On some numerical ODE techniques in nonlinear optimization

Workshop on Control and Optimization with Differential-Algebraic Constraints, Banff International Research Station for Mathematical Innovation and Discovery, Banff, Alberta, Canada

Laurent O. Jay
joint work with Darin G. Mohr (PhD student)

Dept. of Mathematics, The University of Iowa, USA

October 24-29, 2010

Unconstrained nonlinear minimization problem

$$\min_{x \in \mathbb{R}^n} f(x)$$

$f : \mathbb{R}^n \to \mathbb{R}$ is the objective function

we suppose $f \in \mathcal{C}^2$

Line search methods in a nutshell

- Iterations: $x_{k+1} = x_k + \alpha_k p_k$
- Line search direction: $p_k = -{H_k}^{-1} \nabla f(x_k)$

  H_k positive definite

  $\Longrightarrow$ 1. $\nabla f(x_k)^T p_k < 0$, p_k is a descent direction

  2. $f(x_k + \alpha p_k) < f(x_k)$ for $\alpha > 0$ sufficiently small,

  p_k is a direction of decrease

- Steplength: $\alpha_k > 0$

Choices for H_k

H_k is chosen positive definite and symmetric

- Steepest descent: $H_k := I$
- Newton: $H_k := \nabla^2 f(x_k)$ (generally not positive definite!)
- 'Modified Newton': $H_k := \lambda_k I + \nabla^2 f(x_k)$
- Quasi-Newton (secant method in dim $n = 1$):
  - Quasi-Newton condition for H_k:

$$H_k s_{k-1} = y_{k-1}$$

where

$$s_{k-1} := x_k - x_{k-1}, \quad y_{k-1} := \nabla f(x_k) - \nabla f(x_{k-1})$$

  - Example: the BFGS formula:

$$H_k := H_{k-1} + \frac{1}{s_{k-1}^T y_{k-1}} y_{k-1} y_{k-1}^T - \frac{1}{s_{k-1}^T z_{k-1}} z_{k-1} z_{k-1}^T$$

where $z_{k-1} := H_{k-1} s_{k-1}$

Conditions on α_k for global convergence

Let $m(\alpha) := f(x(\alpha))$ where the curve $x(\alpha) := x_k + \alpha p_k$ is a straight line

The Wolfe conditions of sufficient decrease

$$\begin{aligned} m(\alpha_k) &\leq m(0) + \eta_1 \alpha_k m'(0) \\ \eta_2 m'(0) &\leq m'(\alpha_k) \end{aligned}$$

e.g., with parameters $0 < \eta_1 = 10^{-4} < \eta_2 = 0.9 < 1$.

To avoid extra gradient evalutions the Goldstein and Price conditions of sufficient decrease replace $m'(\alpha_k)$ above by

$$\frac{m(\alpha_k) - m(0)}{\alpha_k}$$

Justification of p_k by a quadratic model?

$p_k = -{H_k}^{-1}\nabla f(x_k)$ is the minimizer of the "quadratic model"

$$Q_k(p) := f(x_k) + \nabla f(x_k)^T p + \frac{1}{2} p^T H_k p$$

of the nonlinear function $f(x_k + p)$

- $Q_k(p) \approx f(x_k + p)$ is valid for $\|p\|$ small only! However $\|p_k\|$ is not small in general, unless x_k is already close to a stationary point ($\nabla f(x_k) \approx 0$)
- $\alpha_k p_k$ is the minimizer of $Q_k(p/\alpha_k)$ which has no reason to be a good model of $f(x_k + p)$!

Hence, the choice of the search direction $p_k = -{H_k}^{-1}\nabla f(x_k)$ cannot be justified by $Q_k(p)$ being a good quadratic model. The only possible justification is that of p_k being a direction of decrease

An interpretation of line search methods

The implicit Euler method applied to the (possibly stiff) gradient system

$$\frac{d}{dt}y = -\nabla f(y), \quad y(0) = x_k$$

with artificial time t and stepsize $h_k > 0$ reads

$$y_{k+1} - (x_k - h_k \nabla f(y_{k+1})) = 0$$

One modified Newton iteration with initial guess $y_{k+1}^{(0)} := x_k$ and modified Jacobian $M_k := I + h_k B_k \approx I + h_k \nabla^2 f(y_{k+1})$ leads to

$$\begin{aligned} y_{k+1}^{(1)} &= y_{k+1}^{(0)} - {M_k}^{-1} h_k \nabla f(y_{k+1}^{(0)}) \\ &= x_k - {M_k}^{-1} h_k \nabla f(x_k) = x_k - {H_k}^{-1} \nabla f(x_k) \end{aligned}$$

where $H_k := \lambda_k I + B_k$ and $\lambda_k := 1/h_k$ ($\lambda_k = 0$ corresponds to $h_k = +\infty$!)

An interpretation of line search methods (cont.)

Dense (continuous) output approximate solution passing through x_k and $y_{k+1}^{(1)}$

$$x(\alpha) = x_k - \alpha {H_k}^{-1} \nabla f(x_k) = x_k - \alpha {M_k}^{-1} h_k \nabla f(x_k) \approx y(\alpha h_k)$$

The above interpretation gives a posteriori justifications

- for certain search directions $p_k = -{H_k}^{-1} \nabla f(x_k)$, in particular for the 'modified Newton' method of nonlinear optimization

$$H_k := \lambda_k I + \nabla^2 f(x_k) \text{ instead of } H_k := \nabla^2 f(x_k)$$

  to ensure positive definiteness

- for the initial matrix $H_0 := \lambda_0 I$ of quasi-Newton methods, the choice of λ_0 can be related to the choice of an initial stepsize h_0 through the relation $\lambda_0 = 1/h_0$

Main motivation for numerical ODE methods

Consider the curve $y(t)$ solution to the gradient system

$$\frac{d}{dt}y = -\nabla f(y), \quad y(0) = x_k$$

Theorem

Let $f \in \mathcal{C}^2$ and $\nabla f(x_k) \neq 0 \implies f(y(t)) < f(y(s))$ for $s < t$. Moreover, if the set $\{y \in \mathbb{R}^n \mid f(y) \leq f(x_k)\}$ is compact $\implies$ $\lim_{t\to+\infty} \nabla f(y(t)) = 0$

The curve $y(t)$ is a descent curve for f. It has clearly better properties than the straight line

$$x(\alpha) = x_k + \alpha p_k = x_k - \alpha H_k^{-1} \nabla f(x_k)$$

Descent ODEs for f

- More generally we can consider descent ODEs for f

$$\frac{d}{dt}y = -K_k(t,y)\nabla f(y), \quad y(0) = x_k$$

  with positive definite matrices $K_k(t,y)$
- The flow for $K_k(t,y) := (\nabla^2 f(y))^{-1}$ is called the Newton flow
- Descent curves for f $y(t)$ are more than decent!

Numerical ODE methods in nonlinear optimization

Numerical ODE methods in nonlinear optimization have been considered and rediscovered several times in the past

- Paul T. Boggs (1971, 1977)
- Charalambos-Apostolos Emmanuel Botsaris (1976, 1978)
- Jean-Philippe Vial and Israel Zang (1977)
- Jan A. Snyman (1982)
- Michael C. Bartholomew-Biggs and A. A. Brown (1989)
- Johannes Schropp (1995, 1997, 1999, 2001)
- John A. Ford (1996, 2003)
- William Behrmann (1998, PhD student of Walter Murray)
- Desmond J. Higham (1999)
- Neculai Andrei (2004)
- Pierre-Antoine Absil (2006)
- Tim Kelley et al. (2006, 2009)

A quote from Des Higham (1999)

". . . the possibility of combining optimization and ODE ideas forms an attractive area for future work."

Numerical ODE methods in nonlinear optimization (cont.)

To my knowledge numerical ODE methods are not mentioned in any nonlinear optimization (text)book.

- They may be used only as a last resort when other methods fail
- Their clear advantage: robustness
- Their apparent disadvantage: cost per iteration

What really matters? Efficiency

Personal conviction

By mixing numerical ODE techniques with nonlinear optimization techniques more efficient methods can be obtained.

New Quasi-Newton type condition

The quasi-Newton condition

$$H_k s_{k-1} = y_{k-1}$$

where $s_{k-1} := x_k - x_{k-1}$, $y_{k-1} := \nabla f(x_k) - \nabla f(x_{k-1})$ is originally motivated by the relation

$$\nabla^2 f(x_k) s_{k-1} = y_{k-1} + o(\|s_{k-1}\|)$$

Since we are interested in matrices of the form

$$H_k = \lambda_k I + B_k \approx \lambda_k I + \nabla^2 f(x_k)$$

it is natural to consider the new quasi-Newton type condition

$$H_k s_{k-1} = \lambda_k s_{k-1} + y_{k-1}$$

Theorem

$s_{k-1}^T y_{k-1} > 0$, $\lambda_k > 0$, *and* H_{k-1} *s.p.d.* $\Longrightarrow$ H_k^{BFGS} *s.p.d.*

New Quasi-Newton type condition (cont.)

Similarly for matrices of the form

$$M_k = I + h_k B_k \approx I + h_k \nabla^2 f(x_k)$$

it is natural to consider the new quasi-Newton type condition

$$M_k s_{k-1} = s_{k-1} + h_k y_{k-1}$$

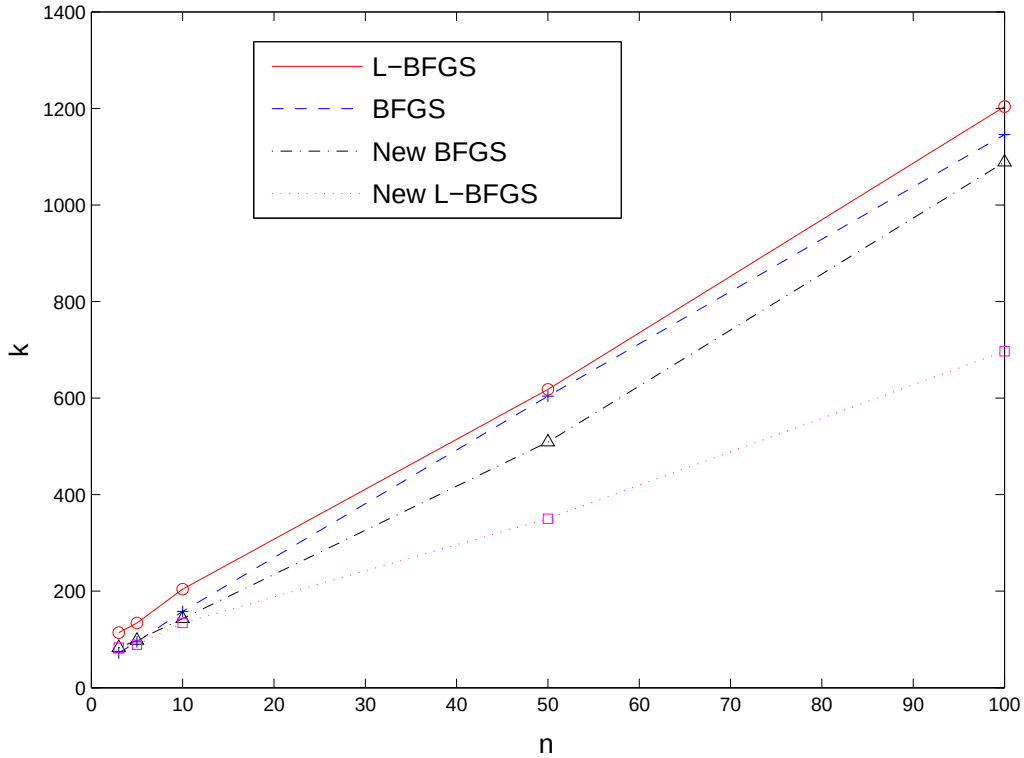
Theorem

$s_{k-1}^T y_{k-1} > 0$, $h_k > 0$, *and* M_{k-1} *s.p.d.* $\Longrightarrow$ M_k^{BFGS} *s.p.d.*

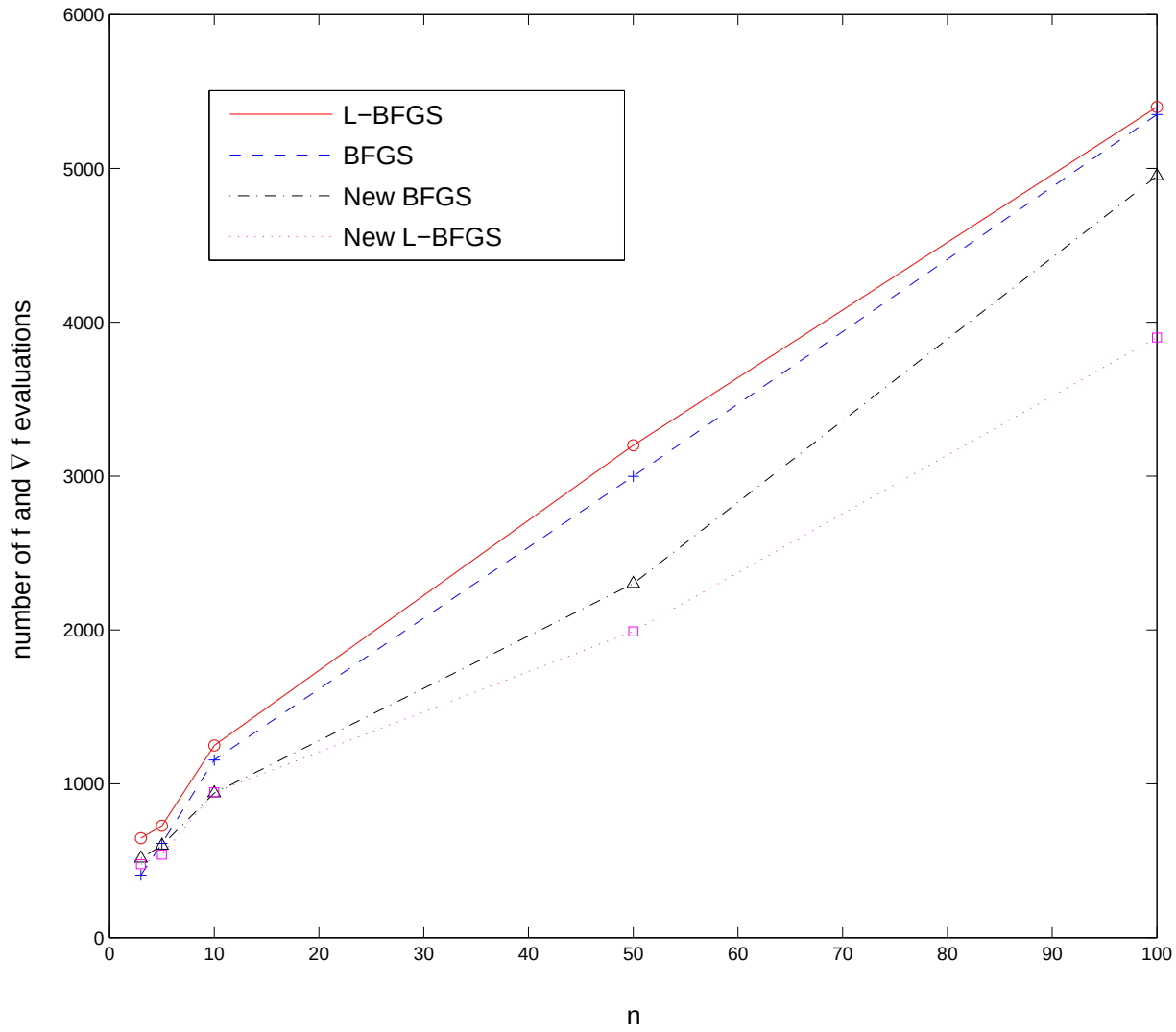
Numerical results for the Rosenbrock problem

$$f(x) := \sum_{i=1}^{n-1} \left((1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2 \right)$$

$H_0 = I$ and
$m = 12$ for limited memory BFGS (L-BFGS)

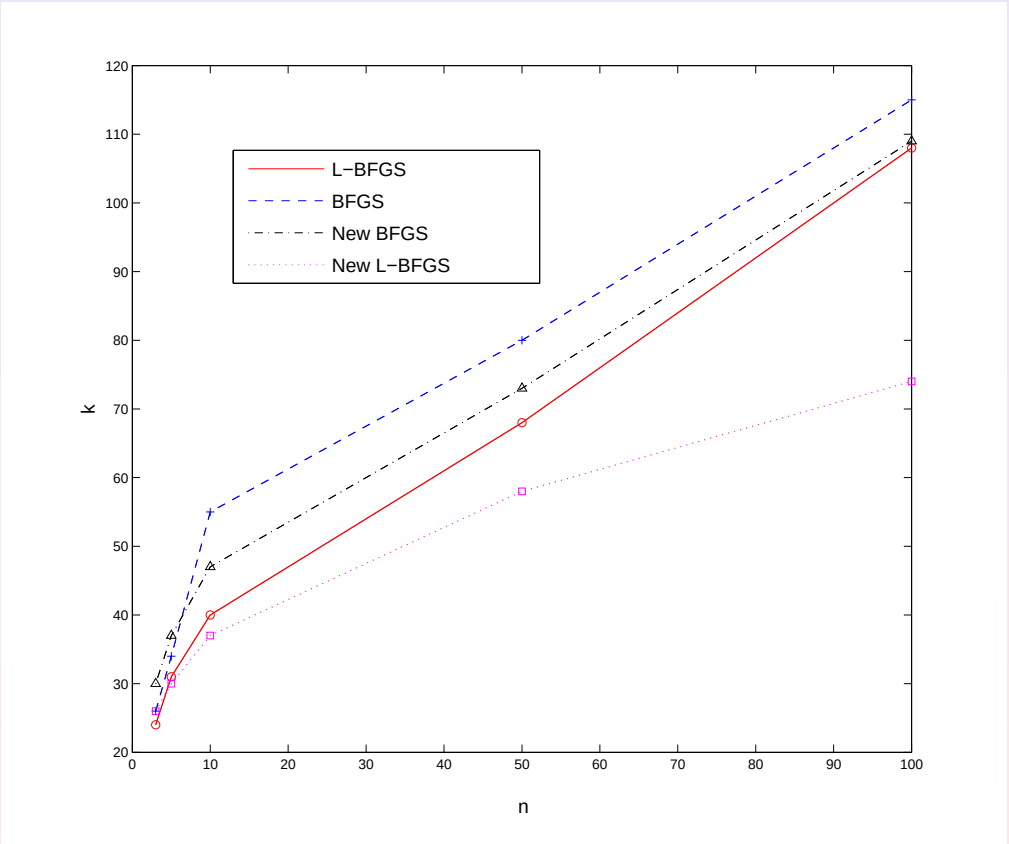


Numerical results for the Rosenbrock problem (cont.)

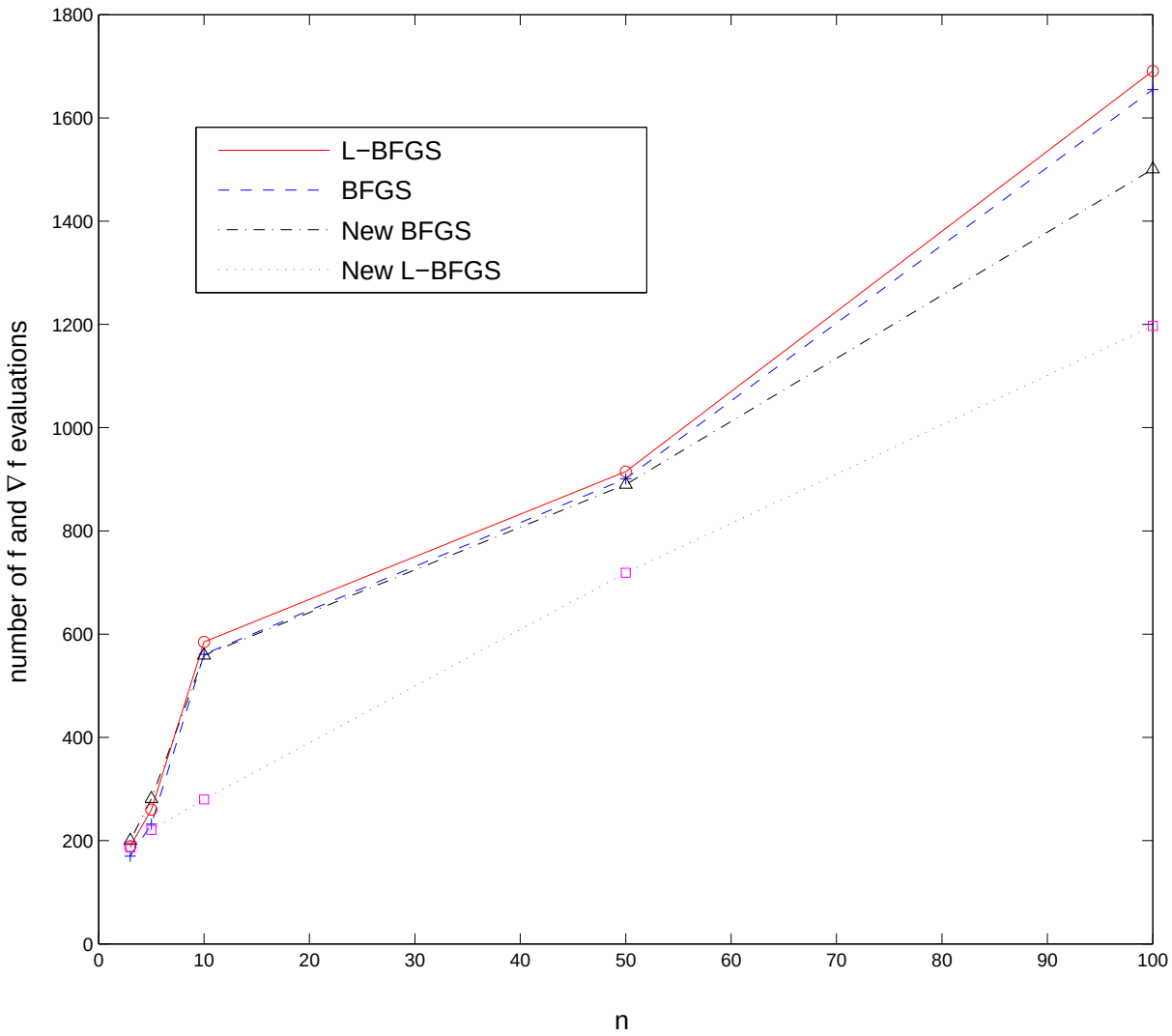


Numerical results for the Zakharov problem

$$f(x) := \sum_{i=1}^{n} x_i^2 + \frac{1}{4}\left(\sum_{i=1}^{n} i \cdot x_i\right)^2 + \frac{1}{16}\left(\sum_{i=1}^{n} i \cdot x_i\right)^4$$



Numerical results for the Zakharov problem (cont.)



2-stage IRK methods and dense output

The implicit Euler method is not the most efficient method for ODEs. It is only of order 1. We can consider dense output curves $x(\alpha)$ of 2-stage implicit Runge-Kutta (IRK) methods

$$\begin{aligned} Y_1 &= x_k - h_k\left(a_{11}\nabla f(Y_1) + a_{12}\nabla f(Y_2)\right) \\ Y_2 &= x_k - h_k\left(a_{21}\nabla f(Y_1) + a_{22}\nabla f(Y_2)\right) \\ x(\alpha) &= x_k - h_k\left(b_1(\alpha)\nabla f(Y_1) + b_2(\alpha)\nabla f(Y_2)\right) \approx y(\alpha h_k) \end{aligned}$$

with the following desirable properties:

- stage order 2 ($\equiv$ simplifying assumption $C(2) \equiv$ collocation)
- dense output of order 2 (the collocation polynomial)
- coefficient matrix A with real positive eigenvalues
- stiffly accurate for $\alpha = 1$, i.e., $x(1) = Y_2$.
- L-stable for $\alpha = 1$

2-stage IRK methods and dense output (cont.)

We obtain a one-parameter family of order 2 methods with

$$c_1 \in]0, 3-2\sqrt{2}] \cup [3+2\sqrt{2}, +\infty[\approx]0, 0.17157] \cup [5.8284, +\infty[$$

- Methods for $c_1 \in]0, 3-2\sqrt{2}]$ and $\widetilde{c}_1 \in [3+2\sqrt{2}, +\infty[$ are in fact equivalent! The latter corresponds to $\widetilde{c}_1 = 1/c_1$, $\widetilde{Y}_1 = Y_2$, $\widetilde{Y}_2 = Y_1$, and stepsize $\widetilde{h}_k = c_1 h_k$.
- For $c_1 := 1/8 = 0.125$ we obtain

$$\begin{array}{c|cc} c_1 & a_{11} & a_{12} \\ c_2 & a_{21} & a_{22} \\ \hline & b_1(\alpha) & b_2(\alpha) \end{array} = \begin{array}{c|cc} 1/8 & 15/112 & -1/112 \\ 1 & 4/7 & 3/7 \\ \hline & (16\alpha - 8\alpha^2)/14 & (-\alpha + 4\alpha^2)/7 \end{array}$$

$$\mu_1(A) \approx 0.15240, \quad \mu_2(A) \approx 0.41009$$

- The 2-stage Radau IIA IRK method has $c_1 = 1/3 = 0.\overline{3}$

2-stage IRK methods and dense output (cont.)

- A 2-stage IRK method enables the construction of an error estimator of order 1 for stepsize selection; error estimation for the implicit Euler method is more difficult!
- We use the IRK framework for approximately solving the gradient system, but not at the cost of solving accurately large systems of equations at each step. Only one modified Newton iteration per step can be made which requires 2 evaluations of ∇f that can be done in parallel
- Initial guesses for $Y_1^{(0)}$ and $Y_2^{(0)}$ can be obtained by using a dense output approximation from the previous step or from a starting algorithm

Curve search

- The points $Y_1^{(1)}$ and $Y_2^{(1)}$ are directly available and are of great interest as trial points. The monotonicity conditions

$$f(x_k) > f(Y_1^{(1)}) > f(Y_2^{(1)})$$

  can be tested
- The dense output curve $x(\alpha)$ can be considered for any $\alpha \geq 0$ not just for $\alpha \in [0, 1]$
- We can do a curve search instead of a line search still based on the same Wolfe or Goldstein and Price conditions of sufficient decrease with $m(\alpha) := f(x(\alpha))$.

Quasi-Newton type formulas for 2-stage IRK methods

For the 2 modified Jacobians

$$M_{1,k+1} \approx I + \mu_1 h_{k+1} \nabla^2 f(x_k), \quad M_{2,k+1} \approx I + \mu_2 h_{k+1} \nabla^2 f(x_k)$$

at least 2 quasi-Newton type conditions per step are possible, not just 1, based on

$$M_{1,k+1} s_{1k} = s_{1k} + \mu_1 h_{k+1} y_{1k}, \quad M_{1,k+1} s_{2k} = s_{2k} + \mu_1 h_{k+1} y_{2k}$$
$$M_{2,k+1} s_{1k} = s_{1k} + \mu_2 h_{k+1} y_{1k}, \quad M_{2,k+1} s_{2k} = s_{2k} + \mu_2 h_{k+1} y_{2k}$$

where

$$s_{1k} := Y_1^{(0)} - x_k, \qquad y_{1k} := \nabla f(Y_1^{(0)}) - \nabla f(x_k)$$
$$s_{2k} := Y_2^{(0)} - Y_1^{(0)}, \qquad y_{2k} := \nabla f(Y_2^{(0)}) - \nabla f(Y_1^{(0)})$$

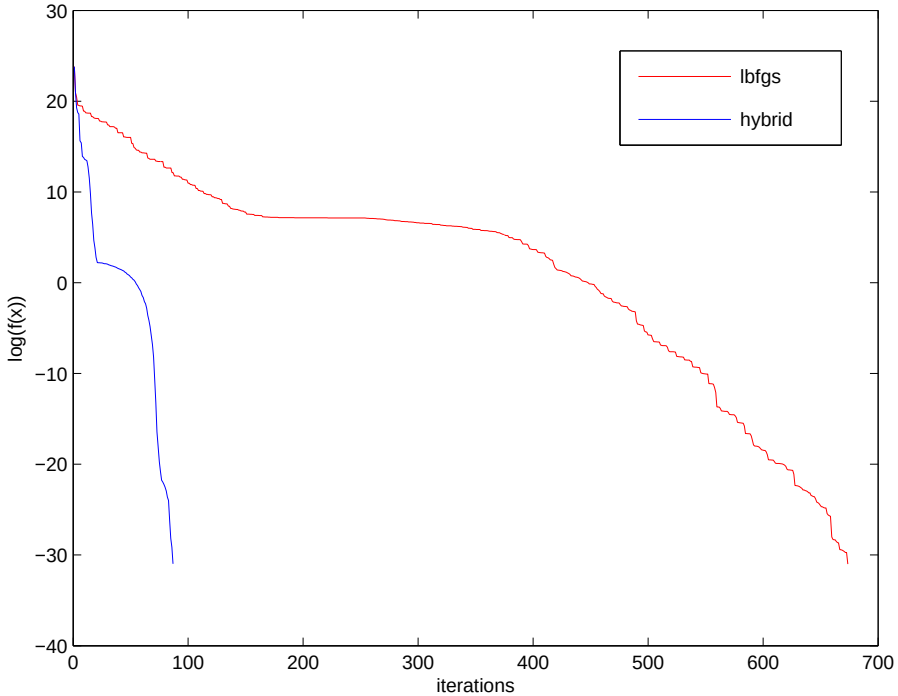
provided that

$$s_{1k}^T y_{1k} > 0, \qquad s_{2k}^T y_{2k} > 0.$$

Numerical results for the Rosenbrock problem

$$f(x) := \sum_{i=1}^{n-1} \left((1 - x_i)^2 + 100(x_{i+1} - x_i^2)^2 \right), \qquad n = 10$$

$H_0 = I$ and
$m = 12$ for limited memory BFGS (L-BFGS)



Fnal remarks and future work

- Search directions can be justified by a gradient ODE system, not by a quadratic model unless the gradient is small
- New quasi-Newton type conditions are promising
- Order two IRK type methods may be more efficient
- Good predictors for $Y_i^{(0)}$ are important
- Curve search strategy is also important (avoid gradients?)
- Preconditioned 2-stage Radau IIA IRK method?
- Extension to nonlinear optimization problems with equality contraints $\longrightarrow$ Numerical DAE methods.