i\cdot}\|^2$$

$$= \frac{1}{2}\langle \Delta, \Delta\rangle - \langle \Delta, W_{\cdot i}H_{i\cdot}\rangle + \frac{1}{2}\langle W_{\cdot i}H_{i\cdot}, W_{\cdot i}H_{i\cdot}\rangle$$

$$= \frac{1}{2}\|\Delta\|^2 - \langle W_{\cdot i}, \Delta H_{i\cdot}^T\rangle + \frac{1}{2}\|W_{\cdot i}\|^2\|H_{i\cdot}\|^2$$

Application 4: NMF

We first try to write the objective function as a function of $W_{\cdot i}$.

$$\frac{1}{2}\|V - WH\|^2 = \frac{1}{2}\|\overbrace{V - W_{\cdot(-i)}H_{(-i)\cdot}}^{:=\Delta} - W_{\cdot i}H_{i\cdot}\|^2$$
$$= \frac{1}{2}\langle\Delta, \Delta\rangle - \langle\Delta, W_{\cdot i}H_{i\cdot}\rangle + \frac{1}{2}\langle W_{\cdot i}H_{i\cdot}, W_{\cdot i}H_{i\cdot}\rangle$$
$$= \frac{1}{2}\|\Delta\|^2 - \langle W_{\cdot i}, \Delta H_{i\cdot}^T\rangle + \frac{1}{2}\|W_{\cdot i}\|^2\|H_{i\cdot}\|^2$$

Consider minimizing a block $W_{\cdot i}$ with all other blocks fixed.

$$W_{\cdot i} = \arg\min_x \frac{1}{2}\|x\|^2 - \langle x, \frac{\Delta H_{i\cdot}^T}{\|H_{i\cdot}\|^2}\rangle + \delta_{\mathbb{R}_+^m}(x)$$
$$\overset{\text{check!}}{\Longleftrightarrow} W_{\cdot i} = P_{\delta_{\mathbb{R}_+^m}}\left(\frac{\Delta H_{i\cdot}^T}{\|H_{i\cdot}\|^2}\right) = \Pi_{\mathbb{R}_+^m}\left(\frac{\left(V - W_{\cdot(-i)}H_{(-i)\cdot}\right)H_{i\cdot}^T}{\|H_{i\cdot}\|^2}\right)$$

Application 4: NMF

We first try to write the objective function as a function of $W_{\cdot i}$.

$$
\begin{aligned}
&\frac{1}{2}\|V - WH\|^2 = \frac{1}{2}\|\overbrace{V - W_{\cdot(-i)}H_{(-i)\cdot}}^{:=\Delta} - W_{\cdot i}H_{i\cdot}\|^2 \\
&= \frac{1}{2}\langle \Delta, \Delta\rangle - \langle \Delta, W_{\cdot i}H_{i\cdot}\rangle + \frac{1}{2}\langle W_{\cdot i}H_{i\cdot}, W_{\cdot i}H_{i\cdot}\rangle \\
&= \frac{1}{2}\|\Delta\|^2 - \langle W_{\cdot i}, \Delta H_{i\cdot}^T\rangle + \frac{1}{2}\|W_{\cdot i}\|^2\|H_{i\cdot}\|^2
\end{aligned}
$$

Consider minimizing a block $W_{\cdot i}$ with all other blocks fixed.

$$
W_{\cdot i} = \arg\min_x \frac{1}{2}\|x\|^2 - \langle x, \frac{\Delta H_{i\cdot}^T}{\|H_{i\cdot}\|^2}\rangle + \delta_{\mathbb{R}_+^m}(x)
$$

$$
\overset{\text{check!}}{\Longleftrightarrow} W_{\cdot i} = P_{\delta_{\mathbb{R}_+^m}}\left(\frac{\Delta H_{i\cdot}^T}{\|H_{i\cdot}\|^2}\right) = \Pi_{\mathbb{R}_+^m}\left(\frac{\left(V - W_{\cdot(-i)}H_{(-i)\cdot}\right)H_{i\cdot}^T}{\|H_{i\cdot}\|^2}\right)
$$

Similarly, block $H_{i\cdot}$ is updated by

$$
H_{i\cdot} = \Pi_{\mathbb{R}_+^n}\left(\frac{W_{\cdot i}^T\left(V - W_{\cdot(-i)}H_{(-i)\cdot}\right)}{\|W_{\cdot i}\|^2}\right)
$$

References i

P. Tseng.
Convergence of a block coordinate descent method for nondifferentiable minimization.
Journal of Optimization Theory and Applications, 109(3):475, 2001.