Two properties of an augmented Lagrangian method for solving strictly convex quadratic programs

J. Ch. GILBERT
INRIA-Rocquencourt
Jean-Charles.Gilbert@inria.fr

Joint work with

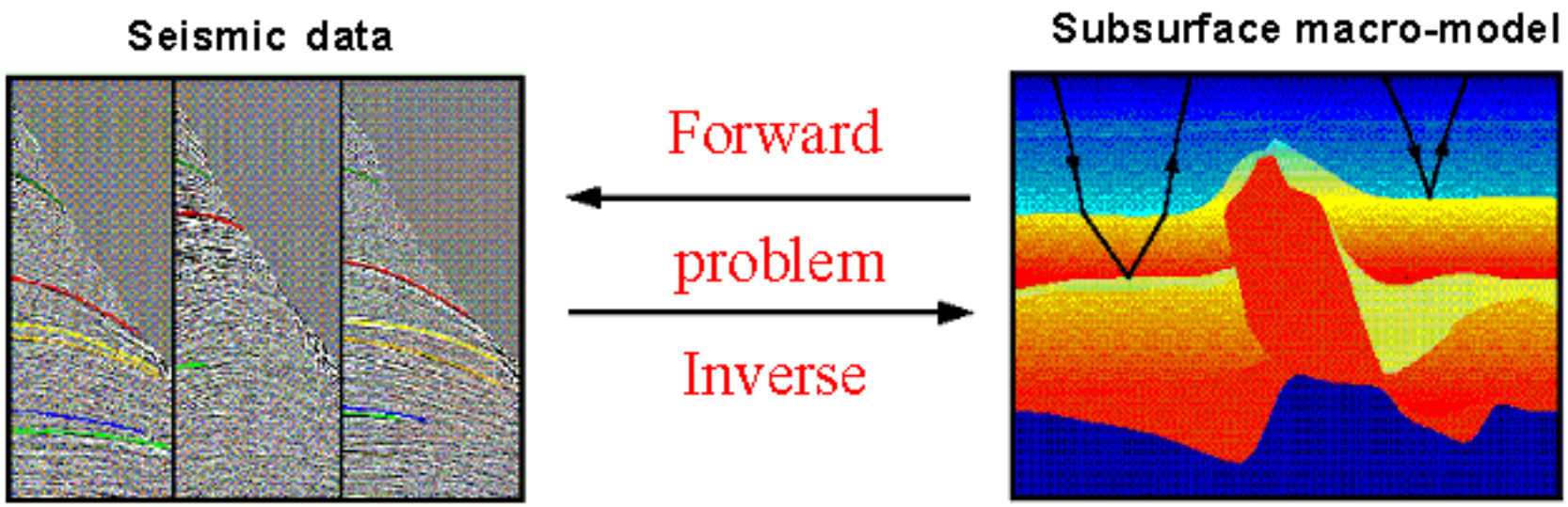
F. DELBOS, R. MASSON, D. SINOQUET, Institut Français du Pétrole (IFP)
R. GLOWINSKI, University of Huston

A geophysical motivation

- Reconstruction of the geophysical structure of the substratum
  (interface positions and velocities).

- From seismic data
  (receivers measuring the wave traveltimes).

- Useful for the detection of petroleum reservoirs.


Seismic data
Subsurface macro-model
Forward
problem
Inverse

Mathematical formulation

- A least-squares problem:

$$\min_m \ \frac{1}{2}\|T(m) - T^{obs}\|_2^2,$$

where

  - m is the model of the subsurface structure (typically 10^5 variables),
  - $T(m)$ is the computed traveltimes,
  - T^{obs} is the observed traveltimes (typically 10^6 observations).

- A need of regularization:

$$\min_m \ \frac{1}{2}\|T(m) - T^{obs}\|_2^2 + \frac{\varepsilon}{2} m^\top R m,$$

where

  - $R \succcurlyeq 0$ is such that the Gauss-Newton Hessian approximation satisfies

$$J^\top J + \varepsilon R \succ 0.$$

  - Here $J := T'(m)$.

Introduction of constraints

- Only linear constraints:

$$\begin{cases} \min_m \ \frac{1}{2}\|T(m) - T^{obs}\|_2^2 \\ c_E(m) = 0 \\ l \leq c_I(m) \leq u. \end{cases}$$

- Solution technique $\longrightarrow$ SQP, which requires the solution to QP's

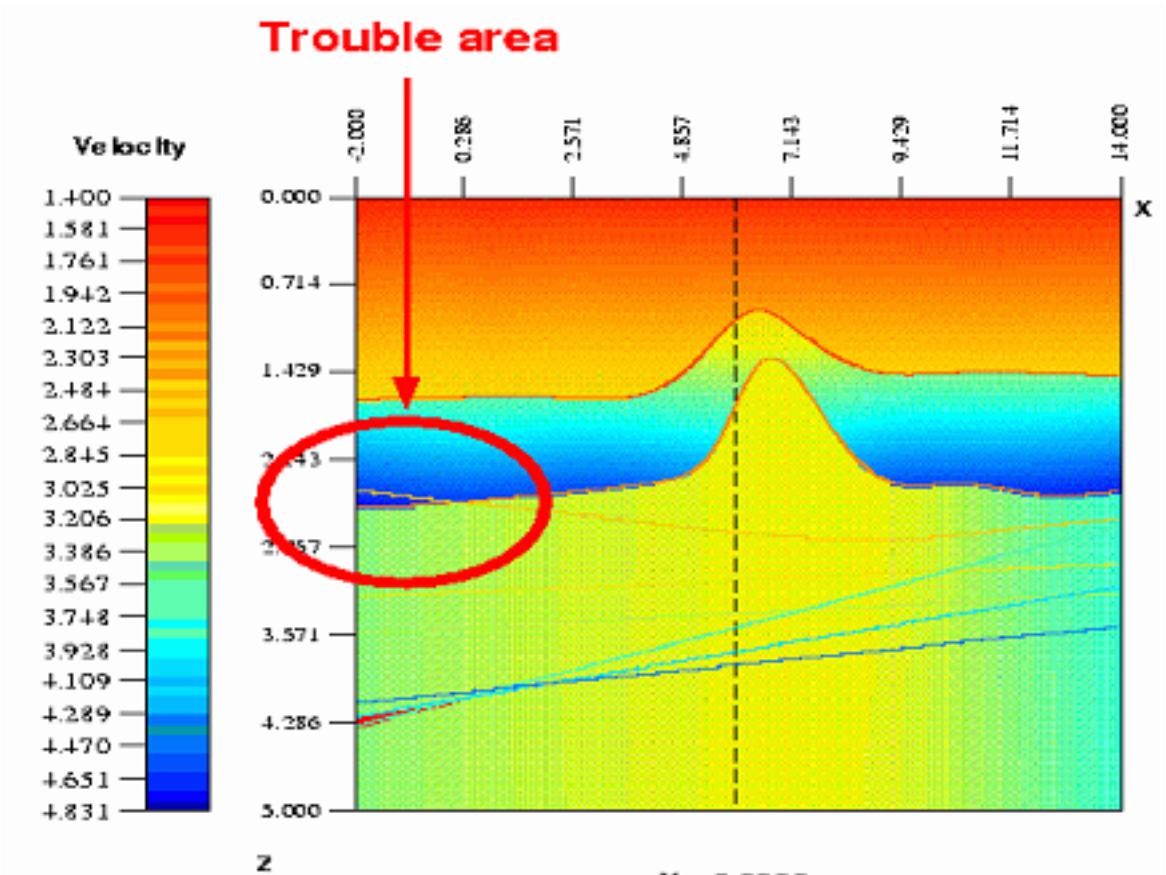
$$\begin{cases} \inf_x \ \frac{1}{2}x^\top Q x + q^\top x \\ C_E x = b \\ l \leq C_I x \leq u, \end{cases}$$

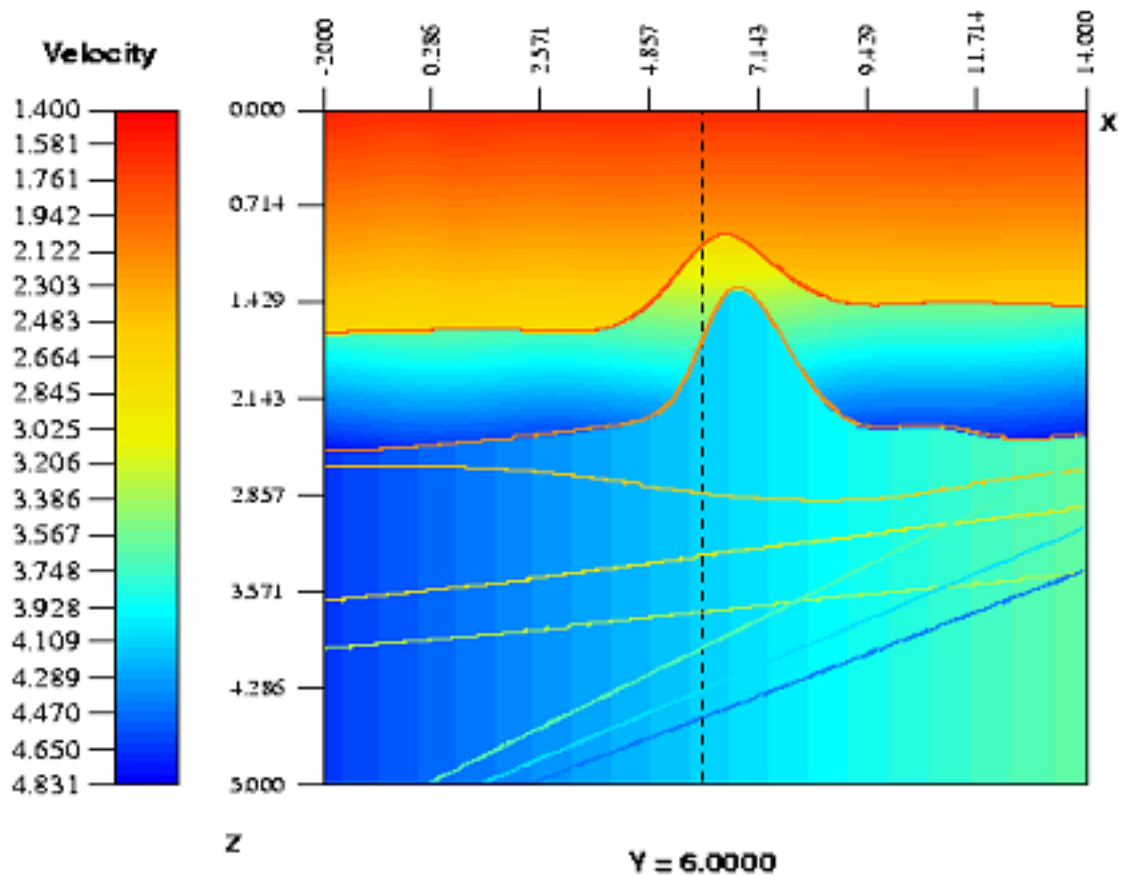
  where

$$Q := J^\top J + \varepsilon R \succ 0.$$

- For "historical" reasons,
  we solve these QP's using the augmented Lagrangian (AL) algorithm:

It is very efficient !


Trouble area
Velocity
1.400
1.581
1.761
1.942
2.122
2.303
2.484
2.664
2.845
3.025
3.206
3.386
3.567
3.748
3.928
4.109
4.289
4.470
4.651
4.831
-2.000
0.286
2.571
4.857
7.143
9.429
11.714
14.000
x
0.000
0.714
1.429
3.571
4.286
5.000
z


Velocity
1.400
1.581
1.761
1.942
2.122
2.303
2.483
2.664
2.845
3.025
3.206
3.386
3.567
3.748
3.928
4.109
4.289
4.470
4.650
4.831
-2.000
0.286
2.571
4.857
7.143
9.429
11.714
14.000
X
0.000
0.714
1.429
2.143
2.857
3.571
4.286
5.000
Z
Y = 6.0000

Solving QP's with an augmented Lagrangian (AL) approach

- **The problem to solve:**

$$\text{(P)} \quad \begin{cases} \inf_x \ \frac{1}{2} x^\top Q x + q^\top x \\ l \leq Cx \leq u, \end{cases}$$

where

  - Q is $n \times n$ and positive definite,
  - $q \in \mathbb{R}^n$,
  - C is $m \times n$,
  - $l \leq Cx \leq u$ means $l_i \leq (Cx)_i \leq u_i$ for $i = 1, \ldots, m$,
  - it is assumed that $l < u$ (l_i and u_i can be infinite).

- **Comments:**

  - $Q \succ 0 \Longrightarrow$ (P) has a unique solution, say $\bar{x}$,
  - $Q \succ 0 \Longrightarrow$ RESTRICTIVE, but may be useful
    - in its own right,
    - for regularized least-squares problems + SQP (then $Q = J^\top J + \varepsilon R$),
    - with SQP + pd-quasi-Newton Hessian approximation,
  - AL approach is useful for LARGE problems, without matrix factorizations.

- **Problem reformulation:**

$$\text{(P)} \quad \begin{cases} \inf_x \ \frac{1}{2}x^\top Q x + q^\top x \\ l \le Cx \le u, \end{cases} \quad \longrightarrow \quad \begin{cases} \inf_{(x,y)} \ \frac{1}{2}x^\top Q x + q^\top x \\ y = Cx \qquad \leftarrow \text{ dualized constraint} \\ l \le y \le u, \end{cases}$$

- **Algorithm AL:** Hestenes (1969), Powell (1969), Rockafellar (1971-1973), . . .
  - given $\lambda_0 \in \mathbb{R}^m$
  - repeat for $k = 0, 1, 2, \ldots$:
    - solve with a certain $r_k > 0$

$$\begin{cases} \inf_{(x,y)} \ \frac{1}{2}x^\top Q x + q^\top x + \lambda_k^\top (y - Cx) + \frac{r_k}{2}\|y - Cx\|^2 \\ l \le y \le u \end{cases} \tag{1}$$

      $\longrightarrow$ unique solution (x_k, y_k)
    - update $\lambda_{k+1} = \lambda_k + r_k(y_k - Cx_k)$
    - stop if $y_k \simeq Cx_k$.

- **Remarks:**
  - role of r_k
  - the Lagrange problem (1) is “easy” to solve (analytically in y or numerically in (x, y)).

Outline

Background: proximality and duality

- The proximal algorithm
- The classical dual function δ_0 and the regularized dual function δ_r
- AL method as a proximal algorithm on δ_0

Property 1: finite time identification of the active constraints at the solution $\bar{x}$ (new ?)

- Define $\mathrm{sign}(t) := -1$ if $t < 0$; $\mathrm{sign}(t) := 0$ if $t = 0$; $\mathrm{sign}(t) := +1$ if $t > 0$.
- If strict complementarity and $\lambda_k \to \bar{\lambda}$ (dual convergence),
  then $\mathrm{sign}(\lambda_k) = \mathrm{sign}(\bar{\lambda})$, for k large enough.

Property 2: GLOBAL linear convergence of $\lambda_k \to \bar{\lambda}$, a dual solution (new ?)

- If r is large enough,
  then $\forall\, \beta > 0$, $\exists\, \rho \in \,]0, 1[$, such that $\mathrm{dist}(\lambda_0, \mathcal{S}_D) \leq \beta$ and $k \geq 0$ imply

$$\|y_{k+1} - Cx_{k+1}\| \leq \rho \|y_k - Cx_k\|.$$

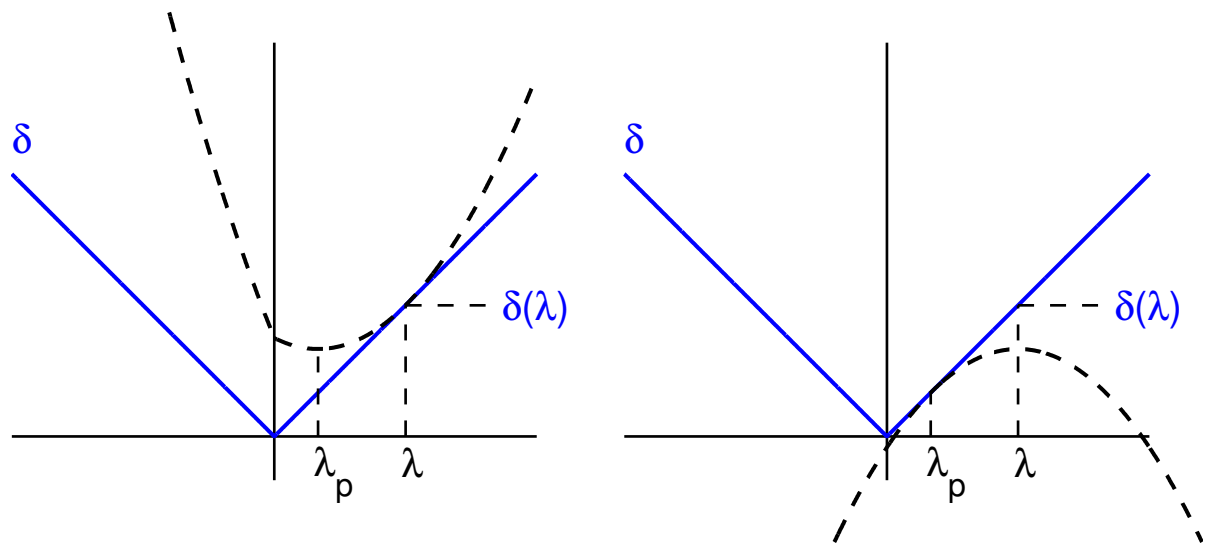
Proximality and duality

Proximality

- Let be given $\delta \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{R}^m)$, $r > 0$ and $\lambda \in \mathbb{R}^m$.
- Then the problem

$$\inf_{\mu \in \mathbb{R}^m} \left(\delta(\mu) + \frac{1}{2r} \|\mu - \lambda\|^2 \right)$$

  has a unique solution, λ_p say, called the proximal point of λ.
- Interpretation: for some $g_p \in \partial\delta(\lambda_p)$, there holds $\lambda_p = \lambda - r g_p$.



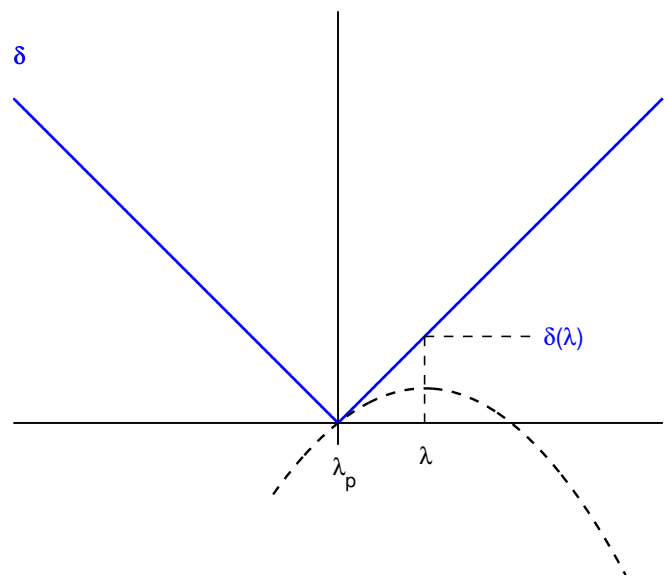
The proximal algorithm for minimizing δ

- **Proximal algorithm:** Bellman et al. (1966), Martinet (1970), Rockafellar (1976), . . .
  - current iterate $\lambda_k \in \mathbb{R}^m$
  - the next iterate λ_{k+1} is the proximal point $(\lambda_k)_p$ of λ_k for a certain $r_k > 0$.

  It is known that, if $\{r_k\}$ is bouded away from zero, then $\lambda_k \to \bar{\lambda}$ (some minimizor of δ).

- **A famous finite time convergence property:**

$$0 \in \operatorname{int} \partial\delta(\bar{\lambda}) \quad \text{(or } \delta'(\bar{\lambda};\mu) > 0, \forall \mu \neq 0\text{)} \quad \Longrightarrow \quad \text{finite termination (or } \lambda_p = \bar{\lambda} \text{ if } \lambda \text{ close to } \bar{\lambda}\text{)}$$

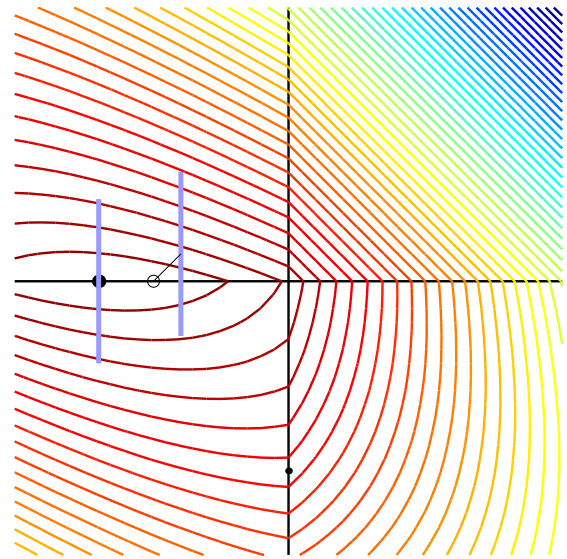


- **A convenient extension of this property:**

If · $\delta \in \mathrm{C}\overline{\mathrm{onv}}(\mathbb{R}^m)$, fixed $r > 0$,
· the proximal algorithm generates a sequence $\{\lambda_k\} \to \bar{\lambda}$,
· there exists a set $\mathcal{A} \subset \mathbb{R}^m$ such that

$$\Omega := \bigcup_{\lambda \in \mathcal{A}} (\lambda + r\,\partial\delta(\lambda))$$

is a neighborhood of $\bar{\lambda}$,
then $\lambda_k \in \mathcal{A}$ for k large.



Proof:

– For k large, $\lambda_k \in \Omega$.
– When $\lambda_k \in \Omega$,
then there is a $\lambda \in \mathcal{A}$ such that $\lambda_k = \lambda + rg$ or $\lambda = \lambda_k - rg$ with $g \in \partial\delta(\lambda)$.
– This implies that $\lambda_{k+1} = \lambda \in \mathcal{A}$.

Special case: $0 \in \operatorname{int} \partial\delta(\bar{\lambda})$.

– Take $\mathcal{A} = \{\bar{\lambda}\}$.
– Since $\Omega = \bar{\lambda} + r\,\partial\delta(\bar{\lambda})$ is a neighborhood of $\bar{\lambda}$, one has $\lambda_k = \bar{\lambda}$ for large k.

Dual functions

- Remind the QP to solve:

$$\text{(P)} \quad \begin{cases} \inf_{(x,y)} \ \frac{1}{2}x^\top Q x + q^\top x \\ y = Cx \\ l \leq y \leq u, \end{cases}$$

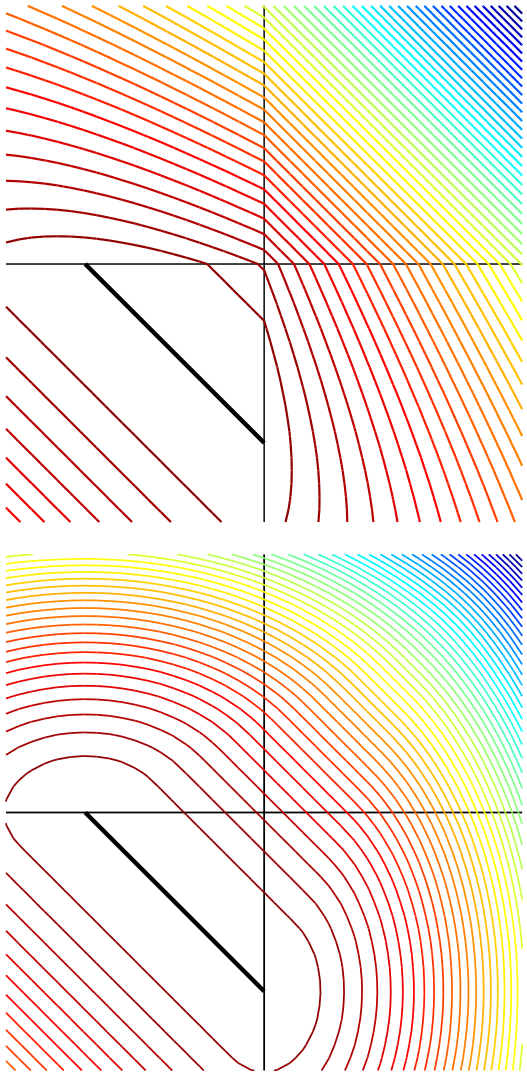
- Classical dual function associated with problem (P):
  for $H := CQ^{-1}C^\top$ and $v := -CQ^{-1}q$

$$\begin{aligned} &\delta_0(\lambda) \\ &:= \quad - \inf_{\substack{x \\ y \in [l,u]}} \left(\frac{1}{2} x^\top Q x + q^\top x + \lambda^\top (y - Cx) \right) \\ &= \quad \max_{\substack{y = (y_i)_{i=1}^m \\ y_i \in \{l_i, u_i\} \text{ if } \lambda_i \neq 0 \\ y_i = 0 \text{ if } \lambda_i = 0}} \left(\frac{1}{2} \lambda^\top H \lambda + (v - y)^\top \lambda + \frac{1}{2} q^\top Q^{-1} q \right). \end{aligned}$$

- Regularized dual function associated with problem (P):

$$\delta_r(\lambda) := - \inf_{\substack{(x,y) \\ l \leq y \leq u}} \left(\frac{1}{2} x^\top Q x + q^\top x + \lambda^\top (y - Cx) + \frac{r}{2} \|y - Cx\|_2^2 \right).$$

This function is $C^{1,1}$.



AL method as a proximal algorithm on δ_0 (Rockafellar, 1973)

- δ_r is the Moreau-Yosida regularization of δ_0:

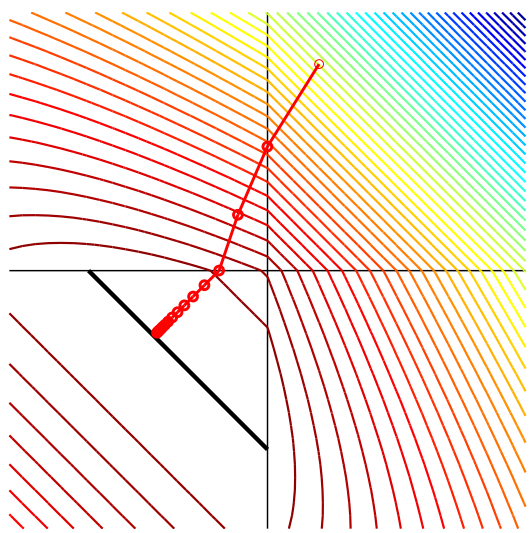
$$\delta_r(\lambda) = \inf_{\mu \in \mathbb{R}^m} \left(\delta_0(\mu) + \frac{1}{2r}\|\mu - \lambda\|^2 \right), \qquad \text{hence} \quad \nabla \delta_r(\lambda) = \tfrac{1}{r}(\lambda - \lambda_p).$$

- On the other hand:

$$\nabla \delta_r(\lambda) = \nabla_\lambda \ell_r(x_{\lambda,r}, y_{\lambda,r}, \lambda) = -(y_{\lambda,r} - C x_{\lambda,r}).$$

- Conclusion:

the proximal point $\lambda_p = \lambda + r(y_{\lambda,r} - C x_{\lambda,r})$ is the new AL iterate.



$r = 0.15$

slow because r is small

Property 1: finite time identification of the active constraints

The result

- Definition:

$$\mathrm{sign}(t) := \begin{cases} -1 & \text{if } t < 0 \\ 0 & \text{if } t = 0 \\ +1 & \text{if } t > 0 \end{cases}$$

Note that for a P-D solution $(\bar{x}, \bar{\lambda})$, there holds

$$\begin{array}{lcl} \bar{\lambda}_i > 0 & \Longrightarrow & \bar{y}_i = l_i \\ \bar{\lambda}_i = 0 & \Longleftarrow & l_i < \bar{y}_i < u_i \\ \bar{\lambda}_i < 0 & \Longrightarrow & \bar{y}_i = u_i \end{array}$$

- The result:

If · fixed $r > 0$,
· the AL algorithm generates a sequence $\{\lambda_k\} \to \bar{\lambda}$,
· strict complementarity at $\bar{\lambda}$,
then $\mathrm{sign}(\lambda_k) = \mathrm{sign}(\bar{\lambda})$ for k large.

The “proof”

- Take as a candidate for the finite time attractor:

$$\mathcal{A} := \{\lambda : \mathrm{sign}(\lambda) = \mathrm{sign}(\bar{\lambda})\}$$

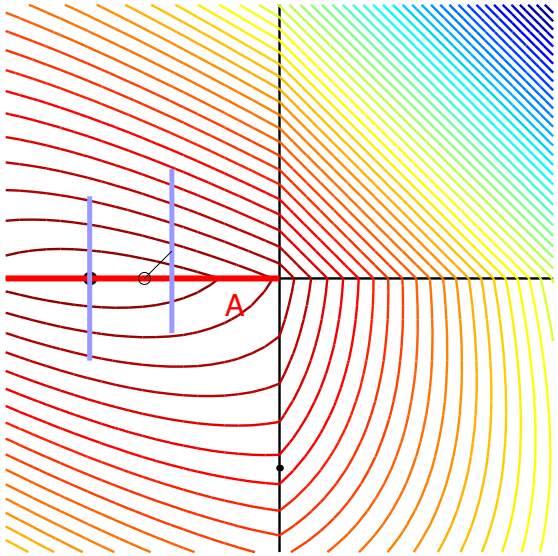
and show that

$$\Omega := \bigcup_{\lambda \in \mathcal{A}} (\lambda + r\,\partial\delta_0(\lambda)) \quad \text{is a neighborhood of } \bar{\lambda}.$$

- To conclude observe that
  - $0 \in \mathrm{ri}\,\partial\delta_0(\bar{\lambda}) = \mathrm{ri}(Cx_{\bar{\lambda}} - Y_{\bar{\lambda}})$

    this is due to the strict complementarity since

$$(Y_\lambda)_i = \begin{cases} [u_i, u_i] & \text{if } \lambda_i < 0 \\ [l_i, u_i] & \text{if } \lambda_i = 0 \\ [l_i, l_i] & \text{if } \lambda_i > 0 \end{cases}$$



  - $\lambda \in \mathcal{A} \mapsto \lambda + r\,\partial\delta_0(\lambda) = \varphi(\lambda) + \mathcal{J}$,

    where φ is an affine mapping ($\varphi(\bar{\lambda}) = \bar{\lambda}$) and $\mathcal{J} := r\partial\delta_0(\bar{\lambda})$.
  - $\varphi(\lambda) = (I + rH)\lambda + \gamma$ has a full rank property.

Observations

- Strict complementarity is essential.

- When r is larger, the set Ω "tends" to be larger.

  $\longrightarrow$ identification "tends" to be faster with large r.

- When r is larger, the convergence of λ_k to $\bar{\lambda}$ is faster.

  $\longrightarrow$ identification "tends" to be faster with large r.

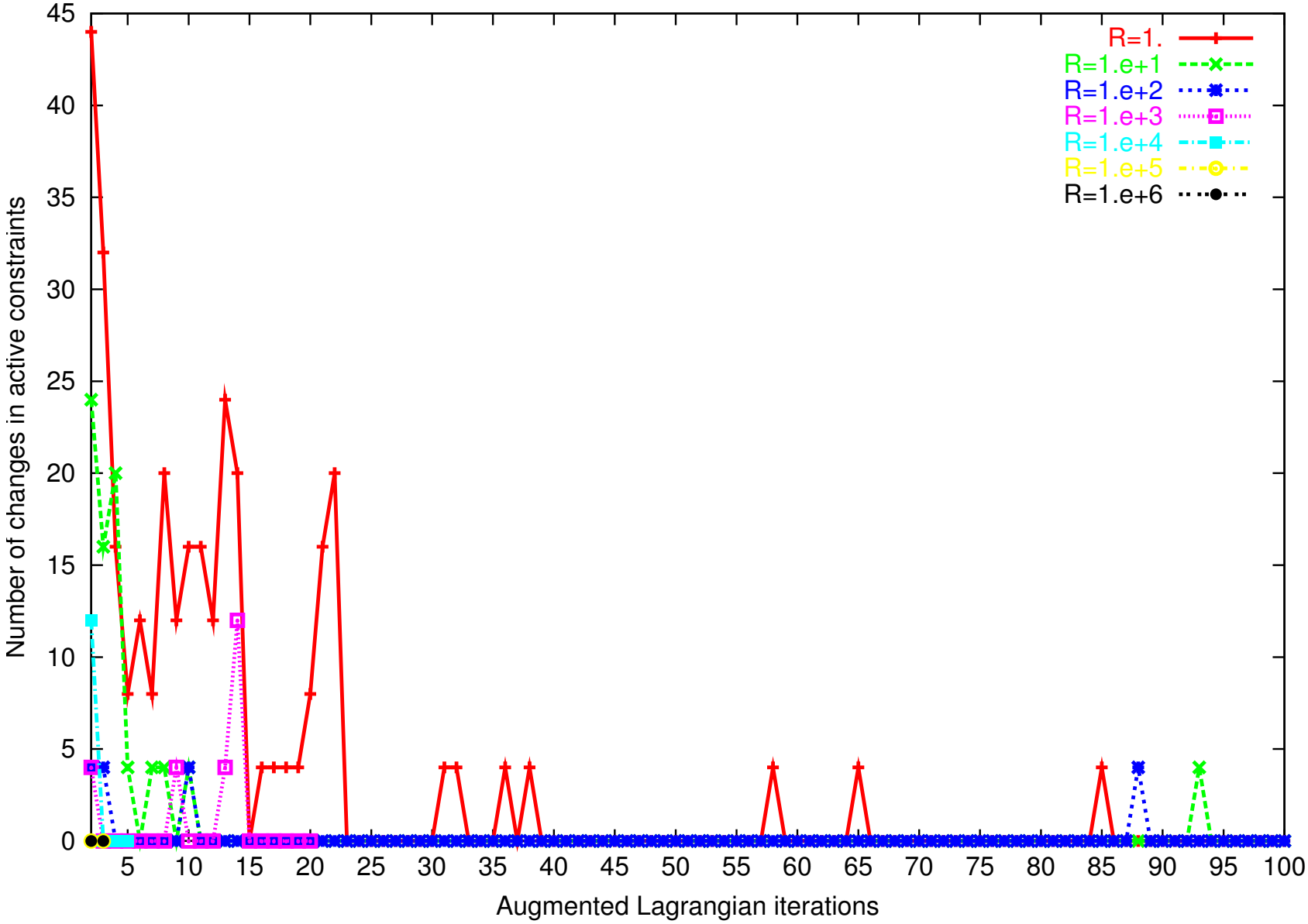
- The algorithm used to solve the Lagrange problem

$$\inf_{\substack{(x,y)\\ l\leq y\leq u}} \left(\frac{1}{2}x^\top Qx + q^\top x + \lambda^\top(y - Cx) + \frac{r}{2}\|y - Cx\|_2^2\right)$$

  should take advantage of this finite time identification property.

  $\longrightarrow$ a possible algorithm is: active set + gradient projection.

Illustrations



Property 2: GLOBAL linear convergence of $\lambda_k \to \bar{\lambda}$

The result

- **The result:**

  If $\{r_k\}$ is "sufficiently" bounded away from zero (say $r_k \geq \bar{r} > 0$, with a large $\bar{r}$), **then** $\forall\, \beta > 0$, $\exists\, \rho \in \,]0, 1[$, such that $\mathrm{dist}(\lambda_0, \mathcal{S}_D) \leq \beta$ implies that $\lambda_k \to \bar{\lambda}$ with

$$\|y_{k+1} - Cx_{k+1}\| \leq \rho \|y_k - Cx_k\|, \qquad \text{for all } k \geq 0.$$

- **Local linear convergence is clear:**

  - Identification of the active constraints in a finite number of iterations.
  - Global linear convergence for equality constrained problems.

- **The interest:**

  In fact ρ decreases when r increases,
  so that the result can be usefull to adapt r from iteration to iteration.

Changing the goal

- Substracting any $\bar{\lambda} \in \mathcal{S}_D$ from the iteration definition

$$\lambda_{k+1} = \lambda_k - r_k g_{k+1}, \quad \text{where} \quad g_{k+1} := -(y_k - Cx_k) \in \partial\delta_0(\lambda_{k+1}),$$

  squaring and using the monotonicity of $\partial\delta_0(\cdot)$, one gets

$$\|\lambda_{k+1} - \bar{\lambda}\|^2 + r_k^2 \|g_{k+1}\|^2 \leq \|\lambda_k - \bar{\lambda}\|^2. \tag{2}$$

- Assume that $(\partial\delta_0)^{-1}$ is radially Lipschitz at 0 in the sense that:

$$\boxed{\operatorname{dist}(\lambda, \mathcal{S}_D) \leq L\|g\|, \quad \forall\, \lambda \in \mathcal{S}_D + \mathcal{B}, \quad \forall\, g \in \partial\delta_0(\lambda).} \tag{3}$$

- Then

$$\|g_{k+1}\| \leq \min\left(\frac{L}{r_k}, 1\right) \|g_k\|,$$

  which implies the linear convergence of $\|y_k - Cx_k\|$ when r_k is sufficiently large.

- Therefore, we have to show (3).

The easy case: equality constrained problems

- Consider the problem

$$\begin{cases} \inf_x \ \frac{1}{2}x^\top Q x + q^\top x \\ Cx = b \end{cases}$$

- The dual function is (denote as before $H := CQ^{-1}C^\top$ and $v := -CQ^{-1}q$):

$$\delta_0(\lambda) = \frac{1}{2}\lambda^\top H\lambda + (v-b)^\top\lambda + \frac{1}{2}q^\top Q^{-1}q.$$

- For $g = \nabla\delta_0(\lambda)$ and $0 = \nabla\delta_0(\bar\lambda)$ can be written

$$g = H\lambda + v - b \qquad \text{and} \qquad 0 = H\bar\lambda + v - b.$$

- Substracting and multiplying by $\lambda - \bar\lambda$ yield

$$\|g\|\,\|\lambda - \bar\lambda\| \geq \boxed{g^\top(\lambda-\bar\lambda) = (\lambda-\bar\lambda)^\top H(\lambda-\bar\lambda).} \tag{4}$$

  - Either $C = 0$ and $b = 0$: then δ_0 is constant, $\mathcal{S}_D = \mathbb{R}^m$, and (3) is trivial.
  - Or $C \neq 0$: then for $\bar\lambda = P_{\mathcal{S}_D}(\lambda)$, $\lambda - \bar\lambda \in N(H)^\perp$, so that

$$(\lambda-\bar\lambda)^\top H(\lambda-\bar\lambda) \geq \sigma_1\|\lambda-\bar\lambda\|^2,$$

  where σ_1 is the smallest nonzero eigenvalue of H. Then (3) follows.

The more difficult case: inequality constrained problems

- The key identity (4) becomes (here $y_\lambda \in Y_\lambda$)

$$\|g\|\,\|\lambda-\bar\lambda\| \geq \boxed{g^\top(\lambda-\bar\lambda) = (\lambda-\bar\lambda)^\top H(\lambda-\bar\lambda) - (y_\lambda-\bar y)^\top(\lambda-\bar\lambda)} \overset{?}{\geq} \sigma\|\lambda-\bar\lambda\|^2.$$

Both terms in the RHS of the squared equation are ≥ 0. None of them can be neglected.

- A difficult situation: the second term vanishes.

This occurs when $\lambda \in \mathcal{O}$ (the optimal orthant).

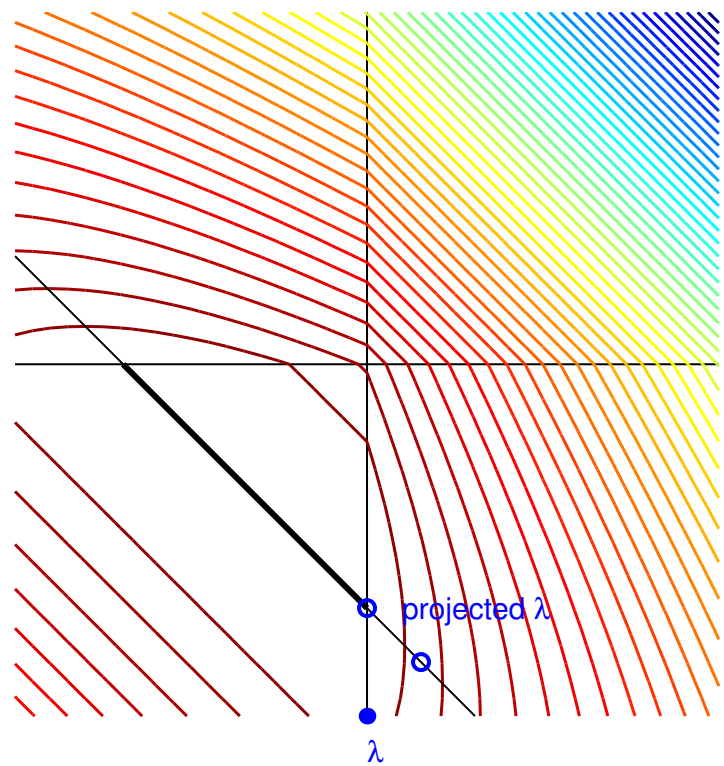
Use of the following result:

If · affine space $\mathcal{A} := \{x \in \mathbb{R}^n : Ax = b\}$,
· polyhedron $\mathcal{X} := \mathcal{A} \cap \mathbb{R}^n_+$,
then $\exists\, \gamma > 0, \forall x \in \mathbb{R}^n_+$:

$$\operatorname{dist}(x, \mathcal{A}) \geq \gamma \operatorname{dist}(x, \mathcal{X})$$

to conclude that

$$(\lambda-\bar\lambda)^\top H(\lambda-\bar\lambda) \geq \sigma\|\lambda-\bar\lambda\|^2.$$

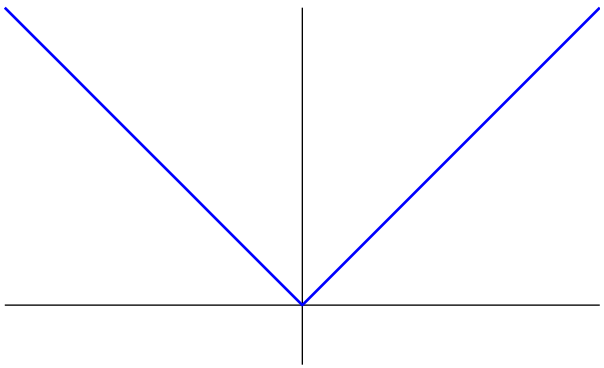


- Another difficult situation: the first term vanishes.

  This situation occurs, for example, when

$$C = 0_{1\times n}, \quad l = -1, \quad u = +1.$$

  Then $\delta_0(\lambda) = |\lambda|$ and $\partial\delta_0^{-1}$ is not *globally* Lipschitz.



  In that case it is possible to show that

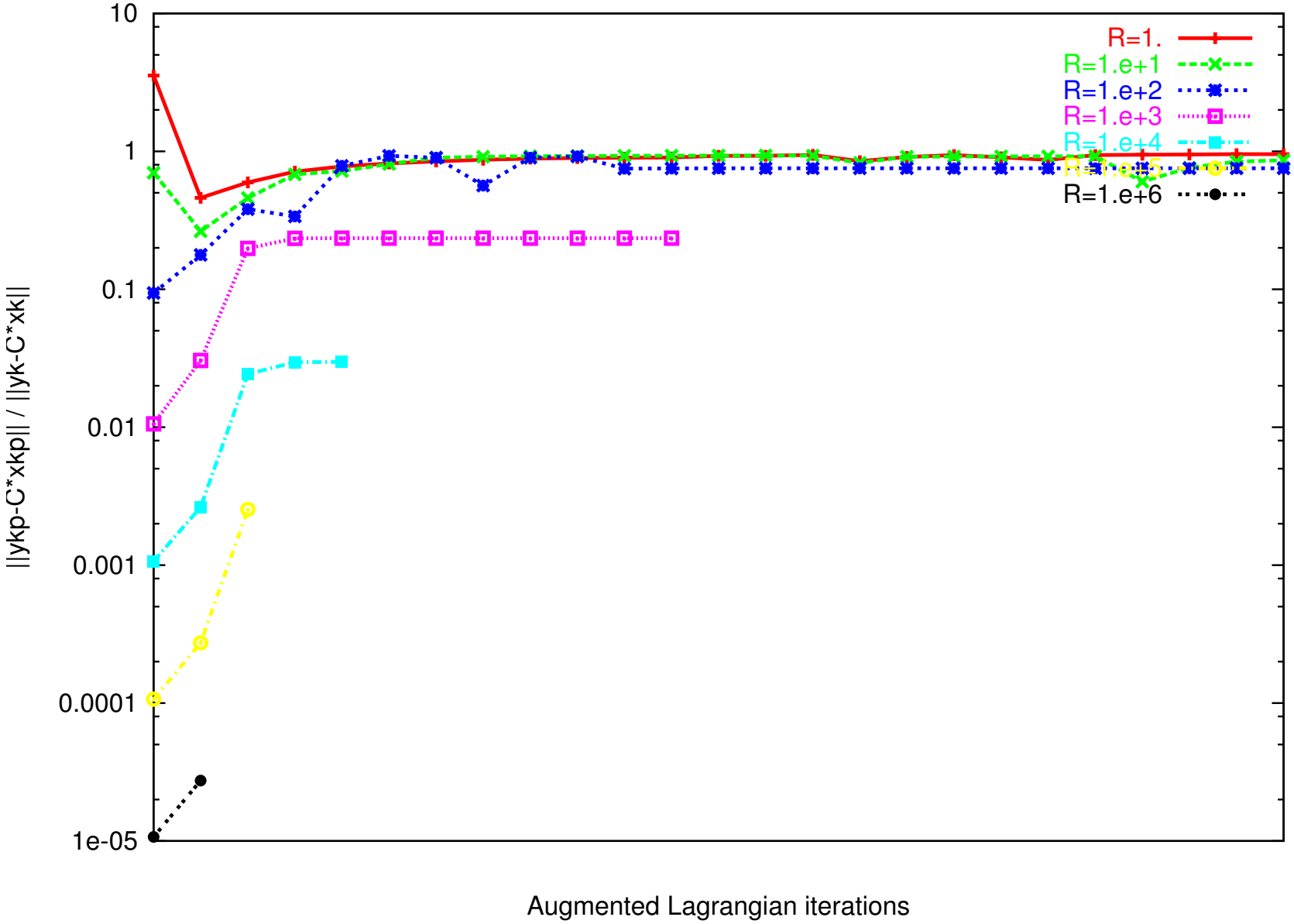
$$\begin{aligned} g^\top(\lambda - \bar{\lambda}) &\geq -(y_\lambda - \bar{y})^\top(\lambda - \bar{\lambda}) \\ &\geq -(y_\lambda - \bar{y})^\top(\lambda - \lambda_{\mathcal{O}}) \\ &\geq \gamma'\|\lambda - \bar{\lambda}\| \\ &\geq \frac{\gamma'}{\beta}\|\lambda - \bar{\lambda}\|^2, \end{aligned}$$

  since $\|\lambda - \bar{\lambda}\| \leq \beta$ (this is the reason for this assumption).

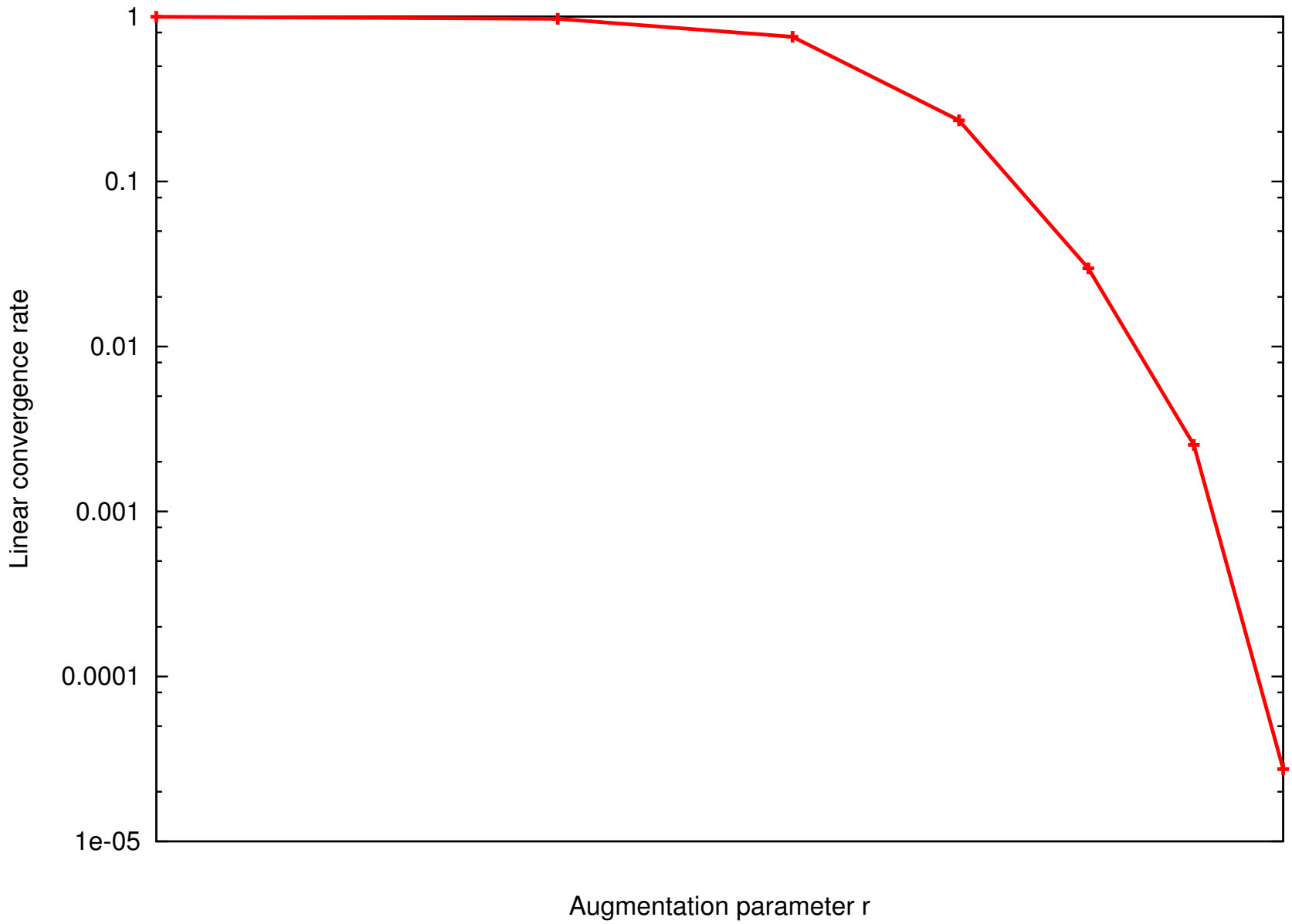
- Then using Cauchy-Schwarz on the LHS, the expected inequality holds in any cases: for some constant γ

$$\|g\| \geq \gamma\|\lambda - \bar{\lambda}\|.$$

An illustration



Linear rate as a function of r (log-scales)



Perspectives and references

- Extending the finite time indentification result to non-quadratic functions.

  F. Delbos and J.Ch. Gilbert (2003). Two properties of an augmented Lagrangian algorithm for solving a strictly convex quadratic program.

  $\longrightarrow$ a comparison with

  R. Mifflin and C. Sagastizábal (2002). Proximal points are on the fast track. *Journal of Convex Analysis* 9, 563-579.

  W. Hare and A. Lewis (2003). Identifying active constraints via partial smoothness and prox-regularity.

- Extending the finite time indentification result to the case of inexact minimization of the AL.

  $\longrightarrow$ see the contribution of

  Z. Dostál, A. Friedlander, and S.A. Santos (2003). Augmented Lagrangian with adaptative precision control for quadratic programming with simple bounds and equality constraints. *SIAM Journal on Optimization* 13, 1120-1140.

- Using the global linear convergence property to design an update rule for r_k.

  $\longrightarrow$ PhD thesis of F. Delbos, Institut Français du Pétrole (2004).