Spectral properties of steplength selections in gradient methods: from unconstrained to constrained optimization

L. Zanni

Department of Physics, Informatics and Mathematics,
University of Modena and Reggio Emilia, Italy

Variational Methods and Optimization in Imaging

IHP - Paris, 4 - 8 February 2019

Joint work with:

S. Crisci, V. Ruggiero, University of Ferrara, Italy

F. Porta, University of Modena and Reggio Emilia, Italy

Outline

1. Gradient methods for unconstrained problems
   - Spectral properties of steplength selections
   - **Design selection rules by exploiting spectral properties**
   - From the quadratic case to general unconstrained problems
2. Gradient projection methods for box-constrained problems
   - Spectral properties of steplengths in the quadratic case
   - **New steplength rules taking into account the constraints**
3. Scaled gradient projection methods
   - Define the diagonal scaling
   - **The steplengths in variable metric approaches**
   - Practical behaviour in imaging
4. Conclusions

Motivation for the steplength analysis

Constrained optimization problems

$$\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x}) \tag{1}$$

- $f : \mathbb{R}^N \longrightarrow \mathbb{R}$ continuously differentiable function
- $\Omega \subset \mathbb{R}^N$, nonempty closed convex set defined by simple constraints

Gradient Projection (GP) methods for $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \vartheta_k \boldsymbol{d}^{(k)} \\ \boldsymbol{d}^{(k)} &= P_\Omega\left(\boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})\right) - \boldsymbol{x}^{(k)} \end{aligned}$$

$$\alpha_k > 0, \qquad \vartheta_k \in (0,1], \qquad P_\Omega(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{z}\in\Omega} \|\boldsymbol{z}-\boldsymbol{x}\|$$

Usually the updating rules for the steplength α_k are those exploited in the unconstrained case: is this a suitable choice?

Spectral analysis of steplength selections

- The unconstrained case
- The box-constrained case
- The Scaled Gradient Projection methods

Steplength selection: **the unconstrained case**

The recipe exploited by state-of-the-art selection rules:

- define steplengths by trying to capture, in an inexpensive way, some second order information
- design selection rules in the strictly convex quadratic case:

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x}, \qquad \text{A symmetric positive definite}$$

  second order information $\leftrightarrow$ spectral properties of A

- design selection rules that generalize, in an inexpensive way, to non-quadratic cases

  $\nabla^2 f(\boldsymbol{x}^{(k)})$ depends on the iterations but $\nabla^2 f(\boldsymbol{x}^{(k)}) \to \nabla^2 f(\boldsymbol{x}^*)$

A popular example: the Barzilai-Borwein (BB) selection rules

Consider the gradient method for the problem $\min f(\boldsymbol{x})$:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)}) \qquad k = 0, 1, \ldots \ ,$$

Suggestion [Barzilai-Borwein, IMA J. Num. Anal. 1988]:

Force the matrix $(\alpha_k I)^{-1}$ to approximate the Hessian $\nabla^2 f(\boldsymbol{x}^{(k)})$ by imposing quasi-Newton properties

$$\alpha_k^{\text{BB1}} = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \|(\alpha I)^{-1} \boldsymbol{s}^{(k-1)} - \boldsymbol{z}^{(k-1)}\| = \frac{{\boldsymbol{s}^{(k-1)}}^T \boldsymbol{s}^{(k-1)}}{{\boldsymbol{s}^{(k-1)}}^T \boldsymbol{z}^{(k-1)}}$$

or

$$\alpha_k^{\text{BB2}} = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \|\boldsymbol{s}^{(k-1)} - (\alpha I) \boldsymbol{z}^{(k-1)}\| = \frac{{\boldsymbol{s}^{(k-1)}}^T \boldsymbol{z}^{(k-1)}}{{\boldsymbol{z}^{(k-1)}}^T \boldsymbol{z}^{(k-1)}}$$

where $\boldsymbol{s}^{(k-1)} = \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right)$, $\quad \boldsymbol{z}^{(k-1)} = (\nabla f(\boldsymbol{x}^{(k)}) - \nabla f(\boldsymbol{x}^{(k-1)}))$.

Spectral properties of the BB steplength rules

Consider a gradient method for the quadratic unconstrained case:

$$\min f(\boldsymbol{x}) \equiv \frac{1}{2}\boldsymbol{x}^T A\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}, \quad A = diag(\lambda_1, \ldots, \lambda_N), \;\; 0 < \lambda_1 < \cdots < \lambda_N$$

$$\boxed{\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)}, \qquad \boldsymbol{g}^{(k)} = \nabla f(\boldsymbol{x}^{(k)}), \qquad k = 0, 1, \ldots}$$

$$\Downarrow$$

$$\boxed{g_i^{(k+1)} = (1 - \alpha_k \lambda_i) g_i^{(k)} \qquad i = 1, \ldots, N}$$

- $\alpha_k = \frac{1}{\lambda_i} \quad \Rightarrow \quad g_i^{(k+1)} = 0 \quad \Rightarrow \quad g_i^{(k+j)} = 0, \quad j = 2, 3 \ldots$
- $\alpha_{k+i-1} = \frac{1}{\lambda_i}, \;\; i = 1, \ldots, N \quad \Rightarrow \quad \boldsymbol{g}^{(k+N)} = 0$ (Finite Termination)

α_k must aim at approximating the inverse of the eigenvalues of A

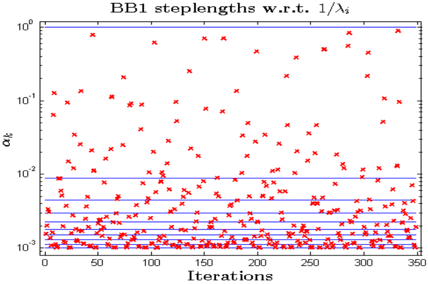
BB rules in the quadratic case

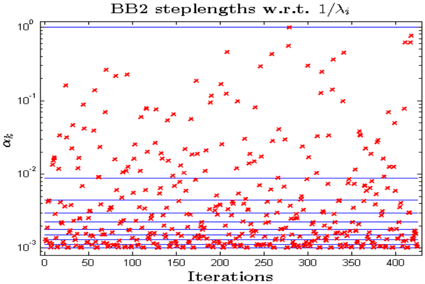
$$\frac{1}{\lambda_N} \leq \alpha_k^{\text{BB2}} = \frac{\boldsymbol{g}^{(k-1)^T} A \boldsymbol{g}^{(k-1)}}{\boldsymbol{g}^{(k-1)^T} A^2 \boldsymbol{g}^{(k-1)}} \leq \alpha_k^{\text{BB1}} = \frac{\boldsymbol{g}^{(k-1)^T} \boldsymbol{g}^{(k-1)}}{\boldsymbol{g}^{(k-1)^T} A \boldsymbol{g}^{(k-1)}} \leq \frac{1}{\lambda_1}$$

Example

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x}$$

- $A = diag(\lambda_1, \ldots, \lambda_{10}), \quad \lambda_i = 111i - 110$
- b random vector; $b_i \in [-10, 10]$
- stopping rule: $\|\boldsymbol{g}^{(k)}\| \leq 10^{-8} \|\boldsymbol{g}^{(0)}\|$





Quadratic case: exploiting spectral properties

In the quadratic case $\left(A = diag(\lambda_1, \ldots, \lambda_N), \quad 0 < \lambda_1 < \cdots < \lambda_N\right)$, we have

- $$g_j^{(k+1)} = (1 - \alpha_k \lambda_j) g_j^{(k)} \quad j = 1, \ldots, N$$

- $$\alpha_k \approx \frac{1}{\lambda_i} \quad \Rightarrow \quad \begin{cases} \left|g_i^{(k+1)}\right| \ll \left|g_i^{(k)}\right| & \text{very useful} \\ \left|g_j^{(k+1)}\right| < \left|g_j^{(k)}\right| \quad \text{if} \quad j < i & \text{useful} \\ \left|g_j^{(k+1)}\right| > \left|g_j^{(k)}\right| \quad \text{if} \quad j > i, \quad \lambda_j > 2\lambda_i & \text{dangerous} \end{cases}$$

- $$\alpha_k^{\text{BB2}} / \alpha_k^{\text{BB1}} = cos^2(\boldsymbol{g}^{(k-1)}, A\boldsymbol{g}^{(k-1)})$$

Idea for improving the BB rules:

- force a sequence of small α_k^{BB2} to reduce $|g_i|$ for large i, leading to gradients in which these components are not dominant
- after a sequence of small α_k, if $\alpha_k^{\text{BB2}} / \alpha_k^{\text{BB1}} \approx 1$, exploit $\alpha^{\text{BB1}} = \frac{\boldsymbol{g}^T \boldsymbol{g}}{\boldsymbol{g}^T A \boldsymbol{g}}$ aiming at obtaining $\alpha^{\text{BB1}} \approx 1/\lambda_i$ for small i

Practical implementations of this idea: ABB and ABB$_{\text{min}}$ rules

Alternate Barzilai-Borwein selection rule [Zhou-Gao-Dai, *COAP (2006)*]

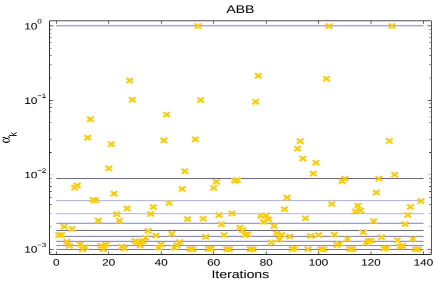
$$\alpha_k^{ABB} = \begin{cases} \alpha_k^{BB2} & \text{if } \quad \dfrac{\alpha_k^{BB2}}{\alpha_k^{BB1}} < \tau, \qquad \tau \in (0,1) \\ \alpha_k^{BB1} & \text{otherwise} \end{cases}$$

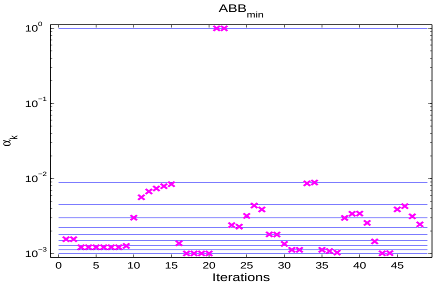
ABB$_{\text{min}}$ rule [Frassoldati-Zanghirati-Zanni, *JIMO (2008)*]

$$\alpha_k^{ABB_{\text{min}}} = \begin{cases} \min\left\{\alpha_j^{BB2} \,|\, j = \max\{1, k - M_\alpha\}, ..., k\right\} & \text{if } \alpha_k^{BB2} / \alpha_k^{BB1} < \tau \\ \alpha_k^{BB1} & \text{otherwise} \end{cases}$$

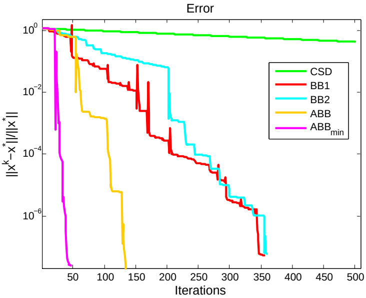
where $M_\alpha > 0$ is a parameter.

ABB and ABB$_{\text{min}}$ rules on the previous toy problem





- Cauchy Steepest Descent (CSD)
  $\alpha_k = \text{argmin}_{\alpha>0}\, f(\boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}^{(k)})$
- BB1 $\rightarrow$ $\alpha_k = \alpha_k^{BB1}$
- BB2 $\rightarrow$ $\alpha_k = \alpha_k^{BB2}$
- ABB $\rightarrow$ alternation
- ABB$_{\text{min}}$ $\rightarrow$ modified alternation

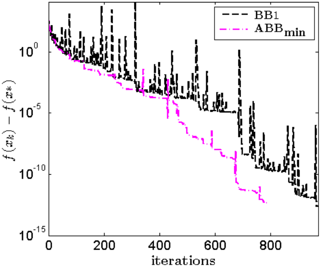


Similar behaviour on randomly generated test problems

Quadratic test problems: $N = 1000$

$$\lambda_1 = 1, \qquad \lambda_N = 10^4,$$

$$\lambda_i,\ i = 2, \ldots, N-1, \text{ log-spaced}$$



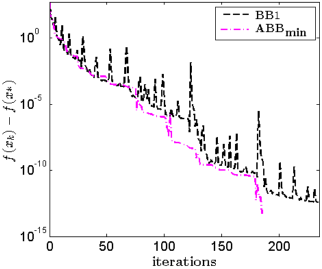
$$\underline{\lambda} = 1, \qquad \overline{\lambda} = 10^3,$$

$$\lambda_i = \underline{\lambda} + (\overline{\lambda} - \underline{\lambda}) * s_i, \quad i = 1, \ldots, N,$$

$$s_i \in (0, 0.2), \quad i = 1, \ldots, N/2,$$

$$s_i \in (0.8, 1),\ i = N/2 + 1, \ldots, N.$$

[Di Serafino-Ruggiero-Toraldo-Z., AMC 2018]



Other efficient steplength rules based on spectral properties

[Pronzato-Zhigljavsky, Comput. Optim. Appl. 50 (2011)]

[Fletcher, Math. Program. Ser. A 135 (2012)]

[Pronzato-Zhigljavsky-Bukina, Acta Appl. Math. 127 (2013)]

[De Asmundis-Di Serafino-Riccio-Toraldo, IMA J. Numer. Anal. 33 (2013)]]

[De Asmundis-Di Serafino-Hager-Toraldo-Zhan, Comput. Optim. Appl. 59 (2014)]

[Gonzaga-Schneider, Comput. Optim. Appl. 63 (2016)]

[Gonzaga, Math. Program. Ser. A 160 (2016)]

- Aimed at breaking the well-known cycling behaviour of the Steepest Descent method
- they share R-linear convergence rate in the quadratic case
- not all these rules easily generalize to general non-quadratic problems (BB-based rules have this crucial property)

General unconstrained problems: $\min_{\boldsymbol{x}\in\mathbb{R}^N} f(\boldsymbol{x})$

Gradient methods with nonmonotone linesearch:

Init.: $\boldsymbol{x}^{(0)} \in \mathbb{R}^N$, $0 < \alpha_{min} \le \alpha_{max}$, $\alpha_0 \in [\alpha_{min}, \alpha_{max}]$, $\delta, \sigma \in (0,1)$, $M \in \mathbb{N}$;

for $k = 0, 1, \ldots$

$\nu_k = \alpha_k$; $f_{ref} = \max\{f(\boldsymbol{x}^{(k-j)}),\ 0 \le j \le \min(k, M)\}$;

while $f(\boldsymbol{x}^{(k)} - \nu_k \boldsymbol{g}^{(k)}) > f_{ref} - \sigma \nu_k {\boldsymbol{g}^{(k)}}^T \boldsymbol{g}^{(k)}$ (line search)

$\nu_k = \delta \nu_k$;

end

$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \nu_k \boldsymbol{g}^{(k)}$;

define a tentative steplength $\alpha_{k+1} \in [\alpha_{min}, \alpha_{max}]$

end

- tentative steplength: exploit effective steplength selections designed for the quadratic case and generalizable in an inexpensive way.
- R-linear convergence of $\{f(\boldsymbol{x}^{(k)})\}$ when f is strongly convex with Lipschitz-cont. gradient ([Dai, JOTA 2002], [Dai-Liao, IMA J.Num.Anal. 2002])

The standard BB rules can be improved

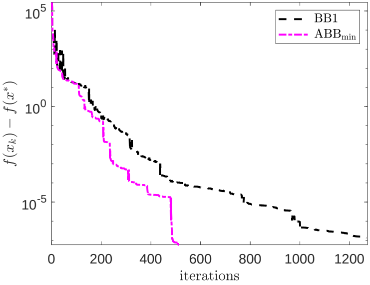
Trigonometric test problems: $n = 50$

$$f(x) = \|\boldsymbol{b} - (A\boldsymbol{v}(\boldsymbol{x}) + B\boldsymbol{u}(\boldsymbol{x}))\|^2,$$

$$\boldsymbol{v}(\boldsymbol{x}) = (\sin(x_1), ..., \sin(x_n))^T,$$

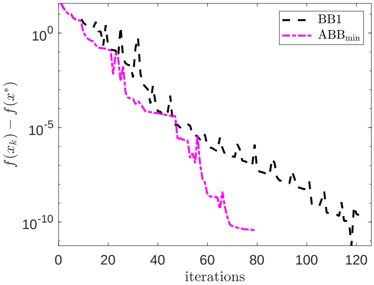
$$\boldsymbol{u}(\boldsymbol{x}) = (\cos(x_1), ..., \cos(x_n))^T,$$

A, B $n \times n$ random matrices
integer entries in $(-100, 100)$



Convex2 test problems: $N = 100$

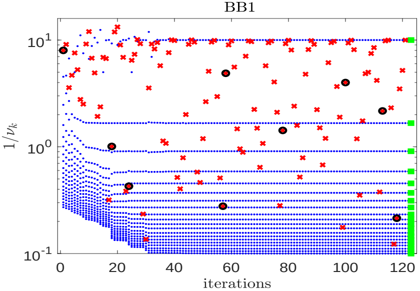
$$f(x) = \sum_{i=1}^{n} \frac{i}{10}(e^{x_i} - x_i);$$

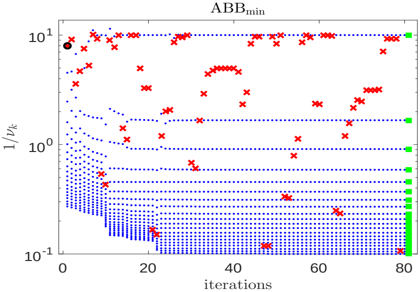


The steplengths mimic the behaviour in the quadratic case

Convex2 test problems: $N = 100$

Green squares: 20 eigenvalues of $\nabla^2 f(\boldsymbol{x}^*)$ with linearly spaced indices
Blue dots: 20 eigenvalues of $\nabla^2 f(\boldsymbol{x}^{(k)})$ with linearly spaced indices
Red cross: $\frac{1}{\nu_k}$ (black circles mean linesearch reductions)





When the Hessian eigenvalues stabilize, the steplengths exhibit the spectral properties observed in the quadratic case, but with respect to the spectrum of the current Hessian.

Spectral analysis of steplength selections

- The unconstrained case
- The box-constrained case
- The Scaled Gradient Projection methods

What about the **constrained case**?

Constrained minimization problems: $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \vartheta_k \boldsymbol{d}^{(k)}, \qquad \boldsymbol{d}^{(k)} = P_\Omega\left(\boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})\right) - \boldsymbol{x}^{(k)}$$

$$P_\Omega(\boldsymbol{x}) = \text{argmin}_{\boldsymbol{z}\in\Omega} \|\boldsymbol{z} - \boldsymbol{x}\|, \quad \Omega \subset \mathbb{R}^N$$

More difficult analysis

- The goal is no more the gradient annihilation
- The gradient projection step makes the relation between successive gradients more complicated

Motivation for generalizing the analysis

- BB rules are considered very effective also in the constrained case and were successfully exploited in many interesting applications
- ABB strategies seem still to outperform standard BB rules

The simplest case: **box-constrained quadratic problems**

$$\min_{\boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}} f(x) \equiv \frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x}, \qquad A \text{ sym. pos. def.}, \quad \boldsymbol{l}, \boldsymbol{u} \in \mathbb{R}^n$$

Gradient Projection (GP) method

Init.: $\boldsymbol{\ell} \le \boldsymbol{x}^{(0)} \le \boldsymbol{u}$, $0 < \alpha_{min} \le \alpha_{max}$, $\alpha_0 \in [\alpha_{min}, \alpha_{max}]$, $\delta, \sigma \in (0,1)$, $M \in \mathbb{N}$;

for $k = 0, 1, \ldots$

$\boldsymbol{d}^{(k)} = P_{\boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}}\left(\boldsymbol{x}^{(k)} - \alpha_k \boldsymbol{g}(x^{(k)})\right) - \boldsymbol{x}^{(k)}$; (gradient projection step)

$\vartheta_k = 1$; $\quad f_{ref} = \max\{f(\boldsymbol{x}^{(k-j)}),\ 0 \le j \le \min(k, M)\}$;

while $f(\boldsymbol{x}^{(k)} + \vartheta_k \boldsymbol{d}^{(k)}) > f_{ref} + \sigma \vartheta_k \boldsymbol{g}^{(k)^T} \boldsymbol{d}^{(k)}$ (line search)

$\vartheta_k = \delta \vartheta_k$;

end

$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \vartheta_k \boldsymbol{d}^{(k)}$;

define the steplength $\alpha_{k+1} \in [\alpha_{min}, \alpha_{max}]$ (steplength updating rule)

end

Box-constrained QP: $\min_{\boldsymbol{\ell} \le \boldsymbol{x} \le \boldsymbol{u}} f(x) \equiv \frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - \boldsymbol{b}^T \boldsymbol{x}$

The solution $\boldsymbol{x}^*$ satisfies
$$\begin{cases} \boldsymbol{g}(\boldsymbol{x}^*)_i = 0 & \text{for} \quad \ell_i < x_i^* < u_i \quad (i \in \mathcal{I}^*) \\ \boldsymbol{g}(\boldsymbol{x}^*)_i \le 0 & \text{for} \quad x_i^* = u_i \quad (i \in \mathcal{J}^*) \\ \boldsymbol{g}(\boldsymbol{x}^*)_i \ge 0 & \text{for} \quad x_i^* = \ell_i \quad (i \in \mathcal{J}^*) \end{cases}$$

Define the set of indices

$$\begin{aligned} \mathcal{J}_{k-1} &= \{i : \ (x_i^{(k-1)} = \ell_i \ \wedge \ g_i^{(k-1)} \ge 0) \ \vee \ (x_i^{(k-1)} = u_i \ \wedge \ g_i^{(k-1)} \le 0)\} \\ \mathcal{I}_{k-1} &= \{1, ..., n\} \setminus \mathcal{J}_{k-1} \end{aligned}$$

Possible idea

Since $\boldsymbol{g}(\boldsymbol{x}^*)_i = 0, \quad i \in \mathcal{I}^*,$ exploit the steplength rules to accelerate the reduction of

$$|g_i^{(k-1)}|, \quad i \in \mathcal{I}_{k-1},$$

as done in the unconstrained case

Are the BB steplength rules useful to this purpose?

$$\alpha_k^{\text{BB1}} = \frac{{\boldsymbol{s}^{(k-1)}}^T \boldsymbol{s}^{(k-1)}}{{\boldsymbol{s}^{(k-1)}}^T \boldsymbol{z}^{(k-1)}}, \quad \alpha_k^{\text{BB2}} = \frac{{\boldsymbol{s}^{(k-1)}}^T \boldsymbol{z}^{(k-1)}}{{\boldsymbol{z}^{(k-1)}}^T \boldsymbol{z}^{(k-1)}}, \quad \begin{array}{l} \boldsymbol{s}^{(k-1)} = (\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}) \\ \boldsymbol{z}^{(k-1)} = (\boldsymbol{g}^{(k)}) - \boldsymbol{g}^{(k-1)}) \end{array}$$

What about $\boldsymbol{s}^{(k-1)}$? (observe that $x_j^{(k)} = x_j^{(k-1)}$, for $j \in \mathcal{J}_{k-1}$)

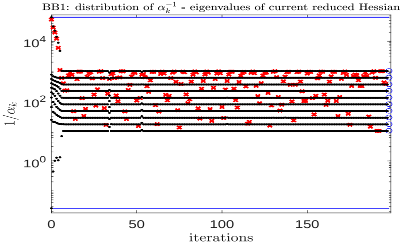
$$\boldsymbol{s}_{\mathcal{J}_{k-1}}^{(k-1)} = \mathbf{0} \;\Rightarrow\; \begin{cases} \alpha_k^{\text{BB1}} = \dfrac{{\boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)}}{{\boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}} = \underset{\alpha\in\mathbb{R}}{\operatorname{argmin}} \|(\alpha_k I)^{-1} \boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)} - \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}\| \\[2ex] \alpha_k^{\text{BB2}} = \dfrac{{\boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}{{\boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)} + {\boldsymbol{z}_{\mathcal{J}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{J}_{k-1}}^{(k-1)}} \end{cases}$$

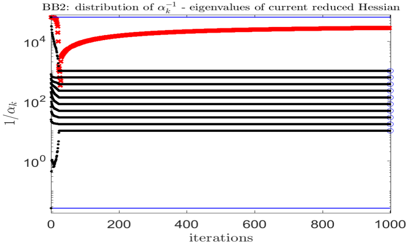
Only the α_k^{BB1} rule is able to capture the spectral properties of the Reduced Hessian $A_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}$ at the current iteration:

$$\lambda_{min}(A_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}) \;\leq\; 1/\alpha_k^{\text{BB1}} \;\leq\; \lambda_{max}(A_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}).$$

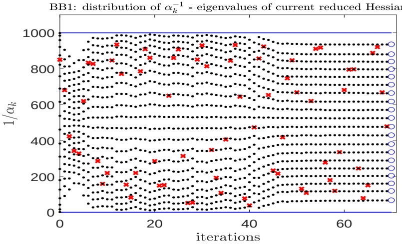
Box-constrained QP: different behaviour of α_k^{BB1} and α_k^{BB2}

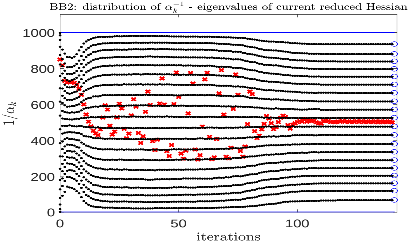
TP1: $n = 1000$, 500 active const., $\lambda_i(A_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}) \in [10, 10^3]$ log-spaced





TP2: $\lambda_i = \frac{M+m}{2} + \frac{M-m}{2} cos(\frac{\pi(i-1)}{n-1})$, $m = 1$, $M = 10^3$





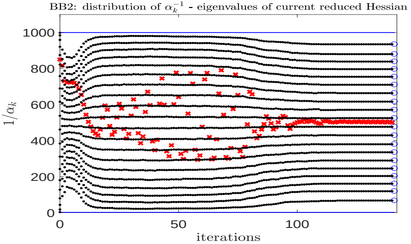
New proposals [Crisci-Ruggiero-Zanni, AMC 2019]

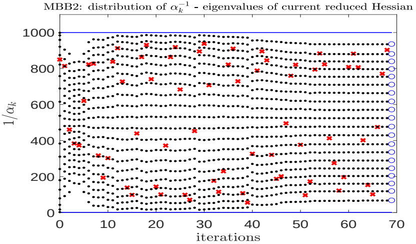
$$\alpha_k^{\text{BB2}} = \frac{{\boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}{{\boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)} + {\boldsymbol{z}_{\mathcal{J}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{J}_{k-1}}^{(k-1)}} \quad \rightarrow \quad \alpha_k^{\text{MBB2}} = \frac{{\boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}{{\boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}$$

Modified BB2 steplength rule

$$\lambda_{min}(A_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}) \le \frac{1}{\alpha_k^{\text{BB1}}} \le \frac{1}{\alpha_k^{\text{MBB2}}} \le \lambda_{max}(A_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}}).$$

TP2: $\lambda_i = \frac{M+m}{2} + \frac{M-m}{2} cos(\frac{\pi(i-1)}{n-1}),\qquad m = 1,\quad M = 10^3$





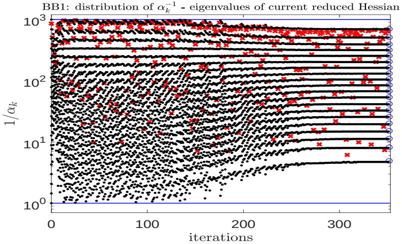
Modified BB2 can be exploited within ABB strategies

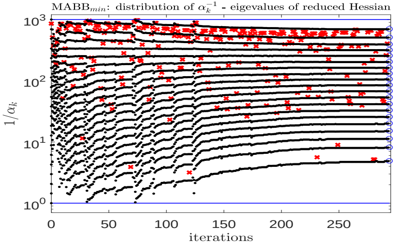
Modified ABB_{min} rule

$$\alpha_k^{MABB_{\text{min}}} = \begin{cases} \min\left\{\alpha_j^{MBB2} \mid j = \max\{1, k - M_\alpha\}, \ldots, k\right\} & \text{if } \alpha_k^{MBB2} / \alpha_k^{BB1} < \tau \\ \alpha_k^{BB1} & \text{otherwise} \end{cases}$$

where $M_\alpha > 0$ is a parameter.

TP3: $n = 1000$, 500 active const., $\lambda_i(A) \in [1, 10^3]$ log-spaced





Performance profile: box-constrained QP test problems

- Test Problems [Moré–Toraldo, Num. Math. 1989]

  108 box-const. QP, $15000 \leq n \leq 25000$,
  $K(A) = 10^4, 10^5, 10^6$, $n_{act} = 0.1n,\ 0.5n,\ 0.9n$

- Methods
  GP method with nonmonotone linesearch and different steplength rules (BB2, MBB2, ABB_{min}, MABB_{min} ...)

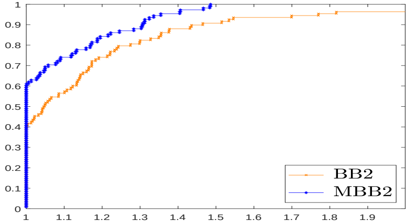
  stopping rules:

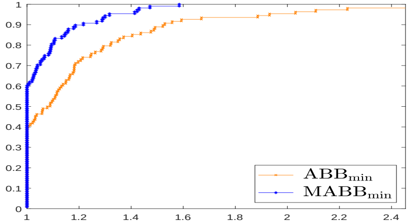
$$\|\varphi(\boldsymbol{x}^{(k)})\|_2 \leq 10^{-5}\|\nabla f(\boldsymbol{x}^{(0)})\|_2,$$

$$(\varphi(\boldsymbol{x}))_i = \begin{cases} (\nabla f(\boldsymbol{x}))_i\,, & \text{for } x_i \neq \ell_i \text{ and } x_i \neq u_i \\ \min\left(0, (\nabla f(\boldsymbol{x}))_i\right), & \text{for } x_i = \ell_i \\ \max\left(0, (\nabla f(\boldsymbol{x}))_i\right), & \text{for } x_i = u_i. \end{cases}$$

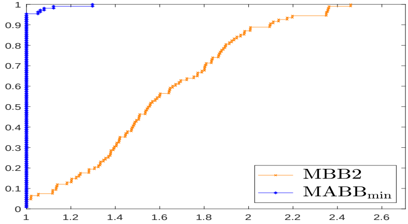
Performance profile ($x \leftarrow \frac{T_{solver}}{T_{min}}$, $y \leftarrow$ %prob. solved within xT_{min})

- The new MBB2 selection outperforms the standard BB2 rule





- Alternated strategies are preferable also in the constrained case

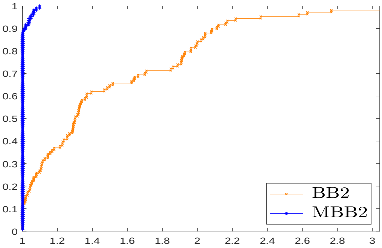


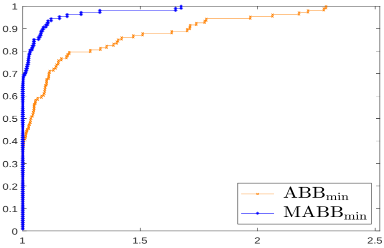
Test Problems [Facchinei-Judice-Soares, ACM TOMS 1997]

$$\min_{\boldsymbol{\ell}\leq\boldsymbol{x}\leq\boldsymbol{u}} f(\boldsymbol{x}) \equiv g(\boldsymbol{x}) + \sum_{i\in L} h_i(\boldsymbol{x_i}) - \sum_{i\in U} h_i(\boldsymbol{x_i}) \qquad \begin{cases} L = \{i \mid x_i^* = \ell_i\} \\ U = \{i \mid x_i^* = u_i\}, \end{cases}$$

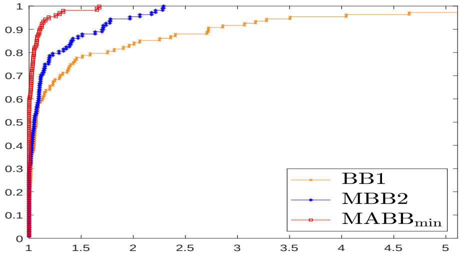
$$g(\boldsymbol{x}) = \begin{cases} \text{Trigonometric} \\ \text{Chained Rosenbrock} \\ \text{Laplace2} \end{cases} \qquad h_i(x_i) = \begin{cases} \beta_i(x_i - x_i^*) \\ \alpha_i\,(x_i - x_i^*)^3 + \beta_i\,(x_i - x_i^*) \\ \alpha_i\,(x_i - x_i^*)^{7/3} + \beta_i\,(x_i - x_i^*) \end{cases}$$

Problem size: $n = 500$; total number of problems: 108





Alternated BB are preferable



Alternated BB in practical applications

- Machine learning: decomp. techniques for training of Support Vector Machines [Serafini-Zanghirati-Z., Par. Comput. 2003, OMS 2005, JMLR 2006]
- Imaging problems in Astronomy, Microscopy, Computed Tomography
  [Bonettini-Zanella-Z., Inv. Prob. 2009], [Loris et al. ACHA 2009],
  [Ruggiero et al. JGO 2010], [Prato et al., A&A 2012],
  [Zanella et al., Sci. Rep. 2013], [Piccolomini et al. COAP 2018]

Spectral analysis of steplength selections

- The unconstrained case
- The box-constrained case
- The Scaled Gradient Projection methods

Basic variable metric approaches

In many imaging applications the behaviour of gradient projection schemes is largely improved by exploiting variable metric approaches:

Scaled Gradient Projection (SGP) methods for $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$

$$
\begin{aligned}
\boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \vartheta_k \boldsymbol{d}^{(k)}, \qquad \vartheta_k \in (0,1], \\
\boldsymbol{d}^{(k)} &= P_{\Omega, D_k^{-1}}\left(\boldsymbol{x}^{(k)} - \alpha_k D_k \nabla f(\boldsymbol{x}^{(k)})\right) - \boldsymbol{x}^{(k)}, \qquad \alpha_k > 0 \\
P_{\Omega, D_k^{-1}}(\boldsymbol{x}) &= \text{argmin}_{\boldsymbol{z}\in\Omega} \|\boldsymbol{z} - \boldsymbol{x}\|_{D_k^{-1}} \qquad D_k \text{ sym. pos. def. matrix} \\
\|\boldsymbol{z} - \boldsymbol{x}\|_{D_k^{-1}} &\equiv \sqrt{(\boldsymbol{z} - \boldsymbol{x})^T D_k^{-1} (\boldsymbol{z} - \boldsymbol{x})}
\end{aligned}
$$

- How can the matrix D_k be chosen?
- How can the steplength rules for α_k be modified for taking into account the scaling matrix?

SGP methods: convergence analysis

$$\begin{aligned} \boldsymbol{x}^{(k+1)} &= \boldsymbol{x}^{(k)} + \vartheta_k \boldsymbol{d}^{(k)}, \qquad \vartheta_k \in (0,1], \\ \boldsymbol{d}^{(k)} &= P_{\Omega, D_k^{-1}} \left(\boldsymbol{x}^{(k)} - \alpha_k D_k \nabla f(\boldsymbol{x}^{(k)}) \right) - \boldsymbol{x}^{(k)}, \qquad \alpha_k > 0 \end{aligned}$$

Analysis of $\{\boldsymbol{x}^{(k)}\}$ and $\{f(\boldsymbol{x}^{(k)})\}$ [Bonettini-Prato, *Inv. Prob. 2015*]

- D_k with eigenvalues in $\left[\frac{1}{\mu}, \mu\right], \ \mu \geq 1$
- $\alpha_k \in [\alpha_{min}, \alpha_{max}], \quad 0 < \alpha_{min} \leq \alpha_{max}$
- $\vartheta_k \ | \ \ f(\boldsymbol{x}^{(k+1)}) \leq f(\boldsymbol{x}^{(k)}) + \sigma \vartheta_k \nabla f(\boldsymbol{x}^{(k)})^T \boldsymbol{d}^{(k)}$

$\Rightarrow$ If $\boldsymbol{x}^{(k_l)} \xrightarrow{l \to \infty} \boldsymbol{x}^*$ then $\nabla f(\boldsymbol{x}^*)^T (\boldsymbol{x} - \boldsymbol{x}^*) \geq 0 \quad \forall \boldsymbol{x} \in \Omega$

- $f(\boldsymbol{x})$ convex, the solution set X^* not empty
- $\mu_k^2 = 1 + \gamma_k, \quad \gamma_k \geq 0, \quad \sum_{k=0}^{\infty} \gamma_k < \infty$
- D_k s.p.d. with eigenvalues in $\left[\frac{1}{\mu_k}, \mu_k\right]$

$\Rightarrow$ $\boldsymbol{x}^{(k)} \xrightarrow{k \to \infty} \boldsymbol{x}^*$, $\boldsymbol{x}^* \in X^*$

- ∇f is Lipschitz on Ω $\quad \Rightarrow \quad f^{(k)} - f^* = \mathcal{O}\left(\frac{1}{k}\right)$

Variable metric updating: the choice of the matrix D_k

- A standard choice: $D_k = \mathsf{diag}\left(D_1^{(k)}, D_2^{(k)}, \ldots, D_N^{(k)}\right)$

$$D_i^{(k)} = \min\left\{\rho, \max\left\{\frac{1}{\rho}, \left(\frac{\partial^2 f(\boldsymbol{x}^{(k)})}{(\partial x_i)^2}\right)^{-1}\right\}\right\}, \quad i = 1, \ldots, N,$$

- Define D_k by exploiting only first-order information

Consider the special non-negatively constrained case: $\min_{\boldsymbol{x} \geq \mathbf{0}} f(\boldsymbol{x})$ and the corresponding KKT conditions

$$\nabla f(\boldsymbol{x}) - \boldsymbol{\xi} = 0, \quad \boldsymbol{x} \geq \mathbf{0}, \quad \boldsymbol{\xi} \geq \mathbf{0}, \quad x_i \xi_i = 0, \quad i = 1, \ldots, N$$

$$\Downarrow$$

$$\boldsymbol{x} \cdot \nabla f(\boldsymbol{x}) = \mathbf{0}, \quad \boldsymbol{x} \geq \mathbf{0}, \quad \nabla f(\boldsymbol{x}) \geq \mathbf{0}$$

" $\cdot$ " denotes the component-wise product

Variable metric updating: the choice of the matrix D_k

Split the gradient [Lantéri-Roche-Aime, *Inv. Prob.* (2002)]:

$$\nabla f(\boldsymbol{x}) = V(\boldsymbol{x}) - U(\boldsymbol{x}), \quad V(\boldsymbol{x}) > 0, \quad U(\boldsymbol{x}) \geq 0$$

and use the splitting in the nonlinear equation $\boldsymbol{x} \cdot \nabla f(\boldsymbol{x}) = \mathbf{0}$:

$$\boldsymbol{x} \cdot V(\boldsymbol{x}) = \boldsymbol{x} \cdot U(\boldsymbol{x}) = \boldsymbol{x} \cdot (-\nabla f(\boldsymbol{x}) + V(\boldsymbol{x})),$$

$$\Downarrow$$

$$\boldsymbol{x} = \boldsymbol{x} - \frac{\boldsymbol{x}}{V(\boldsymbol{x})} \cdot \nabla f(\boldsymbol{x}) = \boldsymbol{x} - D_{\boldsymbol{x}} \nabla f(\boldsymbol{x}), \quad D_{\boldsymbol{x}} = \text{diag}\left(\frac{x_1}{V_1(\boldsymbol{x})}, \ldots, \frac{x_N}{V_N(\boldsymbol{x})}\right)$$

Iterative methods for $\boldsymbol{x} \cdot \nabla f(\boldsymbol{x}) = \mathbf{0}$ based on scaled gradient direction:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - D_k \nabla f(\boldsymbol{x}^{(k)}), \quad D_k = \text{diag}\left(\frac{x_1^{(k)}}{V_1(\boldsymbol{x}^{(k)})}, \ldots, \frac{x_N^{(k)}}{V_N(\boldsymbol{x}^{(k)})}\right)$$

Variable metric updating: the choice of the matrix D_k

The same suggestion arises from a Majorization-Minimiz. (MM) framework [Yang-Oja, *IEEE Trans. Neural Net.*(2011)]

- Consider discrepancy funct. $\mathcal{D}(H\boldsymbol{x}, \boldsymbol{g})$, $H_{i,j} \geq 0$, $x_i > 0$ written as

$$\mathcal{D}(H\boldsymbol{x}, \boldsymbol{g}) = \sum_{d=1}^{p} \sum_{i=1}^{n} \alpha_{d,i} h((Hx)_i, \zeta_d), \quad h(\sigma, t) = \begin{cases} \frac{\sigma^t - 1}{t} & \text{if } t \neq 0 \\ \log(\sigma) & \text{if } t = 0 \end{cases}$$

- A surrogate function $G(\boldsymbol{x}, \bar{\boldsymbol{x}})$ of $\mathcal{D}(H\boldsymbol{x}, \boldsymbol{g})$ at $\bar{\boldsymbol{x}}$ up to an additive constant can be defined in terms of the splitting
  $\nabla\mathcal{D}(H\bar{\boldsymbol{x}}, \boldsymbol{g}) = V(\bar{\boldsymbol{x}}) - U(\bar{\boldsymbol{x}})$, $\quad V(\bar{\boldsymbol{x}}) > 0$, $\quad U(\bar{\boldsymbol{x}}) \geq 0$

$$G(\boldsymbol{x}, \bar{\boldsymbol{x}}) = \sum_{j=1}^{n} \bar{x}_j (V(\bar{\boldsymbol{x}}))_j h\left(\frac{x_j}{\bar{x}_j}, \zeta_{\max}\right) - \bar{x}_j (U(\bar{\boldsymbol{x}}))_j h\left(\frac{x_j}{\bar{x}_j}, \zeta_{\min}\right)$$

  where

$$\zeta_{\max} = \max_{d \in \{1,\dots,p\}} \zeta_d, \qquad \zeta_{\min} = \min_{d \in \{1,\dots,p\}} \zeta_d$$

Variable metric updating: the choice of the matrix D_k

- Since

$$\frac{\partial}{\partial x_j} G(\boldsymbol{x}, \bar{\boldsymbol{x}}) = (V(\bar{\boldsymbol{x}}))_j \left(\frac{x_j}{\bar{x}_j}\right)^{\zeta_{\max}-1} - (U(\bar{\boldsymbol{x}}))_j \left(\frac{x_j}{\bar{x}_j}\right)^{\zeta_{\min}-1}$$

the corresponding MM method (based on $\nabla G(\boldsymbol{x}, \boldsymbol{x}^{(k)}) = 0$) leads to

$$\boldsymbol{x}^{(k+1)} = \operatorname*{argmin}_{\boldsymbol{x} \geq 0} \; G(\boldsymbol{x}, \boldsymbol{x}^{(k)}) = \boldsymbol{x}^{(k)} \left(\frac{U(\boldsymbol{x}^{(k)})}{V(\boldsymbol{x}^{(k)})}\right)^{\frac{1}{\zeta_{\max}-\zeta_{\min}}}$$

- In the special case of Least-Squares or Kullback-Leibler divergence, $(\zeta_{\max} - \zeta_{\min}) = 1$ and

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} \left(\frac{U(\boldsymbol{x}^{(k)})}{V(\boldsymbol{x}^{(k)})}\right) = \boldsymbol{x}^{(k)} - \frac{\boldsymbol{x}^{(k)}}{V(\boldsymbol{x}^{(k)})} \nabla \mathcal{D}(H\boldsymbol{x}^{(k)}, \boldsymbol{g})$$

Thus, the special scaled gradient step is a descent step for $\mathcal{D}(H\boldsymbol{x}^{(k)}, \boldsymbol{g})$.

Variable metric updating: the choice of the matrix D_k

Popular algorithms for imaging problems are based on this special scaling

- Iterative Space Reconstruction Algorithm (ISRA)

$$\min_{\boldsymbol{x}\geq\mathbf{0}} \mathcal{D}(H\boldsymbol{x},\boldsymbol{g}) \equiv \frac{1}{2}\|H\boldsymbol{x}+bg-\boldsymbol{g}\|^2$$

$$\boldsymbol{x}^{(k+1)}=\boldsymbol{x}^{(k)}\frac{H^T\boldsymbol{g}}{H^T(H\boldsymbol{x}^{(k)}+bg)} = \boldsymbol{x}^{(k)}-\frac{\boldsymbol{x}^{(k)}}{H^T(H\boldsymbol{x}^{(k)}+bg)}\nabla\mathcal{D}(H\boldsymbol{x}^{(k)},\boldsymbol{g}),\ \ \boldsymbol{x}^{(0)}>0$$

- Expectation Maximization (EM) or Richardson-Lucy (RL) algorithm

$$\min_{\boldsymbol{x}\geq\mathbf{0}} \mathcal{D}(H\boldsymbol{x},\boldsymbol{g}) \equiv \sum_{i=1}^{n}\left(g_i\log\frac{g_i}{(H\boldsymbol{x}+bg)_i}+(H\boldsymbol{x}+bg)_i-g_i\right)$$

$$\boldsymbol{x}^{(k+1)}=\frac{\boldsymbol{x}^{(k)}}{H^T\mathbf{1}}H^T\frac{\boldsymbol{g}}{H\boldsymbol{x}^{(k)}+bg}=\boldsymbol{x}^{(k)}-\frac{\boldsymbol{x}^{(k)}}{H^T\mathbf{1}}\nabla\mathcal{D}(H\boldsymbol{x}^{(k)},\boldsymbol{g}) \qquad \boldsymbol{x}^{(0)}>0$$

Variable metric updating: the choice of the matrix D_k

The split gradient idea within scaled gradient projection schemes:

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} + \vartheta_k \left(P_{\Omega, D_k^{-1}}(\boldsymbol{x}^{(k)} - \alpha_k D_k \nabla f(\boldsymbol{x}^{(k)})) - \boldsymbol{x}^{(k)} \right)$$

$$D_i^{(k)} = \min \left\{ \mu_k, \max \left\{ \frac{1}{\mu_k}, \frac{x_i^{(k)}}{V_i(\boldsymbol{x}^{(k)})} \right\} \right\}, \quad V_i(\boldsymbol{x}^{(k)}) > 0, \quad i = 1, \dots, N,$$

- similar idea used in [Hager-Mair-Zhang, *Math. Program.* (2009)]:

$$D_i^{(k)} = \frac{\alpha_k x_i^{(k)}}{x_i^{(k)} + \alpha_k \left(\nabla f(\boldsymbol{x}^{(k)}) \right)_i^+}, \quad i = 1, \dots, N, \quad (t)^+ = \max\{0, t\}$$

- works for more general constraints:
  [Hager-Zhang, *COAP* (2014); Bonettini et al. *SIAM J. Sci. Comp.* 2015]

The steplengths in Scaled Gradient Methods

Consider the scaled gradient method: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k D_k \boldsymbol{g}^{(k)}$

Recall the quadratic case: $\min f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T A \boldsymbol{x} - b^T \boldsymbol{x}$

- consider the problem $\tilde{f}(\boldsymbol{y}) = \frac{1}{2}\boldsymbol{y}^T D^{\frac{1}{2}} A D^{\frac{1}{2}} \boldsymbol{y} - b^T D^{\frac{1}{2}} \boldsymbol{y}$ and

$$\boldsymbol{y}^{(k+1)} = \boldsymbol{y}^{(k)} - \alpha_k \tilde{\boldsymbol{g}}^{(k)}, \qquad \tilde{\boldsymbol{g}}^{(k)} = \nabla \tilde{f}(\boldsymbol{y}^{(k)})$$

- Let $\boldsymbol{y}^{(k)} = D^{-\frac{1}{2}} \boldsymbol{x}^{(k)}$; we have $\tilde{\boldsymbol{g}}^{(k)} = D^{\frac{1}{2}} \boldsymbol{g}^{(k)}$ and

$$\boldsymbol{y}^{(k+1)} = D^{-\frac{1}{2}}(\boldsymbol{x}^{(k)} - \alpha_k D \boldsymbol{g}^{(k)}) = D^{-\frac{1}{2}} \boldsymbol{x}^{(k+1)}$$

- gradient step on $\boldsymbol{y}^{(k)}$ $\leftrightarrow$ scaled gradient step on $\boldsymbol{x}^{(k)}$

➤ Exploit the BB rules defined for the preconditioned problems by using

$$\boldsymbol{u}^{(k-1)} = D^{-\frac{1}{2}} \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right), \qquad \boldsymbol{v}^{(k-1)} = D^{\frac{1}{2}} (\boldsymbol{g}^{(k)} - \boldsymbol{g}^{(k-1)})$$

The steplengths in Scaled Gradient Methods

Consider the scaled gradient method: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k D_k \boldsymbol{g}^{(k)}$

The BB rules with scaling:

Let
$$\boldsymbol{s}^{(k-1)} = \left(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)}\right), \qquad \boldsymbol{z}^{(k-1)} = \left(\boldsymbol{g}^{(k)} - \boldsymbol{g}^{(k-1)}\right),$$
$$\boldsymbol{u}^{(k-1)} = D^{-\frac{1}{2}} \boldsymbol{s}^{(k-1)}, \qquad \boldsymbol{v}^{(k-1)} = D^{\frac{1}{2}} \boldsymbol{z}^{(k-1)},$$

define
$$\alpha_k{}^{\mathrm{BB1}} = \frac{\boldsymbol{u}^{(k-1)^T} \boldsymbol{u}^{(k-1)}}{\boldsymbol{u}^{(k-1)^T} \boldsymbol{v}^{(k-1)}} = \frac{\boldsymbol{s}^{(k-1)^T} D_k^{-1} \boldsymbol{s}^{(k-1)}}{\boldsymbol{s}^{(k-1)^T} \boldsymbol{z}^{(k-1)}}$$

$$\alpha_k{}^{\mathrm{BB2}} = \frac{\boldsymbol{u}^{(k-1)^T} \boldsymbol{v}^{(k-1)}}{\boldsymbol{v}^{(k-1)^T} \boldsymbol{v}^{(k-1)}} = \frac{\boldsymbol{s}^{(k-1)^T} \boldsymbol{z}^{(k-1)}}{\boldsymbol{z}^{(k-1)^T} D_k \boldsymbol{z}^{(k-1)}}$$

The steplengths in Scaled Gradient Methods

Consider the scaled gradient method: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k D_k \boldsymbol{g}^{(k)}$

Another interpretation of the scaled BB rules

Force the matrix $(\alpha_k D_k)^{-1}$ to approximate the Hessian $\nabla^2 f(\boldsymbol{x}^{(k)})$

by imposing quasi-Newton properties in variable metric

$$\alpha_k^{\mathsf{BB1}} = \frac{\boldsymbol{s}^{(k-1)^T} D_k^{-1} \boldsymbol{s}^{(k-1)}}{\boldsymbol{s}^{(k-1)^T} \boldsymbol{z}^{(k-1)}} = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \, \|(\alpha_k D_k)^{-1} \boldsymbol{s}^{(k-1)} - \boldsymbol{z}^{(k-1)}\|_{D_k}$$

or

$$\alpha_k^{\mathsf{BB2}} = \frac{\boldsymbol{s}^{(k-1)^T} \boldsymbol{z}^{(k-1)}}{\boldsymbol{z}^{(k-1)^T} D_k \boldsymbol{z}^{(k-1)}} = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \, \|\boldsymbol{s}^{(k-1)} - (\alpha_k D_k) \boldsymbol{z}^{(k-1)}\|_{D_k^{-1}}$$

The steplengths in Scaled Gradient Projection Methods

On the basis of the previous remarks, in case of box-constrained problems, instead of the standard BB2 rule

$$\alpha_k^{\text{BB2}} = \frac{{\boldsymbol{s}^{(k-1)}}^T \boldsymbol{z}^{(k-1)}}{{\boldsymbol{z}^{(k-1)}}^T D_k \boldsymbol{z}^{(k-1)}}$$

try to exploit

$$\alpha_k^{\text{MBB2}} = \frac{{\boldsymbol{s}_{\mathcal{I}_{k-1}}^{(k-1)}}^T \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}{{\boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}^T (D_k)_{\mathcal{I}_{k-1},\mathcal{I}_{k-1}} \boldsymbol{z}_{\mathcal{I}_{k-1}}^{(k-1)}}$$

where

$$\mathcal{I}_{k-1} = \{1, ..., n\} \setminus \mathcal{J}_{k-1}$$

$$\mathcal{J}_{k-1} = \{i : \ (x_i^{(k-1)} = \ell_i \ \wedge \ g_i^{(k-1)} \geq 0) \ \vee \ (x_i^{(k-1)} = u_i \ \wedge \ g_i^{(k-1)} \leq 0)\}$$

[Crisci-Porta-Ruggiero-Zanni, (2019)]

Example: 3D image reconstruction from limited tomographic data

$$\min_{\boldsymbol{x}\geq \boldsymbol{0}} J(\boldsymbol{x}) + \beta J_R(\boldsymbol{x})$$

- Least-squares divergence

$$J(\boldsymbol{x}) = \frac{1}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|^2 \qquad \nabla J(\boldsymbol{x}^{(k)}) = A^T A\boldsymbol{x}^{(k)} - A^T\boldsymbol{b}$$

- Edge preserving regularizer

$$\begin{aligned} J_R(\boldsymbol{x}) &= \textstyle\sum_{j_x=1}^{N_x}\sum_{j_y=1}^{N_y}\sum_{j_z=1}^{N_z}\sqrt{\delta^2 x_{j_x,j_y,j_z} + \delta^2} \\ \delta^2 x_{j_x,j_y,j_z} &= (x_{j_x+1,j_y,j_z} - x_{j_x,j_y,j_z})^2 + (x_{j_x,j_y+1,j_z} - x_{j_x,j_y,j_z})^2 \\ &\quad +(x_{j_x,j_y,j_z+1} - x_{j_x,j_y,j_z})^2 \\ \nabla J_R(\boldsymbol{x}^{(k)}) &= \boldsymbol{V}^{(\boldsymbol{k})} - \boldsymbol{U}^{(\boldsymbol{k})}, \qquad \boldsymbol{V}^{(\boldsymbol{k})} > 0, \qquad \boldsymbol{U}^{(\boldsymbol{k})} \geq 0 \end{aligned}$$

- Scaling matrix derived by the gradient splitting

$$D_k = \min\left(\mu_k, \max\left(\frac{1}{\mu_k}, \mathrm{diag}\left(\frac{\boldsymbol{x}^{(\boldsymbol{k})}}{A^T A\boldsymbol{x}^{(k)} + \beta\boldsymbol{V}^{(\boldsymbol{k})}}\right)\right)\right), \quad \mu_k = \sqrt{1 + \frac{M}{(k+1)^{2.1}}}$$

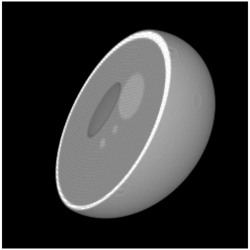
[Piccolomini-Coli-Morotti-Zanni, *Comput. Optim. Appl.* (2018)]

Simulations on the 3D Shepp Logan phantom

Test problem features

- exact volume $\boldsymbol{x}^*$: Shepp Logan phantom with $N_v = 61^3 \approx 226K$ voxels
- projection matrix $A \in M^{N_p \times N_v}$ with $N_\theta = 19 \rightarrow N_p = 61^2 \times N_\theta \approx 70K$

(http://www.imm.dtu.dk/~pcha/TVReg/)



Test platform: Test performed in Matlab 2016a on Intel core I7 6700

Compared methods

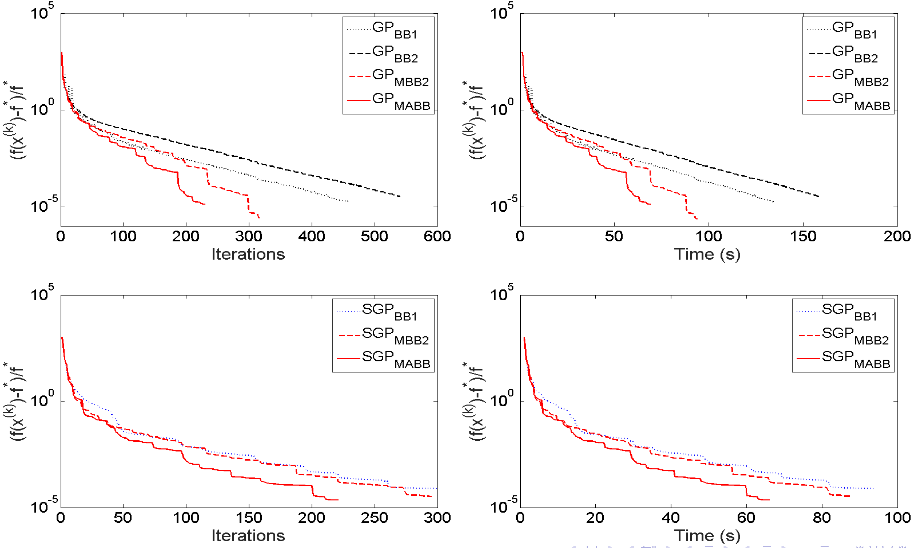
➤ GP equipped with BB1, BB2, MBB2, MABB steplengths

➤ SGP equipped with BB1, MBB2, MABB steplengths

➤ FISTA and Scaled FISTA algorithms

3D CT image reconstruction: the steplength behaviour

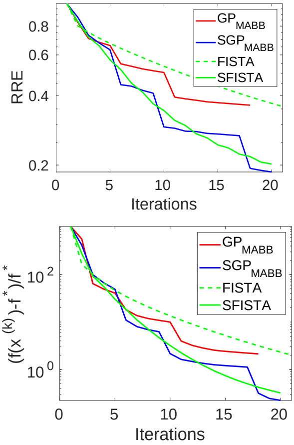
- **GP and SGP:** the new rules within alternated strategies are preferable



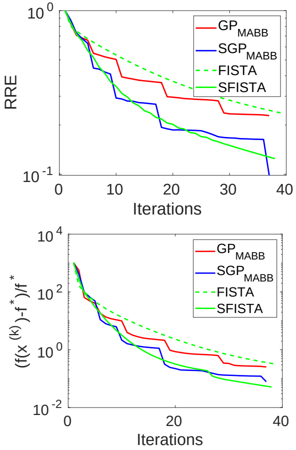
3D CT image reconstruction: the reconstruction error

- SGP and SFISTA preferable when a reconstruction is required in few seconds

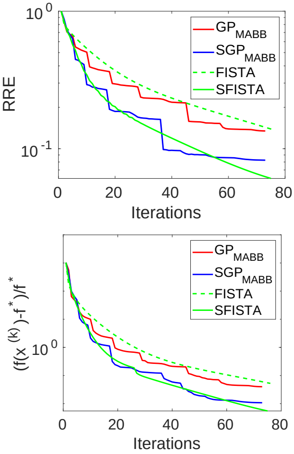
After 5 sec.



After 10 sec.



After 20 sec.



Scaled FISTA for $\min_{\boldsymbol{x}\in\Omega} f(\boldsymbol{x})$

f convex, ∇f Lips. continuous on Ω, dom$(f) \supseteq \Omega$, $X^* \neq \emptyset$

$$\begin{aligned} \boldsymbol{y}^{(k)} &= P_{\Omega, D_k^{-1}}\left(\boldsymbol{x}^{(k)} + \beta_k(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(k-1)})\right) \quad \text{new extrapolation step} \\ \boldsymbol{x}^{(k+1)} &= P_{\Omega, D_k^{-1}}\left(\boldsymbol{y}^{(k)} - \alpha_k D_k \nabla f(\boldsymbol{y}^{(k)})\right) \end{aligned}$$

Convergence analysis [Bonettini-Porta-Ruggiero, *SIAM J. Sci. Comput. 2016*]

- $\alpha_k \;\; \Big| \;\; f(\boldsymbol{x}^{(k+1)}) \leq f(\boldsymbol{y}^{(k)}) + \nabla f(\boldsymbol{y}^{(k)})^T(\boldsymbol{x}^{(k+1)} - \boldsymbol{y}^{(k)}) + \frac{1}{2\alpha_k}\left\|\boldsymbol{x}^{(k+1)} - \boldsymbol{y}^{(k)}\right\|^2_{D_k^{-1}}$
- $\beta_0 = 0, \qquad \beta_k = \frac{k-1}{k+a}, \qquad k = 1, \dots, \qquad a \geq 2$
- $\mu_k^2 = 1 + \gamma_k, \quad \gamma_k \geq 0, \quad \sum_{k=0}^{\infty} \gamma_k < \infty,$
- D_k s.p.d. with eigenvalues in $\left[\frac{1}{\mu_k}, \mu_k\right]$

$$\Downarrow$$

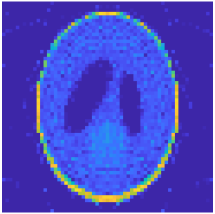
$$f(\boldsymbol{x}^{(k)}) - f^* \leq \frac{C}{(k-1+a)^2}, \qquad k = 1, 2, \dots$$

- $a > 2 \qquad \Rightarrow \qquad \lim_{k\to\infty} x^{(k)} = x^*$

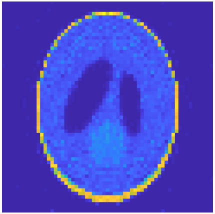
Reconstructions (LS + TV)

- **After 5 sec.**

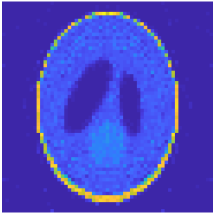
GP_MABB



SGP_MABB

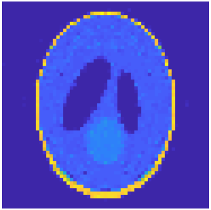


SFISTA



- **After 20 sec.**

GP_MABB



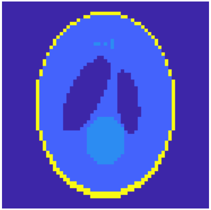
SGP_MABB

SFISTA

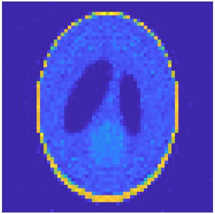
Comparison with the true image

- **After 5 sec.**

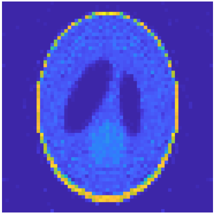
true image



SGP_MABB

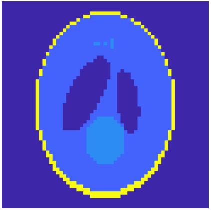


SFISTA



- **After 20 sec.**

true image



SGP_MABB

SFISTA

Conclusions

- Spectral properties of steplength rules in gradient methods:
  - useful for understanding the behaviour of standard rules
  - useful for designing improved selection rules
- Analysis of steplength rules in box-constrained problems:
  - suitable modification of state-of-the-art BB rules are suggested
- Analysis of steplength rules in scaled gradient projection methods:
  - Ad hoc BB rules exploiting spectral properties and scaling matrices

Work in progress

- More general constraints: e.g. $\Omega = \{\boldsymbol{l} \leq \boldsymbol{x} \leq \boldsymbol{u}, \quad \boldsymbol{a}^T\boldsymbol{x} = b\}$
  preliminary results confirm the importance of the spectral analysis
- Possible extension to stochastic gradient approaches

References and software:
www.oasis.unimore.it