SECONDA UNIVERSITÀ DEGLI STUDI DI NAPOLI
SCUOLA POLITECNICA E DELLE SCIENZE DI BASE
DMF – DIPARTIMENTO DI MATEMATICA E FISICA

On the regularization properties of some spectral gradient methods

Daniela di Serafino
Department of Mathematics and Physics, Second University of Naples
daniela.diserafino@unina2.it

contributions from R. De Asmundis, G. Landi, W.W. Hager, G. Toraldo, M. Viola, H. Zhang

PING (Inverse Problems in Geophysics) GNCS Project
Opening Workshop – Florence, April 6, 2016

Outline

Linear discrete inverse problem

$$\mathbf{b} = A\mathbf{x} + \mathbf{n}, \qquad A \in \mathbb{R}^{p \times n}, \ \mathbf{n} \in \mathbb{R}^p, \ \mathbf{x} \in \mathbb{R}^n, \ p \geq n$$

- A and $\mathbf{b}$ known data, A ill-conditioned, with singular values decaying to zero, and full rank
- $\mathbf{n}$ unknown, representing perturbations in the data
- $\mathbf{x}$ unknown, representing the object to be recovered

Linear discrete inverse problem

$$\mathbf{b} = A\mathbf{x} + \mathbf{n}, \qquad A \in \mathbb{R}^{p\times n},\ \mathbf{n} \in \mathbb{R}^p,\ \mathbf{x} \in \mathbb{R}^n,\ p \geq n$$

- A and $\mathbf{b}$ known data, A ill-conditioned, with singular values decaying to zero, and full rank
- $\mathbf{n}$ unknown, representing perturbations in the data
- $\mathbf{x}$ unknown, representing the object to be recovered

Reformulation as linear least squares problem: $\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}} \frac{1}{2}\|\mathbf{b} - A\mathbf{x}\|^2$

Exact least squares solution: $\mathbf{x}^\dagger = A^\dagger \mathbf{b} = \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i}\mathbf{v}_i = \mathbf{x}_{true} + \sum_{i=1}^{n} \frac{\mathbf{u}_i^T \mathbf{n}}{\sigma_i}\mathbf{v}_i$

$A = U\Sigma V^T$, $U = [\mathbf{u}_1, \ldots, \mathbf{u}_p] \in \mathbb{R}^{p\times p}$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_n] \in \mathbb{R}^{n\times n}$, $\Sigma = diag(\sigma_1, \ldots, \sigma_n) \in \mathbb{R}^{p\times n}$

useless, because the noise is amplified!

Filter factors and iterative regularization

Regularization by filter factors:

$$\mathbf{x}_{reg} = \sum_{i=1}^{n} \phi_i \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i$$

choose $\phi_i \approx 1$ to preserve the components of the solution corresponding to large σ_i's, and $\phi_i \approx 0$ to filter out the components corresponding to small σ_i's

Filter factors and iterative regularization

Regularization by filter factors:

$$\mathbf{x}_{reg} = \sum_{i=1}^{n} \phi_i \frac{\mathbf{u}_i^T \mathbf{b}}{\sigma_i} \mathbf{v}_i$$

choose $\phi_i \approx 1$ to preserve the components of the solution corresponding to large σ_i's, and $\phi_i \approx 0$ to filter out the components corresponding to small σ_i's

Iterative regularization methods, with a suitable early stop, can provide useful regularized solutions x_{reg}

Widely investigated classical iterative methods (see, e.g., [Hanke '95; Engl, Hanke & Neubauer '96; Nagy & Palmer '05]):

- Landweber and Steepest Descent (SD): very slow but "stable" convergence, rarely used in practice unless they are coupled with ad hoc preconditioners
- CG (CGLS, LSQR): fast in reducing the error, but too sensitive to stopping criteria (an early or late stopping may significantly deteriorate the solution)

Gradient methods for convex quadratic problems

General framework

choose $\mathbf{x}_0 \in \mathbb{R}^n$; $k = 0$
while (not stop_cond) **do**
  $\mathbf{g}_k = Q\mathbf{x} - \mathbf{c}$
  compute a suitable steplength α_k
  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k$
  $k = k + 1$
end while

$$\text{QP:} \ \underset{\mathbf{x} \in \mathbb{R}^n}{\text{minimize}} \, f(\mathbf{x}) \equiv \frac{1}{2} \mathbf{x}^T Q \mathbf{x} - \mathbf{c}^T \mathbf{x}$$

- old origins [Cauchy 1847; Akaike 1959; Forsythe 1968]
- long considered bad and ineffective because of slow convergence rate and oscillatory behaviour

Gradient methods for convex quadratic problems

General framework

choose $\mathbf{x}_0 \in \mathbb{R}^n$; $k = 0$
while (not stop_cond) **do**
  $\mathbf{g}_k = Q\mathbf{x} - \mathbf{c}$
  compute a suitable steplength α_k
  $\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{g}_k$
  $k = k + 1$
end while

QP: $\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x}) \equiv \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{c}^T\mathbf{x}$

- old origins [Cauchy 1847; Akaike 1959; Forsythe 1968]
- long considered bad and ineffective because of slow convergence rate and oscillatory behaviour

Starting from [Barzilai & Borwein '88], several more efficient gradient methods have been developed, with steplengths related to Hessian spectral properties

[Friedlander, Martínez, Molina & Raydan '99; Dai & Yuan '03, '05; Fletcher '05, '12; Dai, Hager, Schittowski & Zhang '06; Yuan '06, '08; Frassoldati, Zanni & Zanghirati '08; De Asmundis, **dS**, Riccio & Toraldo '13; De Asmundis, **dS**, Hager, Toraldo & Zhang '14; Gonzaga & Schneider '15]

$\Longrightarrow$ interest in the use of the new gradient methods as regularization methods

[Ascher, van den Doel, Huang & Svaiter '09; Cornelio, Porta, Prato & Zanni '13; De Asmundis, **dS** & Landi '16]

Analysis of gradient methods (for linear least squares)

$$\mathbf{g}_k = A^T(A\mathbf{x}_k - \mathbf{b}), \qquad k = 0, 1, 2, \ldots$$

Write $\mathbf{g}_k$ in terms of the SVD of A: if $\mathbf{g}_0 = \sum_{i=1}^{n} \mu_i^0 \mathbf{v}_i$, then

$$\mathbf{g}_k = \sum_{i=1}^{n} \mu_i^k \mathbf{v}_i, \qquad \mu_i^k = \mu_i^0 \prod_{j=0}^{k} (1 - \alpha_j \sigma_i^2)$$

Analysis of gradient methods (for linear least squares)

$$\mathbf{g}_k = A^T(A\mathbf{x}_k - \mathbf{b}), \qquad k = 0, 1, 2, \ldots$$

Write $\mathbf{g}_k$ in terms of the SVD of A: if $\mathbf{g}_0 = \sum_{i=1}^n \mu_i^0 \mathbf{v}_i$, then

$$\mathbf{g}_k = \sum_{i=1}^n \mu_i^k \mathbf{v}_i, \qquad \mu_i^k = \mu_i^0 \prod_{j=0}^k (1 - \alpha_j \sigma_i^2)$$

- if at the k-th iteration $\mu_i^k = 0$ for some i, then $\mu_i^l = 0$ for $l > k$
- $\mu_i^k = 0$ iff $\mu_i^0 = 0$ or $\alpha_j = 1/\sigma_i^2$ for some $j \leq k$
- $\alpha_k \approx \dfrac{1}{\sigma_i^2} \implies \begin{cases} |\mu_i^{k+1}| << |\mu_i^k| & \\ |\mu_r^{k+1}| < |\mu_r^k| & \text{if } r > i \\ |\mu_r^{k+1}| > |\mu_r^k| & \text{if } r < i \text{ and } \lambda_r > 2\sigma_i^2 \end{cases}$

Non-restrictive assumptions: $\sigma_1 > \sigma_2 > \cdots > \sigma_n, \quad \mu_1^0 \neq 0, \quad \mu_n^0 \neq 0$

A framework for building fast gradient methods

A new steplength selection rule

$$\alpha_k = \begin{cases} \alpha_k^{SD} & \text{if } \mathrm{mod}(k, h+m) < h \\ \bar{\alpha}_s & \text{otherwise, with } s = \max\{i \le k \ : \ \mathrm{mod}(i, h+m) = h\} \end{cases}$$

- $h \geq 2$
- α_k^{SD} classical (Cauchy) SD steplength
- $\bar{\alpha}_s$ "special" steplength with spectral properties

In other words: make h consecutive exact line searches and then compute a different steplength, to be kept constant and applied in m consecutive gradient iterations

SDA method

[De Asmundis, **dS**, Riccio & Toraldo '13]

Set $\bar{\alpha}_s = \widetilde{\alpha}_s$, where

$$\widetilde{\alpha}_s = \left(\frac{1}{\alpha_{s-1}^{SD}} + \frac{1}{\alpha_s^{SD}} \right)^{-1}$$

Let $\{\mathbf{x}_k\}$ be the sequence of iterates generated by the SD method applied to the least squares problem, starting from any point $\mathbf{x}_0$. Then

$$\lim_{k\to\infty} \widetilde{\alpha}_k = \frac{1}{\sigma_1^2 + \sigma_n^2}.$$

SDA method

[De Asmundis, **dS**, Riccio & Toraldo '13]

Set $\bar{\alpha}_s = \widetilde{\alpha}_s$, where

$$\widetilde{\alpha}_s = \left(\frac{1}{\alpha_{s-1}^{SD}} + \frac{1}{\alpha_s^{SD}} \right)^{-1}$$

Let $\{\mathbf{x}_k\}$ be the sequence of iterates generated by the SD method applied to the least squares problem, starting from any point $\mathbf{x}_0$. Then

$$\lim_{k\to\infty} \widetilde{\alpha}_k = \frac{1}{\sigma_1^2 + \sigma_n^2} .$$

SDA (SD with Alignment) combines

- the tendency of SD to choose its search direction in $span\{\mathbf{v}_1, \mathbf{v}_n\}$
- the tendency of the gradient method with $\alpha_k = 1/(\sigma_1^2 + \sigma_n^2)$ to align the search direction with $\mathbf{v}_n$,

R-linear conv., but significant improvement of practical convergence speed over SD

SDC method

[De Asmundis, **dS**, Hager, Toraldo & Zhang '14]

Set $\bar{\alpha}_s$ equal to the Yuan steplength [Yuan '06]

$$\alpha_s^Y = 2\left(\sqrt{\left(\frac{1}{\alpha_{s-1}^{SD}} - \frac{1}{\alpha_s^{SD}}\right)^2 + 4\frac{\|\mathbf{g}_s\|^2}{\left(\alpha_{s-1}^{SD}\|\mathbf{g}_{s-1}\|\right)^2}} + \frac{1}{\alpha_{s-1}^{SD}} + \frac{1}{\alpha_s^{SD}}\right)^{-1}$$

Let $\{\mathbf{x}_k\}$ be the sequence generated by the SD method applied to the least squares problem, starting from any point $\mathbf{x}_0$. Then

$$\lim_{k\to\infty} \alpha_k^Y = \frac{1}{\sigma_1^2}.$$

SDC method

[De Asmundis, **dS**, Hager, Toraldo & Zhang '14]

Set $\bar{\alpha}_s$ equal to the Yuan steplength [Yuan '06]

$$\alpha_s^Y = 2\left(\sqrt{\left(\frac{1}{\alpha_{s-1}^{SD}} - \frac{1}{\alpha_s^{SD}}\right)^2 + 4\frac{\|\mathbf{g}_s\|^2}{\left(\alpha_{s-1}^{SD}\|\mathbf{g}_{s-1}\|\right)^2}} + \frac{1}{\alpha_{s-1}^{SD}} + \frac{1}{\alpha_s^{SD}}\right)^{-1}$$

Let $\{\mathbf{x}_k\}$ be the sequence generated by the SD method applied to the least squares problem, starting from any point $\mathbf{x}_0$. Then

$$\lim_{k\to\infty} \alpha_k^Y = \frac{1}{\sigma_1^2}.$$

SDC (SD with Constant – Yuan – steps)

- uses a finite sequence of Cauchy steps in order to force the search in $span\{\mathbf{v}_1, \mathbf{v}_n\}$ and to get a suitable approximation of $1/\sigma_1^2$
- applies this approximation in multiple steps in order to drive toward zero the component of the gradient along $\mathbf{v}_1$

R-linear convergence, but significant improvement of practical convergence speed over SD

Some remarks

- If $\sigma_1 \gg \sigma_n$, then $1/(\sigma_1^2 + \sigma_n^2) \approx 1/\sigma_1^2$ and SDA fosters the elimination of the component of $\mathbf{g}_k$ corresponding to σ_1
- In the ideal case where the component of $\mathbf{g}_k$ along $\mathbf{v}_1$ is completely removed, the problem size decreases by 1, and SDA and SDC tend to drive toward zero the component of $\mathbf{g}_k$ along $\mathbf{v}_2$. The same reasoning applies to $\mathbf{v}_i$ for $i > 2$

Some remarks

- If $\sigma_1 \gg \sigma_n$, then $1/(\sigma_1^2 + \sigma_n^2) \approx 1/\sigma_1^2$ and SDA fosters the elimination of the component of $\mathbf{g}_k$ corresponding to σ_1
- In the ideal case where the component of $\mathbf{g}_k$ along $\mathbf{v}_1$ is completely removed, the problem size decreases by 1, and SDA and SDC tend to drive toward zero the component of $\mathbf{g}_k$ along $\mathbf{v}_2$. The same reasoning applies to $\mathbf{v}_i$ for $i > 2$
- In general SDA and SDC are non-monotone. However,
  - for small values of m, such as $m = 2, 3, 4$, SDA and SDC show monotonicity in practice if h is sufficiently large
  - when very low accuracy is required, as in the regularization of inverse ill-posed problems, h is not required to be "too large" ($h = 2, 3$ and $m = 2$ is a good combination)

Filter factors of SDA and SDC

[De Asmundis, **dS** & Landi '16]

Filter factors of gradient methods

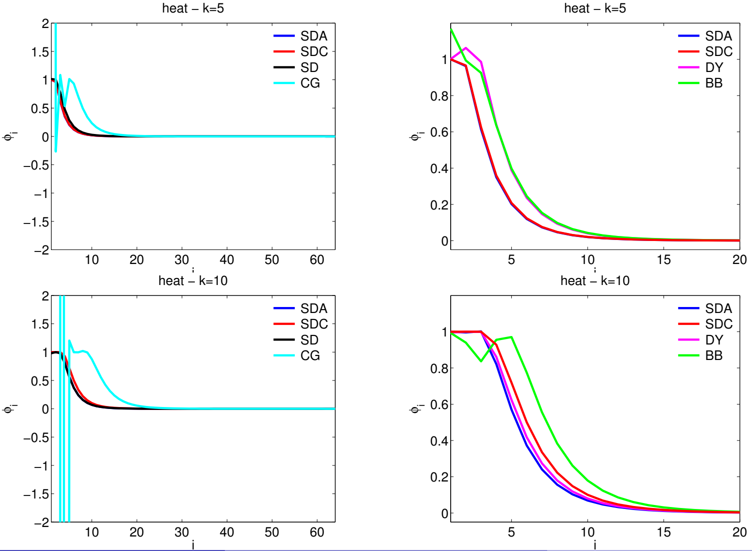
$$\mathbf{x}_{k+1} = \sum_{i=1}^{n} \underbrace{\left(1 - \prod_{r=0}^{k} \left(1 - \alpha_r \sigma_i^2\right)\right)}_{\phi_i^{k+1}} \frac{u_i^T b}{\sigma_i} \mathbf{v}_i, \qquad \mathbf{x}_0 = 0$$

- The better α_r approximates $1/\sigma_i^2$ for some r, the closer ϕ_i^{k+1} will be to 1; multiple values of α_r close to $1/\sigma_i^2$ push ϕ_i^{k+1} to quickly go toward 1
- $1/\alpha_r \gg \sigma_i^2 \quad \Rightarrow \quad \phi_i^k \approx 0$

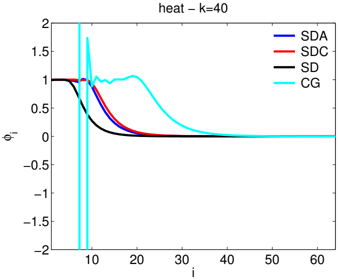
$\Longrightarrow$ the tendency of SDA and SDC to push toward zero the components of the gradient, according to the decreasing order of the singular values, translates into the approximation of the most significant components of the solution

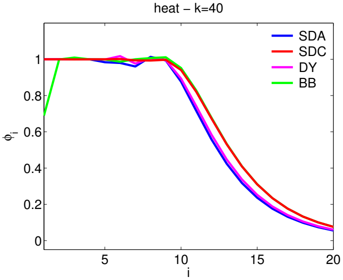
Comparison of filter factors

heat problem from Regularization Tools [Hansen '94], $size(A) = 64 \times 64$, $cond(A) \approx 10^{28}$, Gaussian noise, noise level 0.01, SDA/SDC with $h = 3$ and $m = 2$

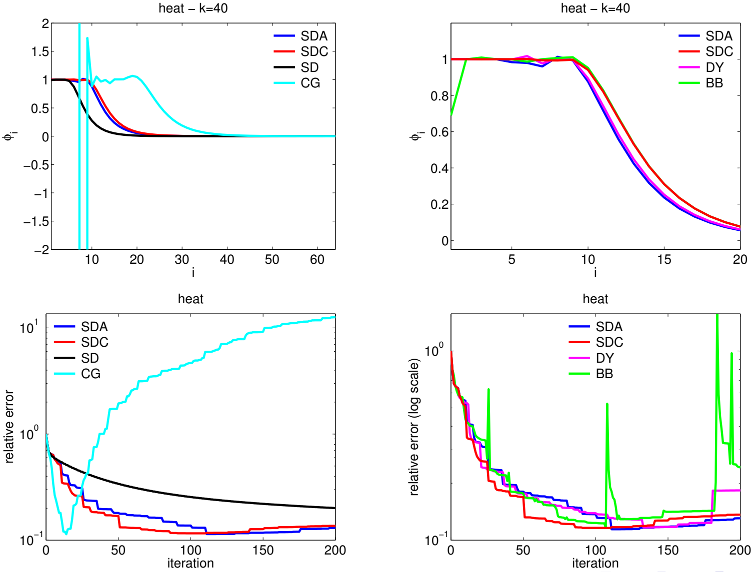


Comparison of filter factors (cont'd)





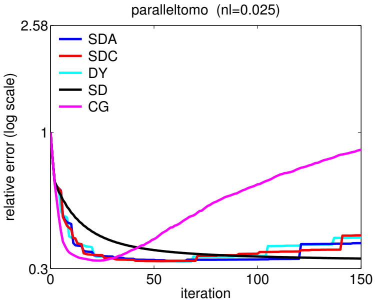
Comparison of filter factors (cont'd) and relative errors

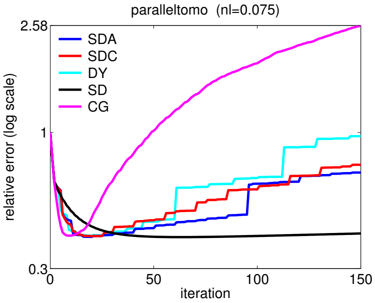


Experiments on image restoration problems: paralleltomo

parallel-beam tomography – AIR Tools [Hansen & Saxild-Hansen '12]

img size $= 50 \times 50$, 36 angles $(0^\circ - 179^\circ)$, 75 parallel rays, $cond(A) \approx 10^{15}$, $h = 3, m = 2$

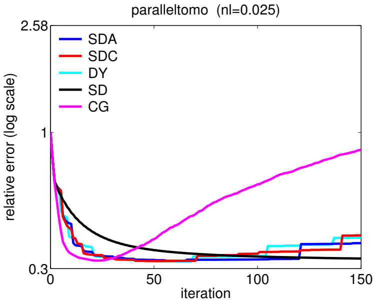


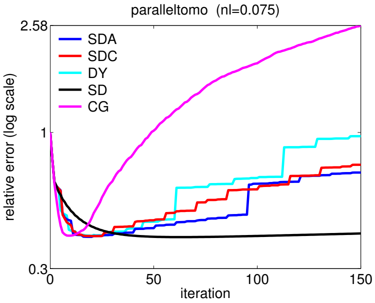


Experiments on image restoration problems: paralleltomo

parallel-beam tomography – AIR Tools [Hansen & Saxild-Hansen '12]

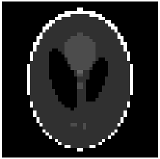
img size $= 50 \times 50$, 36 angles ($0^\circ - 179^\circ$), 75 parallel rays, $cond(A) \approx 10^{15}$, $h = 3, m = 2$



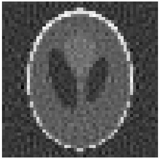


$nl = 0.025$

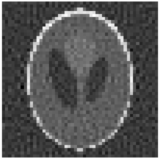
original



SDA



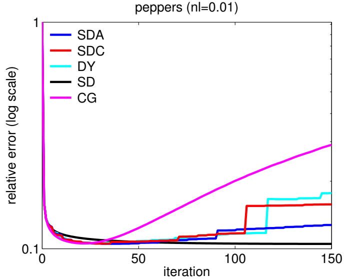
CG



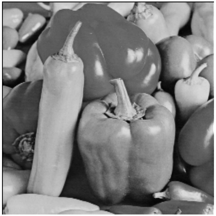
Experiments on image restoration problems: peppers

image deblurring problem, image size $= 256 \times 256$

Gaussian PSF, noise level $nl = 0.01$, $cond(A) \approx 10^{18}$, $h = 3$, $m = 2$



original



noisy & blurred

SDA

CG

Extending SDA and SDC to bound-constrained problems

$$\text{BCQP:} \quad \begin{array}{ll} \text{minimize} & f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{c}^T\mathbf{x} \\ \text{s. t.} & \mathbf{x} \in \Omega, \quad \Omega = \{\mathbf{x} : \mathbf{l} \le \mathbf{x} \le \mathbf{u}\} \end{array}$$

$Q \in \mathbb{R}^{n\times n}$ symmetric (positive definite), $\mathbf{l} \in \{\mathbb{R} \cup \{-\infty\}\}^n$, $\mathbf{u} \in \{\mathbb{R} \cup \{+\infty\}\}^n$

General framework

$\mathbf{x}_0 \in \mathbb{R}^n$; $k = 0$
while (not stop_cond) **do**
  $\mathbf{g}_k = Q\mathbf{x}_k - \mathbf{c}$
  compute a suitable steplength α_k
  $\mathbf{x}_{k+1} = P_\Omega(\mathbf{x}_k - \alpha_k \mathbf{g}_k)$
  $k = k + 1$
end while

Gradient Projection (GP) methods
[Goldstein, 1964; Levitin & Polyak, 1966; Calamai & Moré, 1987]

$$P_\Omega(x) = \text{argmin}\{\|x - z\| : z \in \Omega\}$$

Extending SDA and SDC to bound-constrained problems

$$\text{BCQP:}\quad \begin{array}{ll} \text{minimize} & f(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T Q\mathbf{x} - \mathbf{c}^T\mathbf{x} \\ \text{s. t.} & \mathbf{x} \in \Omega, \quad \Omega = \{\mathbf{x} : \mathbf{l} \le \mathbf{x} \le \mathbf{u}\} \end{array}$$

$Q \in \mathbb{R}^{n\times n}$ symmetric (positive definite), $\mathbf{l} \in \{\mathbb{R} \cup \{-\infty\}\}^n$, $\mathbf{u} \in \{\mathbb{R} \cup \{+\infty\}\}^n$

General framework

$\mathbf{x}_0 \in \mathbb{R}^n$; $k = 0$
while (not stop_cond) **do**
  $\mathbf{g}_k = Q\mathbf{x}_k - \mathbf{c}$
  compute a suitable steplength α_k
  $\mathbf{x}_{k+1} = P_\Omega(\mathbf{x}_k - \alpha_k \mathbf{g}_k)$
  $k = k + 1$
end while

Gradient Projection (GP) methods
[Goldstein, 1964; Levitin & Polyak, 1966; Calamai & Moré, 1987]

$$P_\Omega(x) = \operatorname{argmin}\{\|x - z\| : z \in \Omega\}$$

The spectral properties of SDA and SDC are not preserved!

Two-phase GP algorithm: basics

$\mathbf{x}, \mathbf{x}^* \in \Omega$

- active set at $\mathbf{x}$: $\mathcal{A}(\mathbf{x}) = \{i : x_i = l_i \text{ opp. } x_i = u_i\}$
- projected gradient at $\mathbf{x}$: $(\nabla_\Omega f(\mathbf{x}))_i = \begin{cases} \partial f_i(\mathbf{x}), & x_i \in (l_i, u_i) \\ \min\{\partial f_i(\mathbf{x}),\ 0\}, & x_i = l_i \\ \max\{\partial f_i(\mathbf{x}),\ 0\}, & x_i = u_i \end{cases}$
- binding set at $\mathbf{x}$: $\mathcal{B}(\mathbf{x}) = \{i : (x_i = l_i \text{ and } \partial f_i(\mathbf{x}) \geq 0) \text{ or } (x_i = u_i \text{ e } \partial f_i(\mathbf{x}) \leq 0)\}$
- $\mathbf{x}^*$ nondegenerate stationary point: $\partial f_i(\mathbf{x}^*) \neq 0 \ \ \forall i \in \mathcal{A}(\mathbf{x}^*)$

Identification of the active constraints at the solution [Calamai & Moré, 1987]:
if $\{\mathbf{x}_k\}$ converges to nondegenerate $\mathbf{x}^* \in \Omega$ and $\{\|\nabla_\Omega f(\mathbf{x}_k)\|\}$ converges to 0, then $\mathcal{A}(\mathbf{x}_k) = \mathcal{A}(\mathbf{x}^*)$ for all sufficiently large k

Two-phase GP algorithm: basics

$\mathbf{x}, \mathbf{x}^* \in \Omega$

- active set at $\mathbf{x}$: $\mathcal{A}(\mathbf{x}) = \{ i : x_i = l_i \text{ opp. } x_i = u_i \}$
- projected gradient at $\mathbf{x}$: $(\nabla_\Omega f(\mathbf{x}))_i = \begin{cases} \partial f_i(\mathbf{x}), & x_i \in (l_i, u_i) \\ \min\{\partial f_i(\mathbf{x}),\ 0\}, & x_i = l_i \\ \max\{\partial f_i(\mathbf{x}),\ 0\}, & x_i = u_i \end{cases}$
- binding set at $\mathbf{x}$: $\mathcal{B}(\mathbf{x}) = \{ i : (x_i = l_i \text{ and } \partial f_i(\mathbf{x}) \geq 0) \text{ or } (x_i = u_i \text{ e } \partial f_i(\mathbf{x}) \leq 0) \}$
- $\mathbf{x}^*$ nondegenerate stationary point: $\partial f_i(\mathbf{x}^*) \neq 0 \ \ \forall i \in \mathcal{A}(\mathbf{x}^*)$

Identification of the active constraints at the solution [Calamai & Moré, 1987]:
if $\{\mathbf{x}_k\}$ converges to nondegenerate $\mathbf{x}^* \in \Omega$ and $\{\|\nabla_\Omega f(\mathbf{x}_k)\|\}$ converges to 0, then $\mathcal{A}(\mathbf{x}_k) = \mathcal{A}(\mathbf{x}^*)$ for all sufficiently large k

Basic idea (as in [Moré & Toraldo, 1991]):

- use a GP method to select a "candidate" active set
- use SDA/SDC to explore the face of Ω identified by GP (unconstr. subprob.)
  $\Longrightarrow$ "preserve" the spectral properties of the new gradient methods

Two-phase GP algorithm

[dS, Toraldo, Viola, work in progress]

Sketch of the algorithm

$\mathbf{x}_0 \in \mathbb{R}^n$; $k = 0$

while (not stop_cond) **do**

- apply a GP method to BCQP:

  starting from $\mathbf{y}_0 = \mathbf{x}_k$, generate $\{\mathbf{y}_j\}$ until cond1 is satisfied
- $\bar{\mathbf{x}}_k = \mathbf{y}_{j_k}$, where $\mathbf{y}_{j_k} =$ last $\mathbf{y}_j$
- apply SDA/SDC to $\min\{f_k(\mathbf{d}) \equiv f(\bar{\mathbf{x}}_k + \mathbf{d}) : d_i = 0 \ \forall\, i \in \mathcal{A}(\bar{\mathbf{x}}_k)\}$:

  starting from $\mathbf{d}_0 = \mathbf{0}$, generate $\{\mathbf{d}_j\}$ until cond2 is satisfied
- $\mathbf{x}_{k+1} = P_\Omega(\bar{\mathbf{x}}_k + \alpha_k \mathbf{d}_{r_k})$, with $\mathbf{d}_{r_k}$= last $\mathbf{d}_k$ and α_k computed by a projected search
- if $\mathcal{A}(\mathbf{x}_{k+1}) = \mathcal{B}(\mathbf{x}_{k+1})$, then continue with SDA/SDC

end while

cond1: $\mathcal{A}(\mathbf{y}_j) = \mathcal{A}(\mathbf{y}_{j-1})$ or $f(\mathbf{y}_{j-1}) - f(\mathbf{y}_j) \le \eta_2 \max\{f(\mathbf{y}_{l-1}) - f(\mathbf{y}_l),\ 1 \le l < j\}$

cond2: $f_k(\mathbf{d}_{j-1}) - f_k(\mathbf{d}_j) \le \eta_1 \max\{f_k(\mathbf{d}_{l-1}) - f_k(\mathbf{d}_l),\ 1 \le l < j\}$

Two-phase GP algorithm: convergence

Projected search along $-\nabla f(x_k)$ e d_k:

generate a sequence of "trial" steplengths such that

- $\alpha_k^{(l+1)} \in \left[\gamma_1 \alpha_k^{(l)}, \gamma_2 \alpha_k^{(l)}\right], \quad 0 < \gamma_1 < \gamma_2 < 1, \quad \alpha_k^{(0)} > 0$
- $\alpha_k = \alpha_k^{(r)}$ satisfying an Armijo-like condition for f

Convergence:

if Q is spd and $\mathbf{x}^*$ is the solution of BCQP, then any sequence $\{\mathbf{x}_k\}$ generated by the two-phase GP algorithm is such that

- either $x_k = x^*$ after a finite number of iterations
- or $x_k \to x^*$

Some numerical experiments (Matlab)

random Q with $n = 10^4$ and varying $\kappa(Q)$; bounds: $-\beta \le x_i \le \beta$, $\beta = 1, 5, 9$

$\mathbf{x}_0 = 0$; stop crit. $\|\nabla_\Omega f(x_k)\| \le 10^{-5}\|\nabla f(x_0)\|$; SDC with $h = m = 4$

$\kappa(Q)$	η_1	η_2	# MAT-VET PRODUCTS					
			10% active constr.		50% active constr.		90% active constr.	
			GPSDC	GPCG	GPSDC	GPCG	GPSDC	GPCG
10^3	0.10	0.10	433	289	400	306	304	135
10^3	0.25	0.10	472	592	572	351	393	130
10^3	0.10	0.25	406	260	345	323	165	130
10^3	0.25	0.25	560	336	377	286	198	130
10^6	0.10	0.10	3781	4002	2453	5922	451	1160
10^6	0.25	0.10	3555	4708	3652	3322	548	477
10^6	0.10	0.25	3635	4612	3004	8127	561	1092
10^6	0.25	0.25	3815	4687	2836	4740	565	538
10^9	0.10	0.10	3470	3445	5780	12521	528	869
10^9	0.25	0.10	2697	3949	6730	7593	472	570
10^9	0.10	0.25	3524	3121	5484	15629	559	783
10^9	0.25	0.25	3267	3008	5109	7378	605	635

- GPSDC competitive with GPCG, especially on the most difficult problems
- GPSDC less sensitive to η_1 ed η_2 than GPCG

Exploiting gradient methods for QP/BCQP in nonlinear inverse problems – 1

$(\mathbf{h}, \mathbf{b})$ = data, $\mathbf{x}$ = parameters to be estimated,
$\mathbf{m}(\mathbf{x}, \mathbf{h})$ = model function, $\mathbf{r}(\mathbf{x}) = \mathbf{b} - \mathbf{m}(\mathbf{x}, \mathbf{h})$ = error in model prediction

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x}), \qquad f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|_2^2 = \sum_{i=1}^{m} r_i^2(\mathbf{x})$$

Exploiting gradient methods for QP/BCQP in nonlinear inverse problems – 1

$(\mathbf{h}, \mathbf{b})$ = data, $\mathbf{x}$ = parameters to be estimated,

$\mathbf{m}(\mathbf{x}, \mathbf{h})$ = model function, $\mathbf{r}(\mathbf{x}) = \mathbf{b} - \mathbf{m}(\mathbf{x}, \mathbf{h})$ = error in model prediction

$$\underset{\mathbf{x}\in\mathbb{R}^n}{\text{minimize}}\, f(\mathbf{x}), \qquad f(\mathbf{x}) = \frac{1}{2}\|\mathbf{r}(\mathbf{x})\|_2^2 = \sum_{i=1}^{m} r_i^2(\mathbf{x})$$

Regularized Gauss-Newton method

$\mathbf{x}_0 \in \mathbb{R}^n$; $k = 0$
while (not stop_cond) **do**
  compute J_k Jacobian of $\mathbf{r}$ at $\mathbf{x}_k$
  compute $\mathbf{d}_k$ regularized solution of $\underset{\mathbf{d}\in\mathbb{R}^n}{\text{minimize}}\, \|J_k\mathbf{d} + \mathbf{r}(\mathbf{x}_k)\|_2^2$
  compute α_k by a suitable line search
  $\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{d}_k$
  $k = k + 1$
end while

Exploiting gradient methods for QP/BCQP in nonlinear inverse problems – 1 (cont'd)

Compute a regularized solution of

$$\underset{\mathbf{d}\in\mathbb{R}^n}{\text{minimize}} \, \|J_k \mathbf{d} + \mathbf{r}(\mathbf{x}_k)\|_2^2$$

- [Deidda, Fenu & Rodriguez, 2014]:
  compute a TSVD solution (or a TGSVD one, in order to introduce a regularization matrix)
- Possible alternative:
  use SDA/SDC to compute a regularized solution
  - Less sensitive to the estimate of the noise norm
  - Easy to use in a matrix-free regime
  - Effective? Efficient?

Exploiting gradient methods for QP/BCQP in nonlinear inverse problems – 2

$$\begin{array}{ll} \text{minimize} & f(\mathbf{u}) \equiv f^{fit}(\mathbf{u}) + \lambda f^{reg}(\mathbf{u}) \\ \text{s. t.} & \mathbf{u} \geq \mathbf{0} \end{array}$$

$f^{fit}(\mathbf{u}) = KL(A\mathbf{u}, \mathbf{b})$ Kullback-Leibler divergence

$f^{reg}(\mathbf{u}) = TV(\mathbf{u})$ or $f^{reg}(\mathbf{u}) = \|W\mathbf{u}\|_1$ (frame-based regularization)

Solve the problem by combining

- Iteratively Reweighted Norm approach
  [Wolke & Schwetlick, 1988; Rodriguex & Wohlberg, 2009]
- Weighted Least Squares approximation of KL fidelity term
  [Shen, Yin, Zhang, 2015]

[Work in progress (just started), with G. Landi]

Exploiting gradient methods for QP/BCQP in nonlinear inverse problems – 2 (cont'd)

Algorithm (sketch)

$\mathbf{u}_0 \in \mathbb{R}^n$; $k = 0$

while (not stop_cond) **do**

1. compute $f_k^{fit}(\mathbf{u})$ quadratic approx of $f^{fit}(\mathbf{u})$ (using $\mathbf{u}_k$)
2. compute $f_k^{reg}(\mathbf{u})$ quadratic approx of $f^{reg}(\mathbf{u})$ (using $\mathbf{u}_k$)
3. compute $\mathbf{u}_{k+1} \approx \text{argmin}_{\mathbf{u} \geq \mathbf{0}} f_k^{fit}(\mathbf{u}) + \lambda f_k^{reg}(\mathbf{u})$
4. $k = k + 1$

end while

1. Weighted Least Squares approximation
2. Iteratively Reweighted Norm approach
3. Two-phase GP algorithm

CAN WE EFFICIENTLY EXPLOIT

SPECTRAL GRADIENT METHODS

IN THE PING PROJECT?