# How much patience do you have?
# Issues in complexity for nonlinear optimization

Philippe Toint (with Coralia Cartis and Nick Gould)

Namur Center for Complex Systems (naXys), University of Namur, Belgium

( philippe.toint@fundp.ac.be )

Cambridge, November 2015

# Thanks


- Leverhulme Trust, UK
- Balliol College, Oxford
- Belgian Fund for Scientific Research (FNRS)
- University of Namur, Belgium

# The problem

We consider the unconstrained nonlinear programming problem:

$$\text{minimize} \quad f(x)$$

for $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \rightarrow \mathbb{R}$ smooth.

Important special case: the nonlinear least-squares problem

$$\text{minimize} \quad f(x) = \tfrac{1}{2}\|F(x)\|^2$$

for $x \in \mathbb{R}^n$ and $F : \mathbb{R}^n \rightarrow \mathbb{R}^m$ smooth.

# A useful observation

Note the following: if

- $f$ has gradient $g$ and globally Lipschitz continuous Hessian $H$ with constant $2L$

Taylor, Cauchy-Schwarz and Lipschitz imply

$$
\begin{aligned}
f(x+s) &= f(x) + \langle s, g(x)\rangle + \tfrac{1}{2}\langle s, H(x)s\rangle \\
&\quad + \int_0^1 (1-\alpha)\langle s, [H(x+\alpha s) - H(x)]s\rangle \, d\alpha \\
&\leq \underbrace{f(x) + \langle s, g(x)\rangle + \tfrac{1}{2}\langle s, H(x)s\rangle + \tfrac{1}{3}L\|s\|_2^3}_{m(s)}
\end{aligned}
$$

$\Longrightarrow$ reducing $m$ from $s=0$ improves $f$ since $m(0) = f(x)$.

# Approximate model minimization

Lipschitz constant $L$ unknown $\Rightarrow$ replace by adaptive parameter $\sigma_k$ in the model :

$$m(s) \stackrel{\text{def}}{=} f(x) + s^T g(x) + \tfrac{1}{2} s^T H(x) s + \tfrac{1}{3}\sigma_k \|s\|_2^3 = T_{f,2}(x,s) + \tfrac{1}{3}\sigma_k \|s\|_2^3$$

Computation of the step:

1. minimize $m(s)$ until an approximate first-order minimizer is obtained:

$$\|\nabla_s m(s)\| \leq \kappa_{\text{stop}} \|s\|^2$$

(s-rule)
Note: no global optimization involved.

# Adaptive Regularization with Cubics (ARC2 or AR2)

**Algorithm 1.1: The ARC2 Algorithm**

Step 0: Initialization: $x_0$ and $\sigma_0 > 0$ given. Set $k = 0$

Step 1: Termination: If $\|g_k\| \leq \epsilon$, terminate.

Step 2: Step computation:

Compute $s_k$ such that $m_k(s_k) \leq m_k(0)$ and $\|\nabla_s m(s_k)\| \leq \kappa_{\text{stop}} \|s_k\|^2$.

Step 3: Step acceptance:

Compute $\rho_k = \dfrac{f(x_k) - f(x_k + s_k)}{f(x_k) - T_{f,2}(x_k, s_k)}$

and set $x_{k+1} = \begin{cases} x_k + s_k & \text{if } \rho_k > 0.1 \\ x_k & \text{otherwise} \end{cases}$

Step 4: Update the regularization parameter:

$$\sigma_{k+1} \in \begin{cases} [\sigma_{\min}, \sigma_k] & = \frac{1}{2}\sigma_k & \text{if } \rho_k > 0.9 & \textit{very successful} \\ [\sigma_k, \gamma_1 \sigma_k] & = \sigma_k & \text{if } 0.1 \leq \rho_k \leq 0.9 & \textit{successful} \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & = 2\sigma_k & \text{otherwise} & \textit{unsuccessful} \end{cases}$$

# Cubic regularization highlights

$$f(x+s) \leq m(s) \equiv f(x) + s^T g(x) + \tfrac{1}{2} s^T H(x) s + \tfrac{1}{3} L \|s\|_2^3$$

- Nesterov and Polyak minimize $m$ globally and exactly
  - N.B. $m$ may be non-convex!
  - efficient scheme to do so if $H$ has sparse factors
- global (ultimately rapid) convergence to a 2nd-order critical point of $f$
- better worst-case function-evaluation complexity than previously known

Obvious questions:

- can we avoid the global Lipschitz requirement? YES!
- can we approximately minimize $m$ and retain good worst-case function-evaluation complexity? YES !
- does this work well in practice? yes

# Evaluation complexity: an important result

How many function evaluations (iterations) are needed to ensure that

$$\boxed{\|g_k\| \leq \epsilon?}$$

If $H$ is globally Lipschitz and the s-rule is applied, the ARC2 algorithm requires at most

$$\left\lceil \frac{\kappa_{\rm S}}{\epsilon^{3/2}} \right\rceil \text{ evaluations}$$

for some $\kappa_{\rm S}$ independent of $\epsilon$.

c.f. Nesterov & Polyak

Note: an $O(\epsilon^{-3})$ bound holds for convergence to second-order critical points.

# Evaluation complexity: proof (1)

$$f(x_k + s_k) \le T_{f,2}(x_k, s_k) + \frac{L_f}{p} \|s_k\|^3$$

$$\|g(x_k + s_k) - \nabla_s T_{f,2}(x_k, s_k)\| \le L_f \|s_k\|^2$$

Lipschitz continuity of $H(x) = \nabla_x^2 f(x)$

$$\forall k \ge 0 \qquad f(x_k) - T_{f,2}(x_k, s_k) \ge \tfrac{1}{6}\sigma_{\min}\|s_k\|^3$$

$$f(x_k) = m_k(0) \ge m_k(s_k) = T_{f,2}(x_k, s_k) + \tfrac{1}{6}\sigma_k \|s_k\|^3$$

# Evaluation complexity: proof (2)

$$\exists \sigma_{\max} \quad \forall k \geq 0 \qquad \sigma_k \leq \sigma_{\max}$$

Assume that $\sigma_k \geq \dfrac{L_f(p+1)}{p\,(1-\eta_2)}$. Then

$$|\rho_k - 1| \leq \frac{|f(x_k+s_k) - T_{f,2}(x_k,s_k)|}{|T_{f,2}(x_k,0) - T_{f,2}(x_k,s_k)|} \leq \frac{L_f(p+1)}{p\,\sigma_k} \leq 1-\eta_2$$

and thus $\rho_k \geq \eta_2$ and $\sigma_{k+1} \leq \sigma_k$.

# Evaluation complexity: proof (3)

$$\forall k \ \text{ successful} \qquad \|s_k\| \geq \left( \frac{\|g(x_{k+1})\|}{L_f + \kappa_{\text{stop}} + \sigma_{\max}} \right)^{\frac{1}{2}}$$

$$\begin{aligned}
\|g(x_k+s_k)\| &\leq \|g(x_k+s_k) - \nabla_s T_{f,2}(x_k,s_k)\| \\
&\quad + \left\| \nabla_s T_{f,2}(x_k,s_k) + \sigma_k \|s_k\| s_k \right\| + \sigma_k \|s_k\|^2 \\
&\leq L_f \|s_k\|^2 + \|\nabla_s m(s_k)\| + \sigma_k \|s_k\|^2 \\
&\leq [L_f + \kappa_{\text{stop}} + \sigma_k] \, \|s_k\|^2
\end{aligned}$$

# Evaluation complexity: proof (4)

$$\|g(x_{k+1})\| \leq \epsilon \text{ after at most } \frac{f(x_0) - f_{\rm low}}{\kappa}\,\epsilon^{-3/2} \text{ successful iterations}$$

Let $\mathcal{S}_k = \{j \leq k \geq 0 \mid \text{iteration } j \text{ is successful}\}$.

$$\begin{aligned}
f(x_0) - f_{\rm low} &\geq f(x_0) - f(x_{k+1}) \geq \textstyle\sum_{j\in\mathcal{S}_k} \Big[f(x_i) - f(x_i + s_i)\Big] \\
&\geq \tfrac{1}{10} \sum_{j\in\mathcal{S}_k} \Big[f(x_i) - T_{f,2}(x_i, s_i)\Big] \geq |\mathcal{S}_k| \frac{\sigma_{\min}}{60} \min_i \|s_i\|^3 \\
&\geq |\mathcal{S}_k| \frac{\sigma_{\min}}{60\Big(L_f + \kappa_{\rm stop} + \sigma_{\max}\Big)^{3/2}} \min_i \|g(x_{i+1})\|^{3/2} \\
&\geq |\mathcal{S}_k| \frac{\sigma_{\min}}{60\Big(L_f + \kappa_{\rm stop} + \sigma_{\max}\Big)^{3/2}} \epsilon^{3/2}
\end{aligned}$$

# Evaluation complexity: proof (5)

$$k \le \kappa_u |\mathcal{S}_k|, \quad \text{where} \quad \kappa_u \stackrel{\text{def}}{=} \left(1 + \frac{|\log \gamma_1|}{\log \gamma_2}\right) + \frac{1}{\log \gamma_2} \log\left(\frac{\sigma_{\max}}{\sigma_0}\right),$$

$\sigma_k \in [\sigma_{\min}, \sigma_{\max}]$ + mechanism of the $\sigma_k$ update.

$$\|g(x_{k+1})\| \le \epsilon \text{ after at most } \frac{f(x_0) - f_{\rm low}}{\kappa} \, \epsilon^{-3/2} \text{ successful iterations}$$

One evaluation per iteration (successful or unsuccessful).

# Evaluation complexity: sharpness

Is the bound in $O(\epsilon^{-3/2})$ sharp? YES!!!

Construct a unidimensional example with

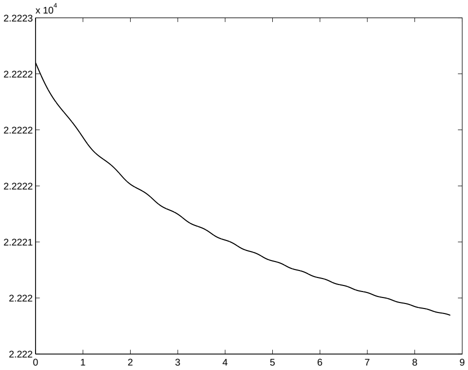
$$x_0 = 0, \quad x_{k+1} = x_k + \left(\frac{1}{k+1}\right)^{\frac{1}{3}+\eta},$$

$$f_0 = \frac{2}{3}\zeta(1+3\eta), \quad f_{k+1} = f_k - \frac{2}{3}\left(\frac{1}{k+1}\right)^{1+3\eta},$$

$$g_k = -\left(\frac{1}{k+1}\right)^{\frac{2}{3}+2\eta}, \quad H_k = 0 \text{ and } \sigma_k = 1,$$

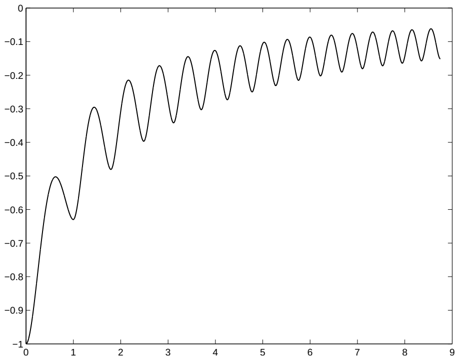
Use Hermite interpolation on $[x_K, x_{k+1}]$.

# An example of slow ARC2 (1)



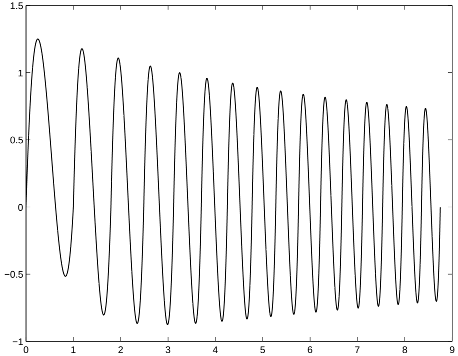

The objective function

# An example of slow ARC2 (2)



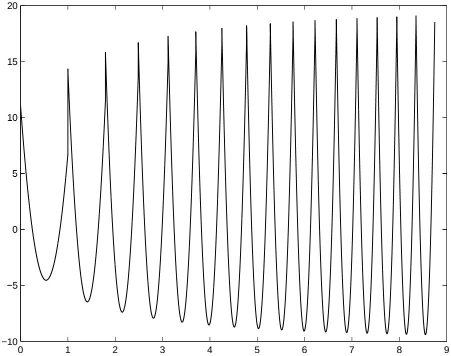

The first derivative

# An example of slow ARC2 (3)




The second derivative

# An example of slow ARC2 (4)




The third derivative

# Slow steepest descent (1)

The steepest descent method with requires at most

$$\left\lceil \frac{\kappa_C}{\epsilon^2} \right\rceil \text{ evaluations}$$

for obtaining $\|g_k\| \leq \epsilon$.

Nesterov

Sharp??? YES

Newton's method (when convergent) requires at most

$$O(\epsilon^{-2}) \text{ evaluations}$$

for obtaining $\|g_k\| \leq \epsilon$ !!!!

# Slow Newton (1)

Choose $\tau \in (0,1)$

$$g_k = -\begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ \left(\frac{1}{k+1}\right)^2 \end{pmatrix} \qquad H_k = \begin{pmatrix} 1 & 0 \\ 0 & \left(\frac{1}{k+1}\right)^2 \end{pmatrix},$$

for $k \geq 0$ and

$$f_0 = \zeta(1+2\eta) + \frac{\pi^2}{6}, \quad f_k = f_{k-1} - \frac{1}{2}\left[\left(\frac{1}{k+1}\right)^{1+2\eta} + \left(\frac{1}{k+1}\right)^2\right] \text{ for } k \geq 1,$$

$$\eta = \eta(\tau) \stackrel{\text{def}}{=} \frac{\tau}{4-2\tau} = \frac{1}{2-\tau} - \frac{1}{2}.$$

# Slow Newton (2)

$$H_k s_k = -g_k,$$

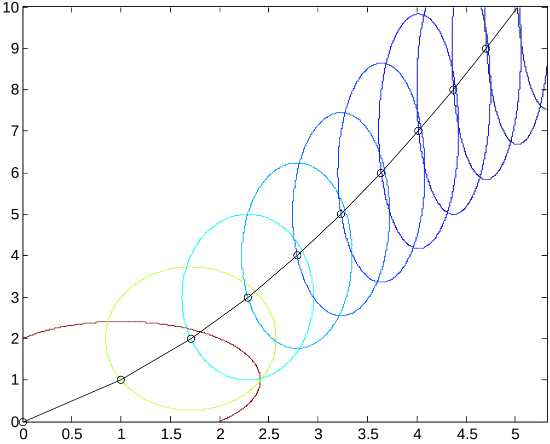
and thus

$$s_k = \begin{pmatrix} \left(\frac{1}{k+1}\right)^{\frac{1}{2}+\eta} \\ 1 \end{pmatrix},$$

$$x_0 = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad x_k = \begin{pmatrix} \sum_{j=0}^{k-1} \left(\frac{1}{j+1}\right)^{\frac{1}{2}+\eta} \\ k \end{pmatrix}.$$

# Slow Newton (3)

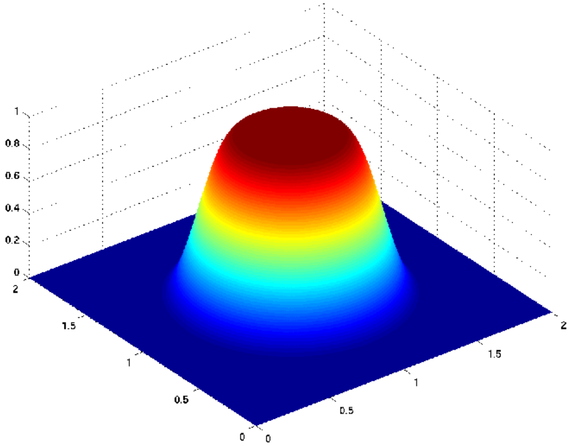
$$q_k(x_{k+1}, y_{k+1}) = f_k + \langle g_k, s_k\rangle + \tfrac{1}{2}\langle s_k, H_k s_k\rangle = f_{k+1}$$




The shape of the successive quadratic models

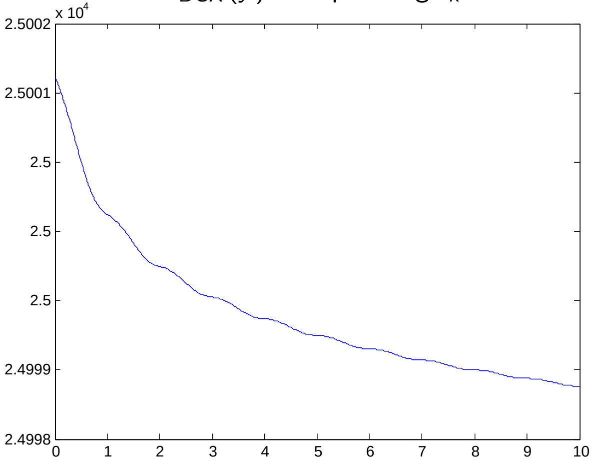
# Slow Newton (4)

Define a support function $s_k(x, y)$ around $(x_k, y_k)$



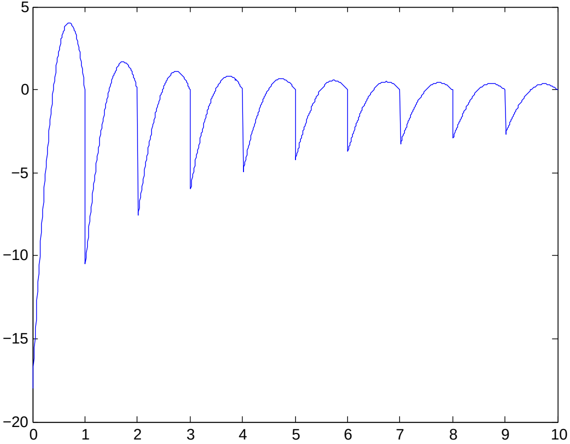
# Slow Newton (5)

A background function $f_{BCK}(y)$ interpolating $f_k$ values...



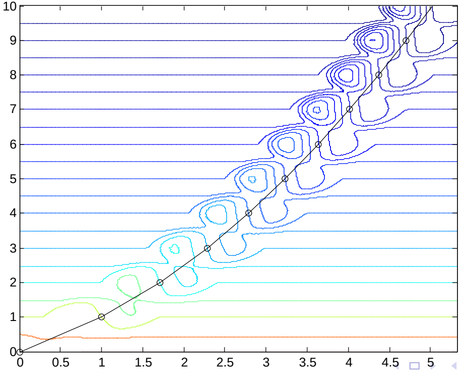
# Slow Newton (6)

...with bounded third derivative



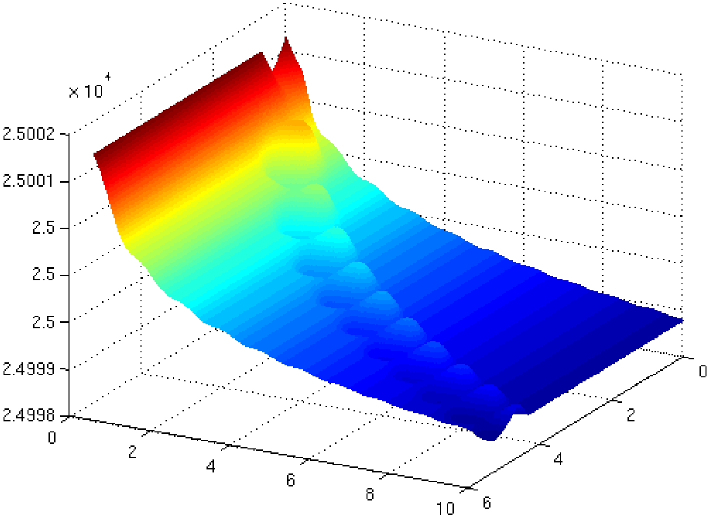
# Slow Newton (7)

$$f_{SN1}(x,y) = \sum_{k=0}^{\infty} s_k(x,y) q_k(x,y) + \left[1 - \sum_{k=0}^{\infty} s_k(x,y)\right] f_{BCK}(x,y)$$



# Slow Newton (8)

Some steps on a sandy dune. . .



# More general second-order methods

Assume that, for $\beta \in (0,1]$, the step is computed by

$$(H_k + \lambda_k I)s_k = -g_k \quad \text{and} \quad 0 \le \lambda_k \le \kappa_s \|s_k\|^\beta$$

(ex: Newton, ARC2, Levenberg-Morrison-Marquardt, (trust-region), ...)

The corresponding method terminates in at most

$$\left\lceil \frac{\kappa_C}{\epsilon^{(\beta+2)/(\beta+1)}} \right\rceil \text{ evaluations}$$

to obtain $\|g_k\| \le \epsilon$ on functions with bounded and (segment-wise) $\beta$-Hölder continuous Hessians.

Note: ranges form $\epsilon^{-2}$ to $\epsilon^{-3/2}$

ARC2 is optimal within this class

# High-order models (1)

What happens if one considers the model

$$m_k(s) = T_{f,p}(x_k, s) + \frac{\sigma_k}{p!}\|s\|_2^{p+1}$$

where

$$T_{f,p}(x,s) = f(x) + \sum_{j=1}^{p} \frac{1}{j!}\nabla_x^j f(x)[s]^j$$

terminating the step computation when

$$\|\nabla_s m(s_k)\| \leq \kappa_{\text{stop}}\|s_k\|^p$$

???

now the AR$p$ method!

# High-order models (2)

$\epsilon$-approx 1rst-order critical point after at most

$$\frac{f(x_0) - f_{\rm low}}{\kappa} \, \epsilon^{-\frac{p+1}{p}}$$

successful iterations

Moreover

$\epsilon$-approx "$q$-th order critical point" after at most

$$\frac{f(x_0) - f_{\rm low}}{\kappa} \, \epsilon^{-\frac{p+1}{p+1-q}}$$

successful iterations

# The constrained case

**Can we apply regularization to the constrained case?**

Consider the constrained nonlinear programming problem:

$$\begin{array}{ll} \text{minimize} & f(x) \\ & x \in \mathcal{F} \end{array}$$

for $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ smooth, and where

$$\mathcal{F} \text{ is convex.}$$

**Ideas:**

- exploit (cheap) projections on convex sets
- use appropriate termination criterion

$$\chi_f(x_k) \stackrel{\text{def}}{=} \left| \min_{x+d \in \mathcal{F}, \|d\| \leq 1} \langle \nabla_x f(x_k), d \rangle \right|,$$

# Constrained step computation

$$\min_s \quad T_{f,2}(x,s) + \tfrac{1}{3}\sigma\|s\|^3$$

subject to

$$x + s \in \mathcal{F}$$

- minimization of the cubic model until an approximate first-order critical point is met, as defined by

$$\chi_m(s) \leq \kappa_{\text{stop}}\|s\|^2$$

c.f. the "s-rule" for unconstrained

Note: OK at local constrained model minimizers

# A constrained regularized algorithm

**Algorithm 4.1: ARC for Convex Constraints (ARC2CC)**

Step 0: Initialization. $x_0 \in \mathcal{F}$, $\sigma_0$ given. Compute $f(x_0)$, set $k = 0$.

Step 1: Termination. If $\chi_f(s_k) \leq \epsilon$, terminate.

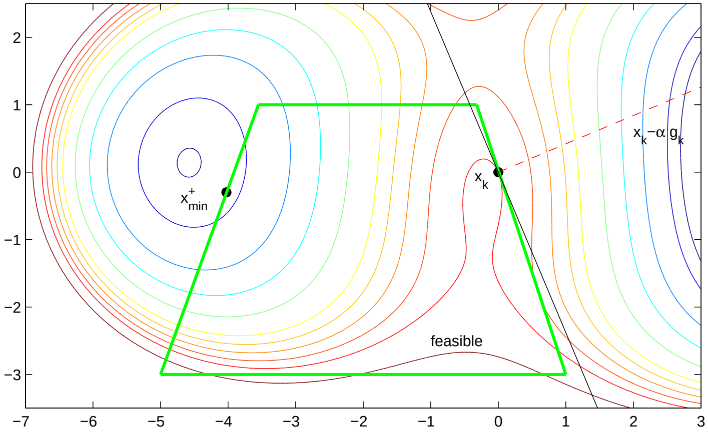
Step 2: Step calculation. Compute $s_k$ and $x_k^+ \stackrel{\rm def}{=} x_k + s_k \in \mathcal{F}$ such that $\chi_m(s_k) \leq \kappa_{\rm stop} \|s_k\|^2$.

Step 3: Acceptance of the trial point. Compute $f(x_k^+)$ and $\rho_k$. If $\rho_k \geq \eta_1$, then $x_{k+1} = x_k + s_k$; otherwise $x_{k+1} = x_k$.

Step 4: Regularisation parameter update. Set

$$\sigma_{k+1} \in \left\{ \begin{array}{ll} [\sigma_{\min}, \sigma_k] & \text{if } \rho_k \geq \eta_2, \\ [\sigma_k, \gamma_1 \sigma_k] & \text{if } \rho_k \in [\eta_1, \eta_2), \\ [\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{if } \rho_k < \eta_1. \end{array} \right.$$

# Walking through the pass...




A “beyond the pass” constrained problem with

$$m(x,y) = -x - \tfrac{42}{100}y - \tfrac{3}{10}x^2 - \tfrac{1}{10}y^3 + \tfrac{1}{3}[x^2+y^2]^{\frac{3}{2}}$$

# Evaluation Complexity for ARC2CC

The ARC2CC algorithm requires at most

$$\left\lceil \frac{\kappa_{\rm C}}{\epsilon^{3/2}} \right\rceil \text{ evaluations}$$

(for some $\kappa_{\rm C}$ independent of $\epsilon$) to achieve $\chi_f(x_k) \leq \epsilon$

Caveat: cost of solving the subproblem!

Higher-order models/critical points: $\left\lceil \dfrac{\kappa_{\rm C}}{\epsilon^{(p+1)/(p+1-q)}} \right\rceil$ evaluations

Identical to the unconstrained case!!!

# The general constrained case

Consider now the general NLO (slack variables formulation):

$$\begin{array}{ll} \text{minimize}_x & f(x) \\ \text{such that} & c(x) = 0 \quad \text{and} \quad x \in \mathcal{F} \end{array}$$

Ideas for a second-order algorithm:

1. get $\|c(x)\| \leq \epsilon$ (if possible) by minimizing $\|c(x)\|^2$ such that $x \in \mathcal{F}$ (getting $\|J(x)^T c(x)\|$ small unsuitable!)
2. track the "trajectory"

$$\mathcal{T}(t) \stackrel{\text{def}}{=} \{x \in \mathbf{R}^n \mid c(x) = 0 \quad \text{and} \quad f(x) = t\}$$

for values of $t$ decreasing from $f$(first feasible iterate) while preserving $x \in \mathcal{F}$

# First-order complexity for general NLO (1)

Sketch of a two-phases algorithm:

feasibility: apply ARC2CC to

$$\min_x \nu(x) \stackrel{\text{def}}{=} \|c(x)\|^2 \quad \text{such that} \quad x \in \mathcal{F}$$

at most $O(\epsilon_P^{-1/2}\epsilon_D^{-3/2})$ evaluations

tracking: successively

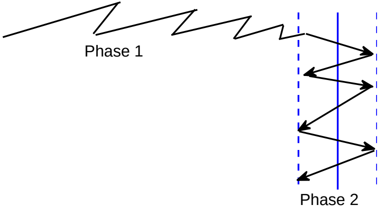
- apply ARC2CC (with specific termination test) to

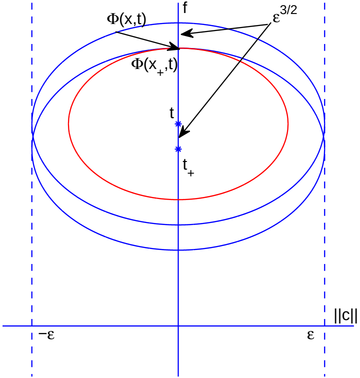
$$\min_x \mu(x) \stackrel{\text{def}}{=} \|c(x)\|^2 + (f(x) - t)^2 \quad \text{such that} \quad x \in \mathcal{F}$$

- decrease $t$ (proportionally to the decrease in $\phi(x)$)

at most $O(\epsilon_P^{-1/2}\epsilon_D^{-3/2})$ evaluations

# A view of Algorithm ARC2CC



# First-order complexity for general NLO (2)

Under the "conditions stated above", the ARC2CC algorithm takes at most

$$O(\epsilon_P^{-1/2}\epsilon_D^{-3/2}) \text{ evaluations}$$

to find an iterate $x_k$ with either

$$\|c(x_k)\| \leq \delta\epsilon_P \quad \text{and} \quad \chi_{\mathcal{L}} \leq \|(y,1)\|\epsilon_D$$

for some Lagrange multiplier $y$ and where

$$\mathcal{L}(x,y) = f(x) + \langle y, c(x)\rangle,$$

or

$$\|c(x_k)\| > \delta\epsilon \quad \text{and} \quad \chi_{\|c\|} \leq \epsilon.$$

# Conclusions

- Complexity analysis for first-order points using second-order methods

$$O(\epsilon^{-3/2}) \;\; \text{(unconstrained, convex constraints)}$$
$$O(\epsilon_p^{-1/2}\epsilon_d^{-3/2}) \;\; \text{(equality and general constraints)}$$

- Available also for $p$-th order methods :

$$O(\epsilon^{-(p+1)/p}) \;\; \text{(unconstrained, convex constraints)}$$
$$O(\epsilon_p^{-1/p}\epsilon_d^{-(p+1)/p}) \;\; \text{(equality and general constraints)}$$

- Jarre's example $\Rightarrow$ global optimization much harder
- ARC2 is optimal amongst second-order method
- More also known (DFO, non-smooth, etc)

**Many thanks for your attention!**