R-Linear Convergence of Limited Memory Steepest Descent

Frank E. Curtis, Lehigh University

joint work with

Wei Guo, Lehigh University

15th EUROPT Workshop on Advances in Continuous Optimization
Montréal, Québec, Canada

12 July 2017

Outline

Outline

Introduction

Limited Memory Steepest Descent (LMSD)

R-Linear Convergence of LMSD

Numerical Demonstrations

Summary

Unconstrained optimization: Steepest descent

Consider the unconstrained optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{where} \quad f : \mathbb{R}^n \to \mathbb{R} \quad \text{is} \quad \mathcal{C}^1.$$

Let us focus exclusively on a steepest descent framework:

Algorithm SD Steepest Descent

Require: $x_1 \in \mathbb{R}^n$

for $k \in \mathbb{N}$ **do**

    Compute $g_k \leftarrow \nabla f(x_k)$

    Choose $\alpha_k \in (0, \infty)$

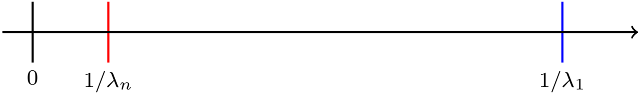
    Set $x_{k+1} \leftarrow x_k - \alpha_k g_k$

end for

All that remains to be determined are the stepsizes $\{\alpha_k\}$.

Minimizing strongly convex quadratics

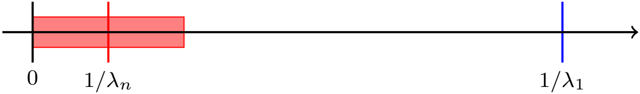
Suppose $f(x) = \frac{1}{2}x^T Ax - b^T x$, where A has eigenvalues $\lambda_n \geq \cdots \geq \lambda_1 > 0$.



Convergence (rate) of the algorithm depends on choices for $\{\alpha_k\}$.

Minimizing strongly convex quadratics

Suppose $f(x) = \frac{1}{2}x^T Ax - b^T x$, where A has eigenvalues $\lambda_n \geq \cdots \geq \lambda_1 > 0$.



Choosing $\alpha_k \leftarrow 1/\lambda_n$ leads to Q-linear convergence with constant $(1 - \lambda_1/\lambda_n)$

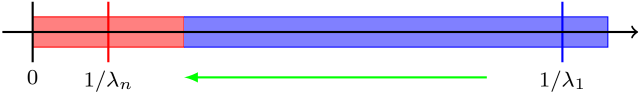
Minimizing strongly convex quadratics

Suppose $f(x) = \frac{1}{2}x^T Ax - b^T x$, where A has eigenvalues $\lambda_n \geq \cdots \geq \lambda_1 > 0$.

...but certain "components" of the gradient vanish in a larger range.

Minimizing strongly convex quadratics

Suppose $f(x) = \frac{1}{2}x^T Ax - b^T x$, where A has eigenvalues $\lambda_n \geq \cdots \geq \lambda_1 > 0$.



Goal: Allow large stepsizes, shrink range (automatically) to catch entire gradient.

Contributions

Consider Fletcher's limited memory steepest descent (LMSD) method.

- Extends the Barzilai-Borwein (BB) "two-point stepsize strategy".
- BB methods known to have R-linear convergence rate; Dai and Liao (2002).
- We prove that LMSD also attains R-linear convergence.

Although proved convergence rate is not necessarily better than that for BB, one can see reasons for improved empirical performance.

Outline

Decomposition

$$\boxed{\min_{x \in \mathbb{R}^n} f(x), \quad \text{where} \quad f(x) = \tfrac{1}{2} x^T A x - b^T x}$$

Let A have the eigendecomposition $A = Q \Lambda Q^T$, where

$$\begin{aligned} Q &= \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} \quad && \text{is orthogonal} \\ \text{and} \quad \Lambda &= \operatorname{diag}(\lambda_1, \ldots, \lambda_n) \quad && \text{with} \quad \lambda_n \geq \cdots \geq \lambda_1 > 0. \end{aligned}$$

Decomposition

$$\boxed{\min_{x \in \mathbb{R}^n} f(x), \quad \text{where} \quad f(x) = \tfrac{1}{2} x^T A x - b^T x}$$

Let A have the eigendecomposition $A = Q \Lambda Q^T$, where

$$\begin{aligned} Q &= \begin{bmatrix} q_1 & \cdots & q_n \end{bmatrix} && \text{is orthogonal} \\ \text{and} \quad \Lambda &= \operatorname{diag}(\lambda_1, \dots, \lambda_n) && \text{with} \quad \lambda_n \geq \cdots \geq \lambda_1 > 0. \end{aligned}$$

Let $g := \nabla f$. For any $x \in \mathbb{R}^n$, the gradient of f at x can be expressed as

$$g(x) = \sum_{i=1}^{n} d_i q_i, \quad \text{where} \quad d_i \in \mathbb{R} \quad \text{for all} \quad i \in [n] := \{1, \dots, n\}.$$

Recursion

Let $g := \nabla f$. For any $x \in \mathbb{R}^n$, the gradient of f at x can be expressed as

$$g(x) = \sum_{i=1}^{n} d_i q_i, \quad \text{where} \quad d_i \in \mathbb{R} \quad \text{for all} \quad i \in [n] := \{1, \ldots, n\}. \tag{1}$$

Recursion

Let $g := \nabla f$. For any $x \in \mathbb{R}^n$, the gradient of f at x can be expressed as

$$g(x) = \sum_{i=1}^{n} d_i q_i, \quad \text{where} \quad d_i \in \mathbb{R} \quad \text{for all} \quad i \in [n] := \{1, \ldots, n\}. \tag{1}$$

If $x^+ \leftarrow x - \alpha g(x)$, then the weights satisfy the recursive property:

$$d_i^+ = (1 - \alpha \lambda_i) d_i \quad \text{for all} \quad i \in [n].$$

Recursion

Let $g := \nabla f$. For any $x \in \mathbb{R}^n$, the gradient of f at x can be expressed as

$$g(x) = \sum_{i=1}^{n} d_i q_i, \quad \text{where} \quad d_i \in \mathbb{R} \quad \text{for all} \quad i \in [n] := \{1, \ldots, n\}. \tag{1}$$

If $x^+ \leftarrow x - \alpha g(x)$, then the weights satisfy the recursive property:

$$d_i^+ = (1 - \alpha\lambda_i) d_i \quad \text{for all} \quad i \in [n].$$

Proof (Sketch).

Since $g(x) = Ax - b$,

$$\begin{aligned} x^+ &= x - \alpha g(x) \\ Ax^+ &= Ax - \alpha g(x) \\ g(x^+) &= (I - \alpha A) g(x) \\ g(x^+) &= (I - \alpha Q \Lambda Q^T) g(x), \end{aligned}$$

then decompose according to (1).

Recursion

Let $g := \nabla f$. For any $x \in \mathbb{R}^n$, the gradient of f at x can be expressed as

$$g(x) = \sum_{i=1}^{n} d_i q_i, \quad \text{where } \; d_i \in \mathbb{R} \; \text{ for all } \; i \in [n] := \{1, \dots, n\}. \tag{1}$$

If $x^+ \leftarrow x - \alpha g(x)$, then the weights satisfy the recursive property:

$$d_i^+ = (1 - \alpha\lambda_i) d_i \; \text{ for all } \; i \in [n].$$

Proof (Sketch).

Since $g(x) = Ax - b$,

$$\begin{aligned} x^+ &= x - \alpha g(x) \\ Ax^+ &= Ax - \alpha g(x) \\ g(x^+) &= (I - \alpha A) g(x) \\ g(x^+) &= (I - \alpha Q \Lambda Q^T) g(x), \end{aligned}$$

then decompose according to (1).

Idea: Choose stepsizes as reciprocals of (estimates of) eigenvalues of A.

LMSD method: Main idea

Fletcher (2012):

- Repeated cycles (or "sweeps") of m iterations.
- At start of $(k+1)$st cycle, suppose one has the kth cycle values in

$$G_k := \begin{bmatrix} g_{k,1} & \cdots & g_{k,m} \end{bmatrix} \text{ corresponding to } \{x_{k,1}, \dots, x_{k,m}\}.$$

- Iterate displacements lie in Krylov sequence initiated from $g_{k,1}$.

LMSD method: Main idea

Fletcher (2012):

- Repeated cycles (or "sweeps") of m iterations.
- At start of $(k+1)$st cycle, suppose one has the kth cycle values in

$$G_k := \begin{bmatrix} g_{k,1} & \cdots & g_{k,m} \end{bmatrix} \text{ corresponding to } \{x_{k,1}, \dots, x_{k,m}\}.$$

- Iterate displacements lie in Krylov sequence initiated from $g_{k,1}$.
- Performing a QR decomposition to obtain

$$G_k = Q_k R_k,$$

one obtains m eigenvalue estimates (Ritz values) as eigenvalues of

$$\text{(symmetric tridiagonal)} \;\; T_k \leftarrow Q_k^T A Q_k,$$

which are contained in the spectrum of A in an optimal sense (more later).
- One can also obtain these estimates more cheaply and with less storage...

LMSD method: Efficient eigenvalue estimation

Storing the kth cycle reciprocal stepsizes in

$$J_k \leftarrow \begin{bmatrix} \alpha_{k,1}^{-1} & & \\ -\alpha_{k,1}^{-1} & \ddots & \\ & \ddots & \alpha_{k,m}^{-1} \\ & & -\alpha_{k,m}^{-1} \end{bmatrix},$$

one finds that by computing the (partially extended) Cholesky factorization

$$G_k^T \begin{bmatrix} G_k & g_{k,m+1} \end{bmatrix} = R_k^T \begin{bmatrix} R_k & r_k \end{bmatrix},$$

one has

$$T_k \leftarrow \begin{bmatrix} R_k & r_k \end{bmatrix} J_k R_k^{-1}.$$

Long story short: One can obtain Ritz values (and stepsizes) in $\sim \frac{1}{2}m^2 n$ flops

- . . . and this is done only once every m steps.
- Hence, costs $\mathcal{O}(mn)$ per iteration, like limited memory quasi-Newton.

LMSD

Algorithm LMSD Limited Memory Steepest Descent

Require: $x_{1,1} \in \mathbb{R}^n$, $m \in \mathbb{N}$, and $\epsilon \in \mathbb{R}_+$

Choose stepsizes $\{\alpha_{1,j}\}_{j\in[m]} \subset \mathbb{R}_{++}$
Compute $g_{1,1} \leftarrow \nabla f(x_{1,1})$
if $\|g_{1,1}\| \leq \epsilon$, **then return** $x_{1,1}$
for $k \in \mathbb{N}$ **do**
    for $j \in [m]$ **do**
        Set $x_{k,j+1} \leftarrow x_{k,j} - \alpha_{k,j} g_{k,j}$
        Compute $g_{k,j+1} \leftarrow \nabla f(x_{k,j+1})$
        if $\|g_{k,j+1}\| \leq \epsilon$, **then return** $x_{k,j+1}$
    end for
    Set $x_{k+1,1} \leftarrow x_{k,m+1}$ and $g_{k+1,1} \leftarrow g_{k,m+1}$
    Set G_k and J_k
    Compute (R_k, r_k), then compute T_k
    Compute $\{\theta_{k,j}\}_{j\in[m]} \subset \mathbb{R}_{++}$ as the eigenvalues of T_k
    Compute $\{\alpha_{k+1,j}\}_{j\in[m]} \leftarrow \{\theta_{k,j}^{-1}\}_{j\in[m]} \subset \mathbb{R}_{++}$
end for

(Note: There is also a version using harmonic Ritz values.)

Known convergence properties

BB methods ($m = 1$):

- R-superlinear when $n = 2$; Barzilai and Borwein (1988)
- Convergent for any n from any starting point; Raydan (1993)
- R-linear for any n; Dai and Liao (2002)

LMSD methods ($m \geq 1$):

- Convergent for any n from any starting point; Fletcher (2012)
- Prior to our work: Convergence rate not yet analyzed.

Outline

Basic Assumptions

Assumption 1

(i) *Algorithm LMSD is run with* $\epsilon = 0$ *and* $g_{k,j} \neq 0$ *for all* $(k,j) \in \mathbb{N} \times [m]$.

(ii) *For all* $k \in \mathbb{N}$, *the matrix* G_k *has linearly independent columns. Further, there exists* $\rho \in [1, \infty)$ *such that, for all* $k \in \mathbb{N}$,

$$\|R_k^{-1}\| \leq \rho \|g_{k,1}\|^{-1}. \tag{2}$$

To justify (2), note that when $m = 1$, one has

$$Q_k R_k = G_k = g_{k,1} \quad \text{where} \quad Q_k = g_{k,1}/\|g_{k,1}\| \quad \text{and} \quad R_k = \|g_{k,1}\|.$$

Hence, (2) holds with $\rho = 1$.

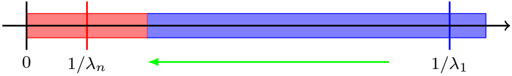
Intuition

Lemma 2

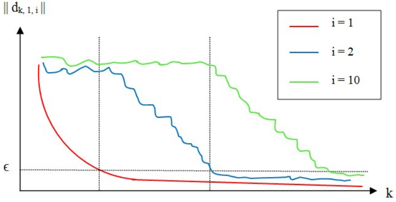
For all $k \in \mathbb{N}$, the eigenvalues of T_k satisfy

$$\theta_{k,j} \in [\lambda_{m+1-j}, \lambda_{n+1-j}] \subseteq [\lambda_1, \lambda_n] \quad \textit{for all} \quad j \in [m].$$

Recall...



We essentially prove that...



Worst-case “blow-up” of weights over a cycle

Lemma 3

For each $(k, j, i) \in \mathbb{N} \times [m] \times [n]$*:*

$$|d_{k,j+1,i}| \leq \delta_{j,i}|d_{k,j,i}| \quad \textit{where} \quad \delta_{j,i} := \max\left\{\left|1 - \frac{\lambda_i}{\lambda_{m+1-j}}\right|, \left|1 - \frac{\lambda_i}{\lambda_{n+1-j}}\right|\right\}.$$

Hence, for each $(k, j, i) \in \mathbb{N} \times [m] \times [n]$*:*

$$|d_{k+1,j,i}| \leq \Delta_i|d_{k,j,i}| \quad \textit{where} \quad \Delta_i := \prod_{j=1}^{m} \delta_{j,i}.$$

Furthermore, for each $(k, j, p) \in \mathbb{N} \times [m] \times [n]$*:*

$$\sqrt{\sum_{i=1}^{p} d_{k,j+1,i}^2} \leq \hat{\delta}_{j,p}\sqrt{\sum_{i=1}^{p} d_{k,j,i}^2} \quad \textit{where} \quad \hat{\delta}_{j,p} := \max_{i\in[p]} \delta_{j,i},$$

while, for each $(k, j) \in \mathbb{N} \times [m]$*:*

$$\|g_{k+1,j}\| \leq \Delta\|g_{k,j}\| \quad \textit{where} \quad \Delta := \max_{i\in[n]} \Delta_i.$$

Q-linear convergence of weight $i = 1$

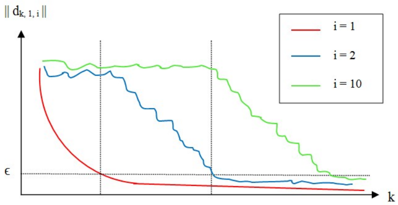
Lemma 4

If $\Delta_1 = 0$, *then* $d_{1+\hat{k},\hat{j},1} = 0$ *for all* $(\hat{k}, \hat{j}) \in \mathbb{N} \times [m]$. *Otherwise, if* $\Delta_1 > 0$, *then:*

(i) *for* $(k, j) \in \mathbb{N} \times [m]$ *with* $d_{k,j,1} = 0$, *it follows that* $d_{k+\hat{k},\hat{j},1} = 0$ *for all* $(\hat{k}, \hat{j}) \in \mathbb{N} \times [m]$;

(ii) *for* $(k, j) \in \mathbb{N} \times [m]$ *with* $|d_{k,j,1}| > 0$ *and any* $\epsilon_1 \in (0, 1)$, *it follows that*

$$\frac{|d_{k+\hat{k},\hat{j},1}|}{|d_{k,j,1}|} \leq \epsilon_1 \quad \textit{for all} \quad \hat{k} \geq 1 + \left\lceil \frac{\log \epsilon_1}{\log \Delta_1} \right\rceil \quad \textit{and} \quad \hat{j} \in [m].$$



Ritz value representation

Lemma 5

For all $(k, j) \in \mathbb{N} \times [m]$, let $q_{k,j} \in \mathbb{R}^m$ denote the unit eigenvector corresponding to the eigenvalue $\theta_{k,j}$ of T_k, i.e., that with $T_k q_{k,j} = \theta_{k,j} q_{k,j}$ and $\|q_{k,j}\| = 1$. Then, defining

$$D_k := \begin{bmatrix} d_{k,1,1} & \cdots & d_{k,m,1} \\ \vdots & \ddots & \vdots \\ d_{k,1,n} & \cdots & d_{k,m,n} \end{bmatrix} \quad \text{and} \quad c_{k,j} := D_k R_k^{-1} q_{k,j},$$

it follows that, with the diagonal matrix of eigenvalues (namely, $\Lambda = Q^T A Q$),

$$\theta_{k,j} = c_{k,j}^T \Lambda c_{k,j} \quad \text{and} \quad c_{k,j}^T c_{k,j} = 1.$$

“If first p weights small, then bound for weight $p+1\ldots$”

(We express $\hat{\delta}_p \in [1,\infty)$ dependent only on m, p, and the spectrum of A.)

Lemma 6 (simplified)

For any $(k,p) \in \mathbb{N} \times [n-1]$, *if there exists* $(\epsilon_p, K_p) \in (0, \frac{1}{2\hat{\delta}_p \rho}) \times \mathbb{N}$ *with*

$$\sum_{i=1}^{p} d_{k+\hat{k},1,i}^2 \leq \epsilon_p^2 \|g_{k,1}\|^2 \quad \textit{for all} \quad \hat{k} \geq K_p,$$

then there exists $K_{p+1} \geq K_p$ *dependent only on* ϵ_p, ρ, *and the spectrum of* A *with*

$$d_{k+K_{p+1},1,p+1}^2 \leq 4\hat{\delta}_p^2 \rho^2 \epsilon_p^2 \|g_{k,1}\|^2;$$

Proof (Key step).

First p elements of $c_{k+\hat{k},j}$ small enough such that

$$\theta_{k+\hat{k},j} = \sum_{i=1}^{n} \lambda_i c_{k+\hat{k},j,i}^2 \geq \tfrac{3}{4}\lambda_{p+1} \quad \text{for} \quad \hat{k} \geq K_p \quad \text{and} \quad j \in [m].$$

"If first p weights small, then bound for all first $p+1$ weights..."

Lemma 7

For any $(k,p) \in \mathbb{N} \times [n-1]$, *if there exists* $(\epsilon_p, K_p) \in (0, \frac{1}{2\hat{\delta}_p \rho}) \times \mathbb{N}$ *with*

$$\sum_{i=1}^{p} d_{k+\hat{k},1,i}^2 \leq \epsilon_p^2 \|g_{k,1}\|^2 \quad \text{for all} \quad \hat{k} \geq K_p,$$

then, with $\epsilon_{p+1}^2 := (1 + 4\max\{1, \Delta_{p+1}^4\}\hat{\delta}_p^2 \rho^2)\epsilon_p^2$ *and* $K_{p+1} \in \mathbb{N}$,

$$\sum_{i=1}^{p+1} d_{k+\hat{k},1,i}^2 \leq \epsilon_{p+1}^2 \|g_{k,1}\|^2 \quad \text{for all} \quad \hat{k} \geq K_{p+1}.$$

R-linear convergence of LMSD

Lemma 8

There exists $K \in \mathbb{N}$ dependent only on the spectrum of A such that

$$\|g_{k+K,1}\| \leq \tfrac{1}{2}\|g_{k,1}\| \quad \textit{for all} \quad k \in \mathbb{N}.$$

Theorem 9

The sequence $\{\|g_{k,1}\|\}$ vanishes R-linearly in the sense that

$$\|g_{k,1}\| \leq c_1 c_2^k \|g_{1,1}\|,$$

where

$$c_1 := 2\Delta^{K-1} \quad \textit{and} \quad c_2 := 2^{-1/K} \in (0,1).$$

Outline

Numerical demonstrations with $n = 100$: $m = 1$ and $m = 5$

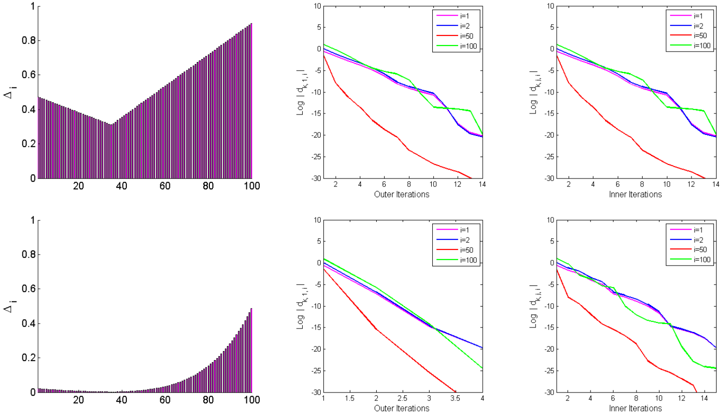


Figure: $\{\lambda_1, \ldots, \lambda_{100}\} \subset [1, 1.9]$

Numerical demonstrations with $n = 100$: $m = 1$ and $m = 5$

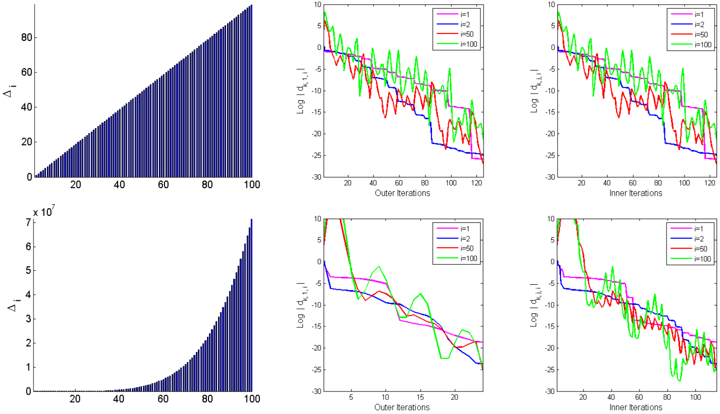


Figure: $\{\lambda_1, \ldots, \lambda_{100}\} \subset [1, 100]$

Numerical demonstrations with $n = 100$: $m = 1$ and $m = 5$

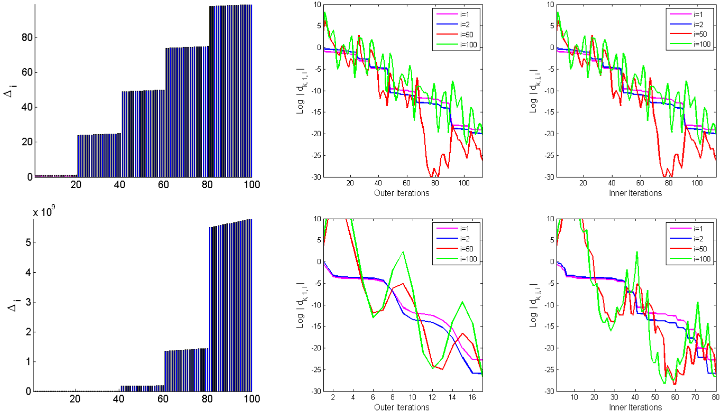


Figure: $\{\lambda_1, \ldots, \lambda_{100}\} \subset 5$ clusters, $m = 5$

Numerical demonstrations with $n = 100$: $m = 1$ and $m = 5$

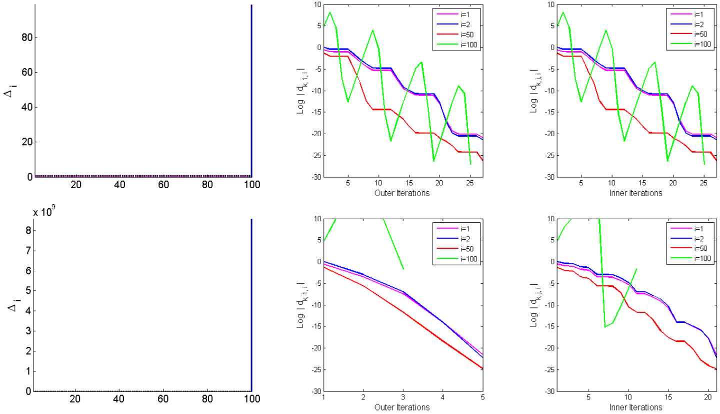


Figure: $\{\lambda_1, \ldots, \lambda_{100}\} \subset 2$ clusters (low heavy), $m = 5$

Numerical demonstrations with $n = 100$: $m = 1$ and $m = 5$

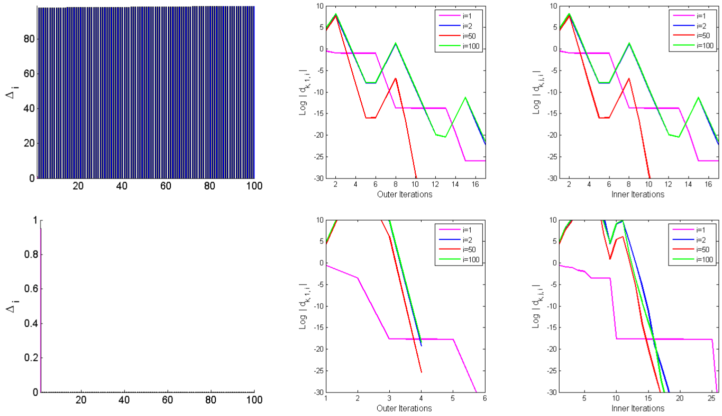


Figure: $\{\lambda_1, \ldots, \lambda_{100}\} \subset 2$ clusters (high heavy), $m = 5$

Outline

Summary

Consider Fletcher's limited memory steepest descent (LMSD) method.

- Extends the Barzilai-Borwein (BB) "two-point stepsize strategy".
- BB methods known to have R-linear convergence rate; Dai and Liao (2002).
- We prove that LMSD also attains R-linear convergence.

Although proved convergence rate is not necessarily better than that for BB, one can see reasons for improved empirical performance.

Soon in *IMA Journal of Numerical Analysis*: `https://doi.org/10.1093/imanum/drx016`