Why adjoint based least squares solving ought to be optimal

Andreas Griewank

Department of Mathematics, Humboldt-Universität zu Berlin, Germany
School of Information Sciences, Yachaytech, Ibarra, Ecuador

September 2, 2015

Numerical Methods for Large-Scale Nonlinear Problems and their Applications
ICERM, Brown University, Providence, RI

with thanks to
Andrea Walther(PAB) and Sebastian Schlenkrich(TUD)
Sandra Schneider(HUB) and Claudia Tutsch(CLU)

Setting

Problem

$$\min \varphi(x) \equiv \tfrac{1}{2}\|F(x)\|_2^2 \quad \text{for} \quad F : \mathbb{R}^n \mapsto \mathbb{R}^m \quad \text{with} \quad n \leq m$$

First order optimality condition (necessary)

$$0 = \nabla\varphi(x_*) \equiv F(x_*)^\top F'(x_*) \in \mathbb{R}^n$$

Second order optimality condition (sufficient)

$$1 > \kappa_* \equiv \|R_*^{-\top} \sum_{i=1}^{m} F_i(x) \nabla^2 F_i(x) R_*^{-1}\|_2 \quad \text{with} \quad F'(x_*) = Q_* \mathbb{R}_*$$

Derivative availability and cost

$$\frac{\mathrm{OPS}\{\dot{y} \equiv F'(x)\dot{x}\}}{\mathrm{OPS}\{y \equiv F(x)\}} \leq 3 \quad , \quad 4 \geq \frac{\mathrm{OPS}\{\bar{x}^\top \equiv \bar{y}^\top F'(x)\}}{\mathrm{OPS}\{y \equiv F(x)\}}$$

Gauss Adjoint Broyden Method

Tangent conditions for $B \approx F'$

$$B_+ s = y \equiv F'(x_+)\, s \in \mathbb{R}^m \quad \text{and} \quad B_+^\top \sigma = F'(x_+)^\top \sigma \in \mathbb{R}^n$$

Transposed Broyden Update

$$B_+ = B + \frac{\sigma\sigma^\top}{\sigma^\top\sigma}(F'(x_+) - B) \quad \textit{for} \quad \sigma = y \quad \text{and} \quad \sigma = r \equiv y - Bs$$

yields rank-two update, which can be implemented in $O(mn)$ operations.

Resulting Properties

Frobenius norm change minimality, domain transformation invariance, and heredity on affine systems $F(x) = Ax - b$.

Quasi-Gauss-Newton Iteration

$$x_+ = x - \alpha(B^\top B)^{-1}\nabla\varphi(x) \quad \text{with} \quad \alpha \quad \text{by Andersen(m=1)}$$

Provable Properties

Global convergence

$$0 = \inf_k \|\nabla\varphi(x_k)\| \quad \Leftarrow \quad x_0 \in \{\varphi(x) \le c\} \text{ compact and } \operatorname{rank}(F'(x)) = n$$

Asymptotic R-rate in overdetermined case $(m > n)$

$$0 = \inf_k \|x_k - x_*\| \;\Rightarrow\; \limsup_{k\to\infty} \|x_k - x_*\|^{\frac{1}{k}} \;\le\; \kappa_* \;<\; 1$$

Asymptotic order in consistent case $(m = n)$

$$\liminf_{k\to\infty} |\log(\|x_k - x_*\|)|^{\frac{1}{k}} \;\ge\; \rho_n \;\approx\; 1 + \frac{\log(n)}{n} \quad \text{with} \quad 1 = \rho_n^{n+1} - \rho_n^n$$

On affine problems

Finite termination in $\le n$ steps, (á la GMRES when $m = n$ and $B_0 = I$.)

Piecewise linearizations of nonsmooth equations and their numerical solution

Andreas Griewank[1,2]
Tom Streubel[1]
Richard Hasenfelder[1]

1) Department of Mathematics, Humboldt University at Berlin
2) School of Information Sciences, Yachaytech, Ibarra, Eucador

Numerical Methods for Large-Scale Nonlinear Problems and Their Applications
ICERM at Brown University

$PC^{1,1}$ evaluation procedures

$$F(x,y,z) = \begin{bmatrix} \sin(|x+y|) + x \\ 2\big|3\cdot|z| - x\big| \end{bmatrix}$$

function expression

v_0	$=$	$x+y$
v_1	$=$	$\lvert v_0\rvert$
v_2	$=$	$\sin(v_1)$
f_0	$=$	$v_2 + x$
v_3	$=$	$\lvert z\rvert$
v_4	$=$	$3v_3$
v_5	$=$	$v_4 - x$
v_6	$=$	$\lvert v_5\rvert$
f_1	$=$	$2v_6$
$F(x,y,z)$	$=$	$(f_0 \quad f_1)$

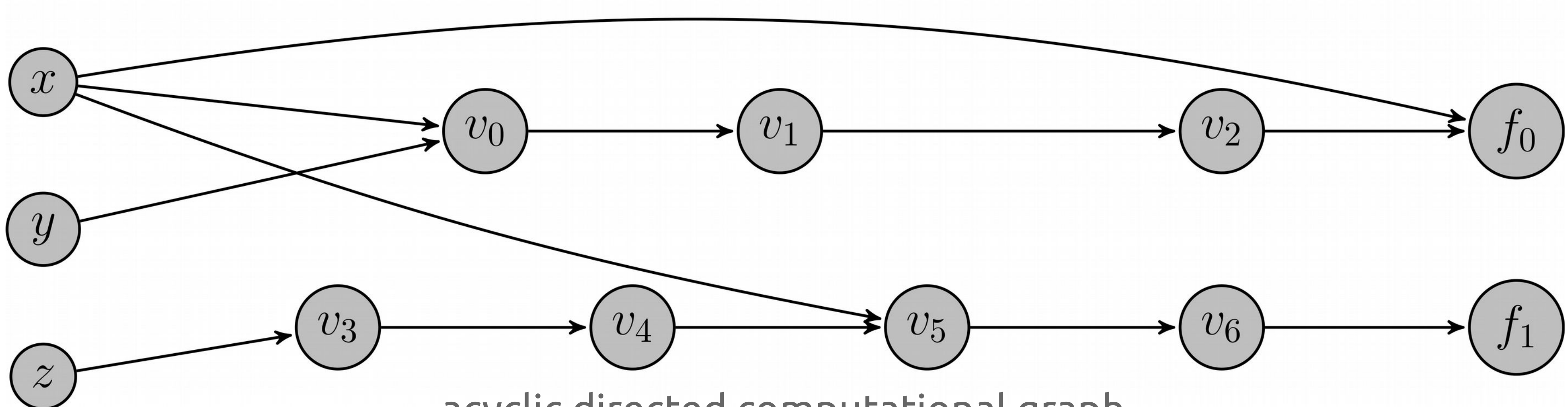
single assignment code

assume $F:\ \mathbb{R}^n \to \mathbb{R}^m$ to be a chain of $C^{1,1}$ functions from some Library Φ and the absolute value function $\mathrm{abs}(\cdot)$

the expression can be recast as single assignment code

$$v_i = \varphi_i(v_j)_{j \prec i} \quad \text{for} \quad i = 1,\ldots,l$$

here $\prec$ is a dependence relation generating a partial order



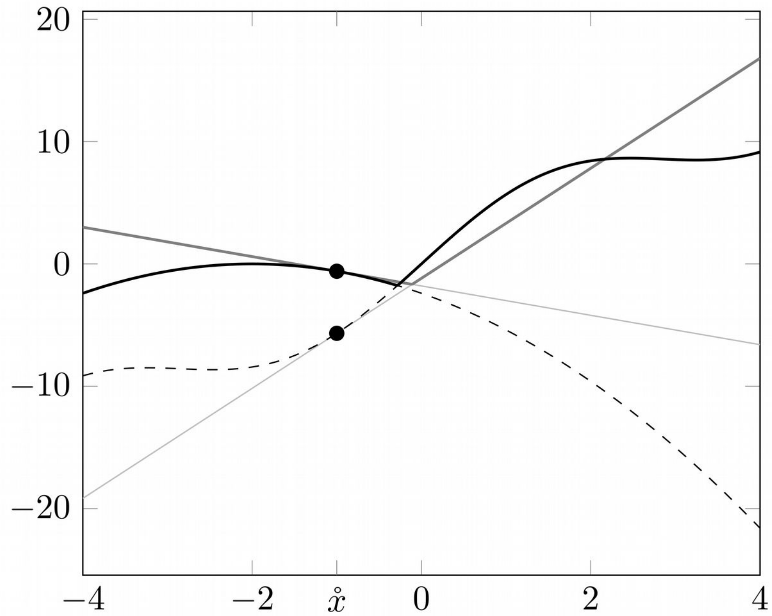
acyclic directed computational graph

Algorithmic Piecewise Linearization - I

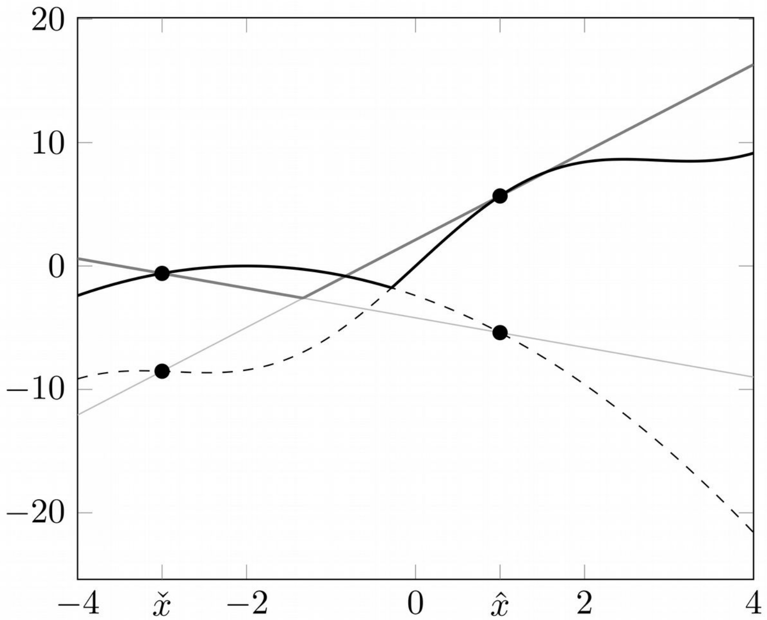
basic idea

- propagate piecewise linear rather than linear approximations
- therefor replace differentiable elementals by its **linear** tangent/secant model
- as well as absolute value function by itself

tangent mode



secant mode



Algorithmic Piecewise Linearization - II

either choose

- one reference point $\mathring{x} \in \mathrm{dom}(F)$ $\quad \mathring{v} = v(\mathring{x})$ (tangent mode)
- two reference points $\check{x}, \hat{x} \in \mathrm{dom}(F)$ $\quad \mathring{v} = \frac{1}{2}(v(\check{x}) + v(\hat{x}))$ (secant mode)

For any single assignment evaluate an increment

$$
\begin{array}{rcll}
\Delta v_i & = & \alpha \Delta v_j \pm \beta \Delta v_k & \text{for } v_i = \alpha v_j \pm \beta v_k \\
\Delta v_i & = & \mathring{v}_j \cdot \Delta v_k + \mathring{v}_k \cdot \Delta v_j & \text{for } v_i = v_j \cdot v_k \\
\Delta v_i & = & c_\varphi \cdot \Delta v_j & \text{for } v_i = \varphi(v_j), \text{ where } \varphi \neq \mathrm{abs} \\
\Delta v_i & = & |\mathring{v}_j + \Delta v_j| - \mathring{v}_i & \text{for } v_i = |v_j|
\end{array}
$$

$$
c_\varphi = \begin{cases} \varphi'(\mathring{v}_j) & \text{tangent mode or } \check{x} = \hat{x} \\ \frac{\varphi(\check{v}_j) - \varphi(\hat{v}_j)}{\check{v}_j - \hat{v}_j} & \text{secant mode and } \check{x} \neq \hat{x} \end{cases}
$$

These increments depends on reference point(s) and preceding increments. So we write

- $\Delta v_i(\mathring{x}; \Delta x) \quad \equiv \Delta v_i$ (tangent mode)
- $\Delta v_i(\check{x}, \hat{x}; \Delta x) \equiv \Delta v_i$ (secant mode)

Algorithmic Piecewise Linearization - III

$\Delta F(\mathring{x};\cdot)$ is called **tangent** piecewise linear model of F at $\mathring{x}$ and satisfies

$$F(x) = F(\mathring{x}) + \Delta F(\mathring{x}; x - \mathring{x}) + \mathcal{O}(\|x - \mathring{x}\|^2)$$

Inhomogeneous tangent model $\Diamond_{\mathring{x}} F(x) = F(\mathring{x}) + \Delta F(\mathring{x}; x - \mathring{x})$

$\Delta F(\check{x}, \hat{x};\cdot)$ is called **secant** piecewise linear model of F at $\check{x}, \hat{x}$ if

$$F(x) = \mathring{F} + \Delta F(\check{x}, \hat{x}; x - \mathring{x}) + \mathcal{O}(\|x - \check{x}\| \cdot \|x - \hat{x}\|)$$

where

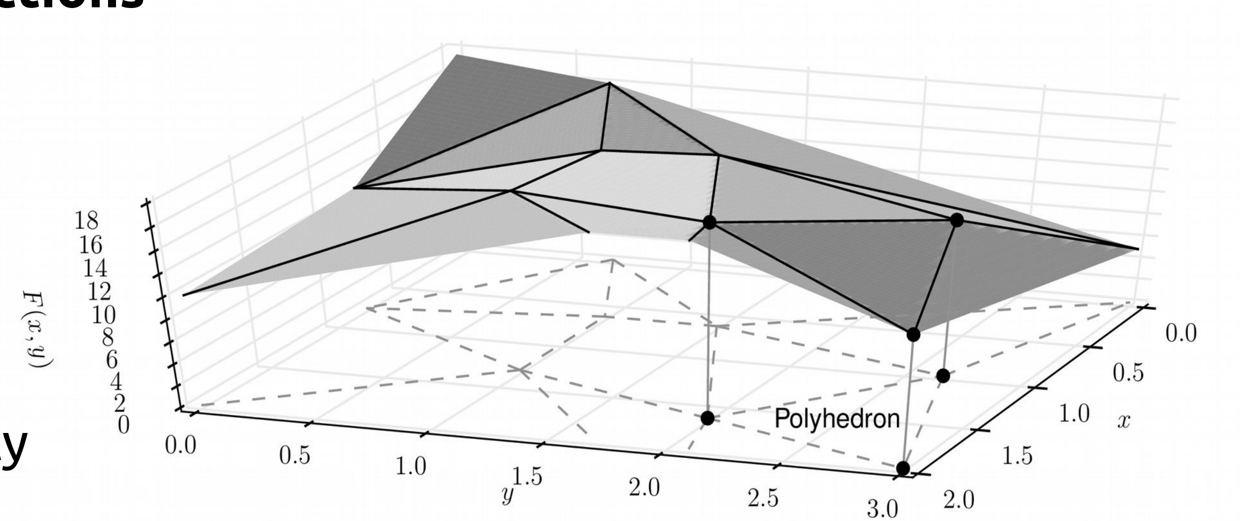
$$\mathring{F} = \frac{1}{2}[F(\check{x}) + F(\hat{x})] \qquad \mathring{x} = \frac{1}{2}(\check{x} + \hat{x})$$

Inhomogeneous secant model $\Diamond_{\check{x}}^{\hat{x}} F(x) = \mathring{F} + \Delta F(\check{x}, \hat{x}; x - \mathring{x})$

Algorithmic Piecewise Linearization - IV

- Algorithmic piecewise linearization can be performed by slight modifications of common AD-Tools (e.g. Adol-C) → see autodiff.org

- **general properties of PL functions**
  - Lipschitz continuous
  - consists of linear and absolute value functions
  - correspond to a polyhedral subdivision
  - a polyhedron with non empty interior is called essential
  - Implication chain (by S. Scholtes):



$$\text{injectivity} \quad \Longrightarrow \quad \text{openness} \quad \Longrightarrow \quad \text{surjectivity}$$

  - openness is equivalent to coherent orientation:

$$\text{w.l.o.g. } \forall \text{ essential linear selection function } d + Jx : \ \det(J) > 0$$

Approximation properties of PL models

$$\|\Diamond_{\mathring{x}} F(x) - \Diamond_{\mathring{y}} F(x)\| \leq \qquad L\left[\|\mathring{x} - \mathring{y}\| \max(\|x - \mathring{x}\|, \|x - \mathring{y}\|)\right]$$

For some (algorithmically computable) Lipschitz constant L

$$\Diamond_{\mathring{x}} F(x) \equiv F(\mathring{x}) + \Delta F(\mathring{x}; x - \mathring{x})$$

$\Uparrow$ simplifies to, if $\check{x} = \hat{x}$

$$\begin{aligned}\|\Diamond_{\check{x}}^{\hat{x}} F(x) - \Diamond_{\check{y}}^{\hat{y}} F(x)\| \leq L \max \Big[& \|\check{x} - \check{y}\| \max(\|x - \hat{x}\|, \|x - \hat{y}\|), \\ & \|\hat{x} - \hat{y}\| \max(\|x - \check{x}\|, \|x - \check{y}\|) \Big]\end{aligned}$$

For some (algorithmically computable) Lipschitz constant L

$$\Diamond_{\check{x}}^{\hat{x}} F(x) \equiv \mathring{F} + \Delta F(\check{x}, \hat{x}; x - \mathring{x})$$

Implications:

- $\check{x} = a = \check{y} \qquad \Longrightarrow \qquad \Diamond_a^{\hat{x}} F(a) = \Diamond_a^{\hat{y}} F(a) = F(a)$
- $\check{x} = a = \hat{x} \qquad \Longrightarrow \qquad \|F(a) - \Diamond_{\check{y}}^{\hat{y}} F(a)\| \leq 2L\|a - \check{y}\|\|a - \hat{y}\|$

Newton via successive piecewise linearization I

Tangent mode

Let x^* be a root of a $PC^{1,1}$ algorithm F.
If $\Diamond_{\mathring{x}}^{-1} F(0) \cap B_\rho(x^*) \neq \emptyset$ for a fixed radius $\rho > 0$ then

$$x_{j+1} \in \arg\min\{\|x - x_j\| \mid x \in \Diamond_{x_j}^{-1} F(0)\}$$

is called **feasible tangent mode iteration**.

Secant mode

if $\Diamond_{\{\check{x},\hat{x}\}}^{-1} F(0) \cap B_\rho(x^*) \neq \emptyset$ again for a fixed radius then

$$x_{j+1} \in \arg\min\{\|x - \mathring{x}_j\| \mid x \in \Diamond_{\{x_{j-1},x_j\}}^{-1} F(0)\}$$

is called **feasible secant mode iteration**, where $\mathring{x}_j = \frac{1}{2}(x_j + x_{j-1})$ and $\Diamond^{-1}$ set-valued inverses

Quadratic or golden ration convergence rate

Tangent mode

assume feasibility of tangent mode iteration as well as

$$\exists c > 0 \, \forall \mathring{x} \in B_\rho(x^*) : \; \|\mathring{x} - x^*\| \leq c \|\lozenge_{x^*} F(\mathring{x})\|$$ (local strong metric regularity)

satisfied, the tangent mode iteration converges quadratically (rate $p = 2$) to x^*

Secant mode

assume feasibility of secant mode iteration as well as

$$\exists c > 0 \, \forall \mathring{x} \in B_\rho(x^*) : \; \|\mathring{x} - x^*\| \leq c \|\lozenge_{x^*} F(\mathring{x})\|$$ (local strong metric regularity)

satisfied, then the secant mode iteration converges with
Golden ratio rate $p = \frac{1}{2}(1 + \sqrt{5})$ to the root x^*

Newton via successive piecewise linearization II

strong metric regularity in x^* i.e.

$$\exists c > 0 \, \forall \mathring{x} \in B_\rho(x^*) : \; \|\mathring{x} - x^*\| \leq c \|\Diamond_{x^*} F(\mathring{x})\|$$

is implied by openness of the restriction of $\Diamond_{x^*} F(\cdot)$ to $B_\rho(x^*)$

So far we know

feasibility of both iterations

$$x_{j+1} \in \arg\min\{\|x - x_j\| \mid x \in \Diamond_{x_j}^{-1} F(0)\}$$

$$x_{j+1} \in \arg\min\{\|x - \mathring{x}_j\| \mid x \in \Diamond_{\{x_{j-1}, x_j\}}^{-1} F(0)\}$$

is implied by injectivity of $\Diamond_{x^*} F(\cdot)$

Open Newton Conjecture:
feasibility is already guaranteed in case of openness of $\Diamond_{x^*} F(\cdot)$

Generalized Trapezoidal Rule

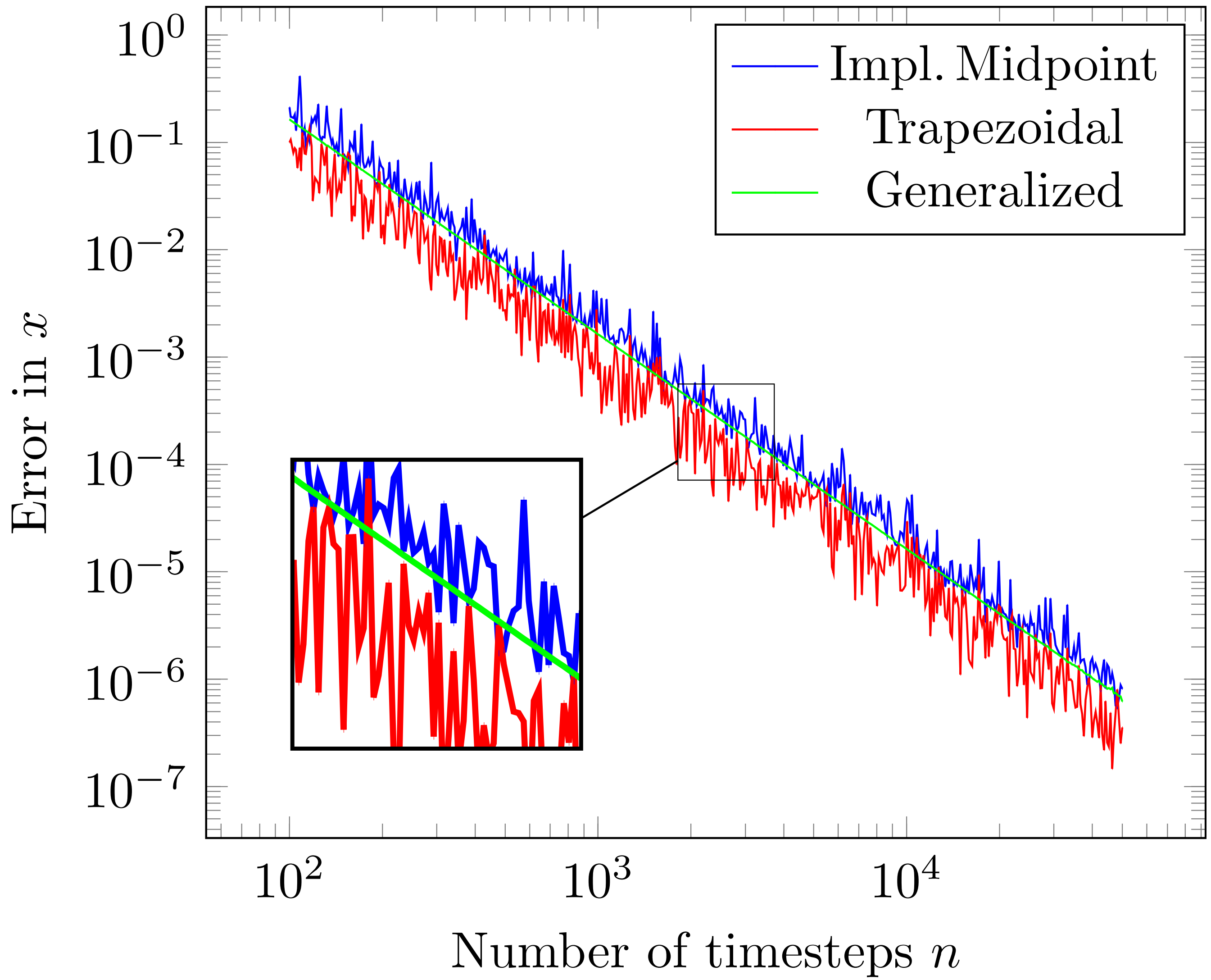
Idea for construction: We integrate the ODE $\dot{x}(t) = F(x(t))$:

$$\hat{x} - \check{x} := x(h) - x(0) = \int_0^h F(x(t))\mathrm{d}t = h \int_{-\frac{1}{2}}^{\frac{1}{2}} \underbrace{F\left(x\left(\tfrac{h}{2} + \tau h\right)\right)}_{F(x) \approx \mathring{\mathrm{F}} + \Delta\mathrm{F}(\hat{x}, \check{x}; t(\hat{x} - \check{x}))} \mathrm{d}\tau$$

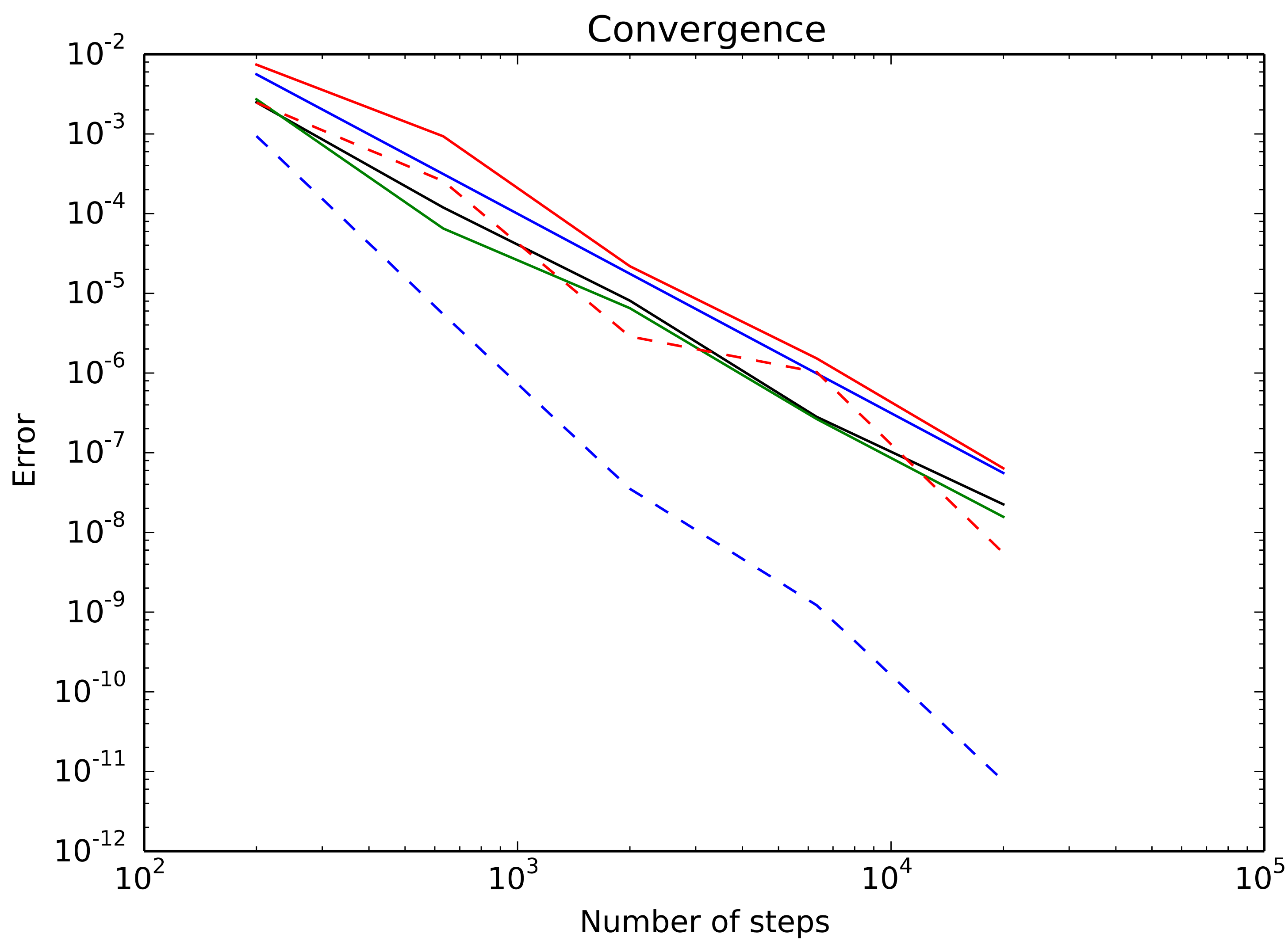
and get

$$\hat{x} - \check{x} = h \int_{-\frac{1}{2}}^{\frac{1}{2}} \mathring{\mathrm{F}} + \Delta\mathrm{F}(\hat{x}, \check{x}; t(\hat{x} - \check{x}))\mathrm{d}t \quad \text{with} \quad \mathring{\mathrm{F}} = \tfrac{1}{2}[\mathrm{F}(\hat{x}) + \mathrm{F}(\check{x})]$$

Convergence

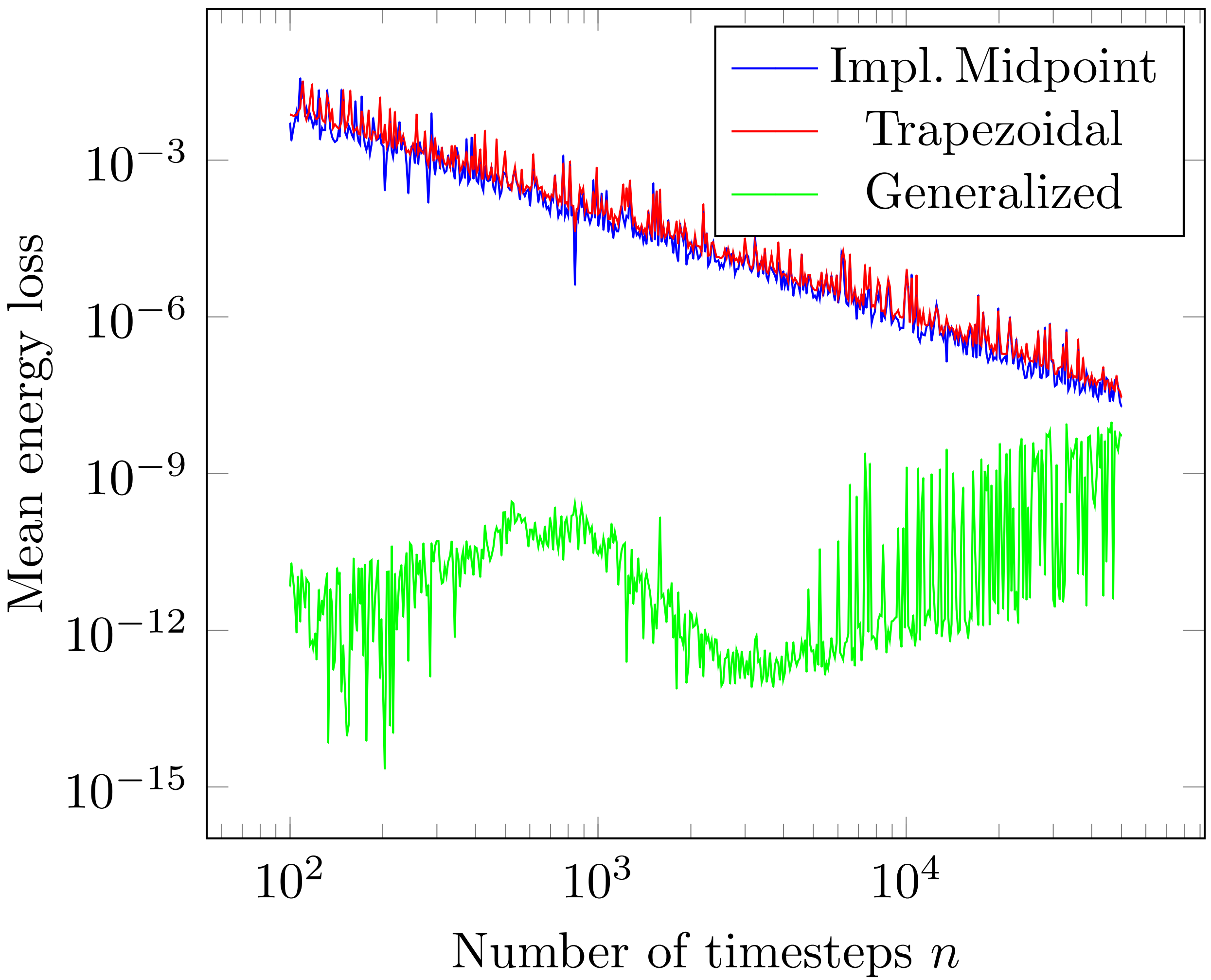


Convergence with Romberg extrapolation

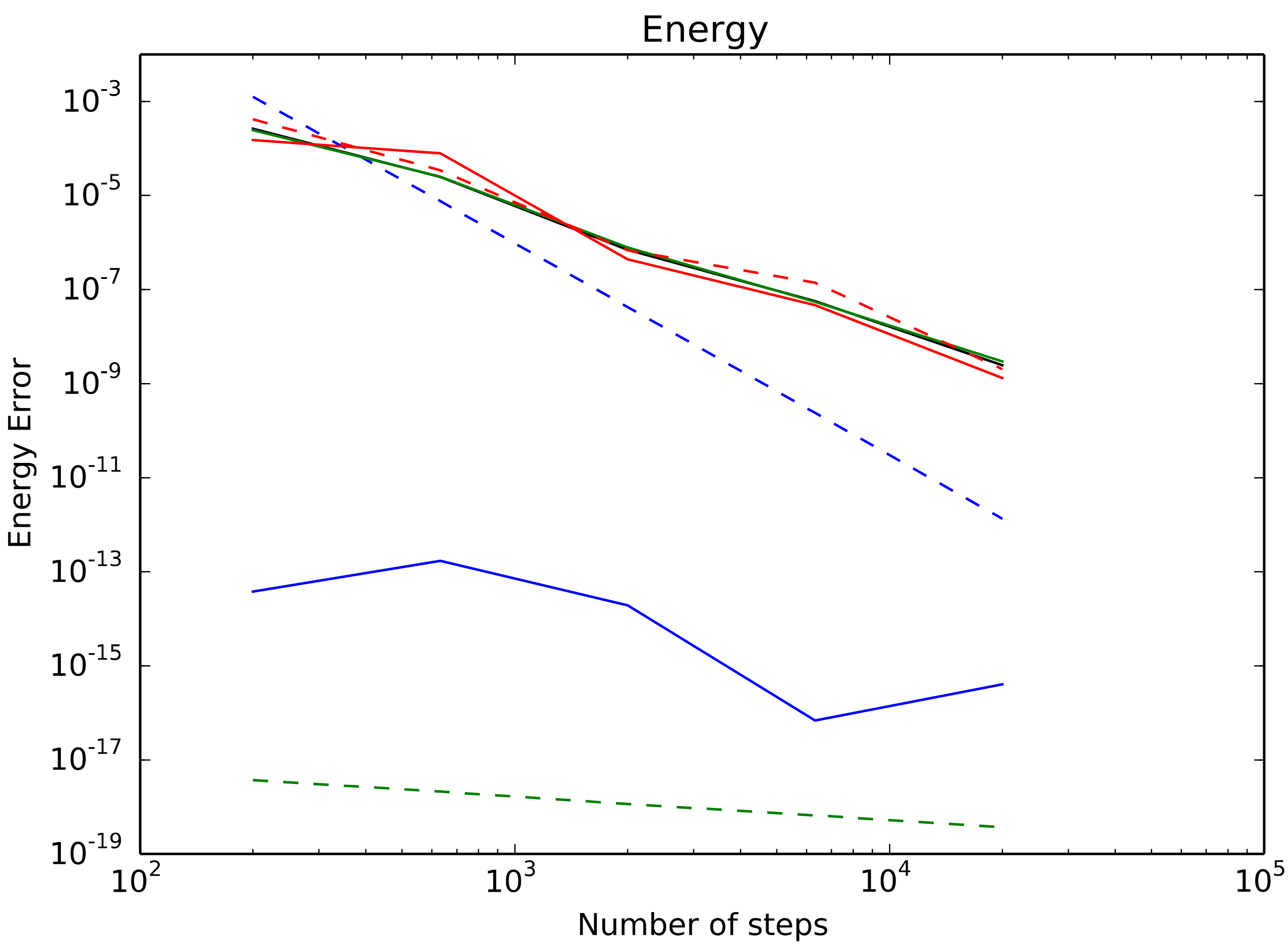


Energy of the System

Summed energy loss: $\left\| V(x) + \frac{1}{2}\dot{x}^2 - \frac{1}{2} \right\|$



Energy with Romberg extrapolation

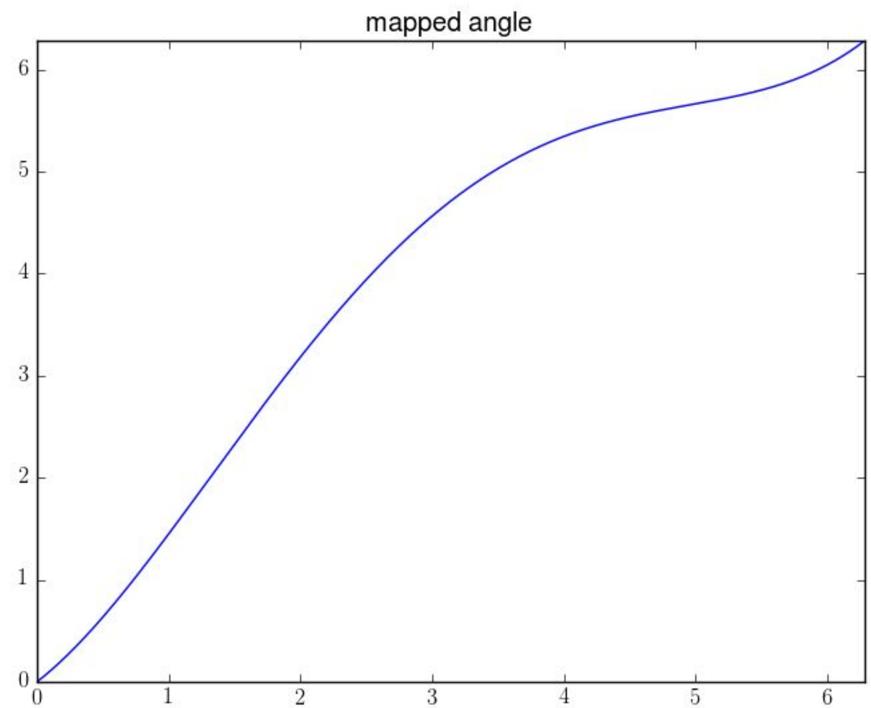


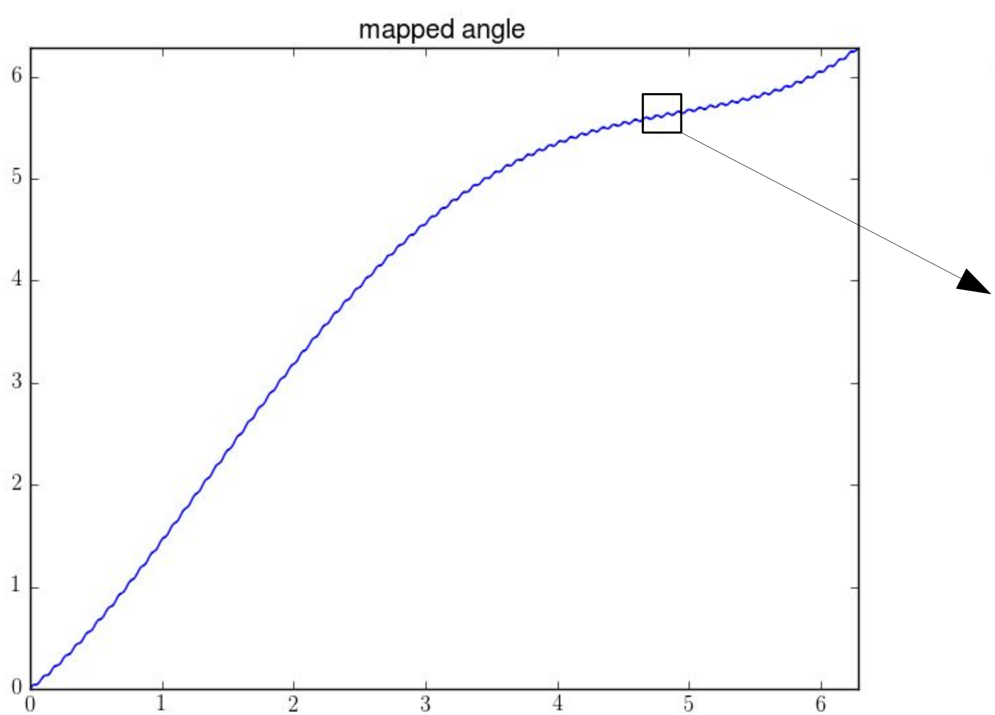
A 2D oscillating test example

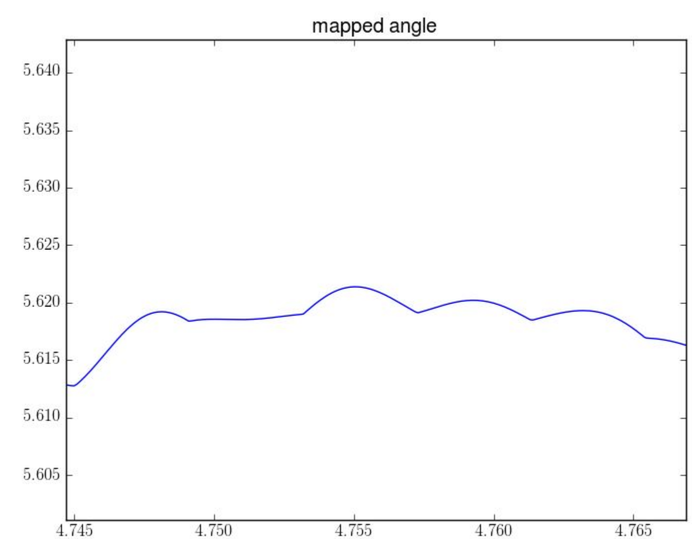
- For any vector $x \in \mathbb{R}^2$ take its angle from polar coordinate representation and map it by some differentiable (right picture) or $PC^{1,1}$ bijective function (picture below)

$$\psi :\ [0, 2\pi[\rightarrow [0, 2\pi[$$

- thereby preserve its euclidean norm

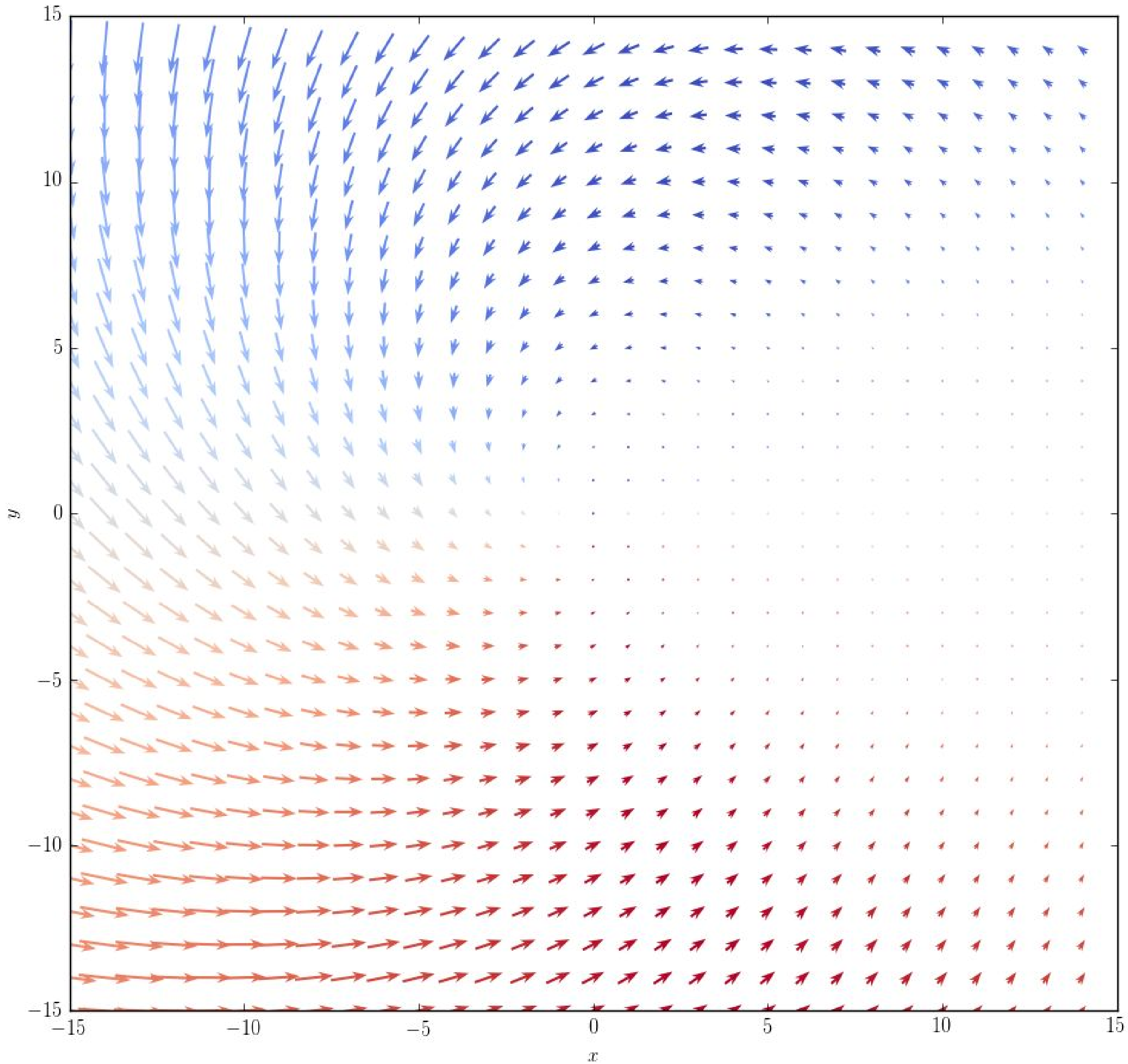




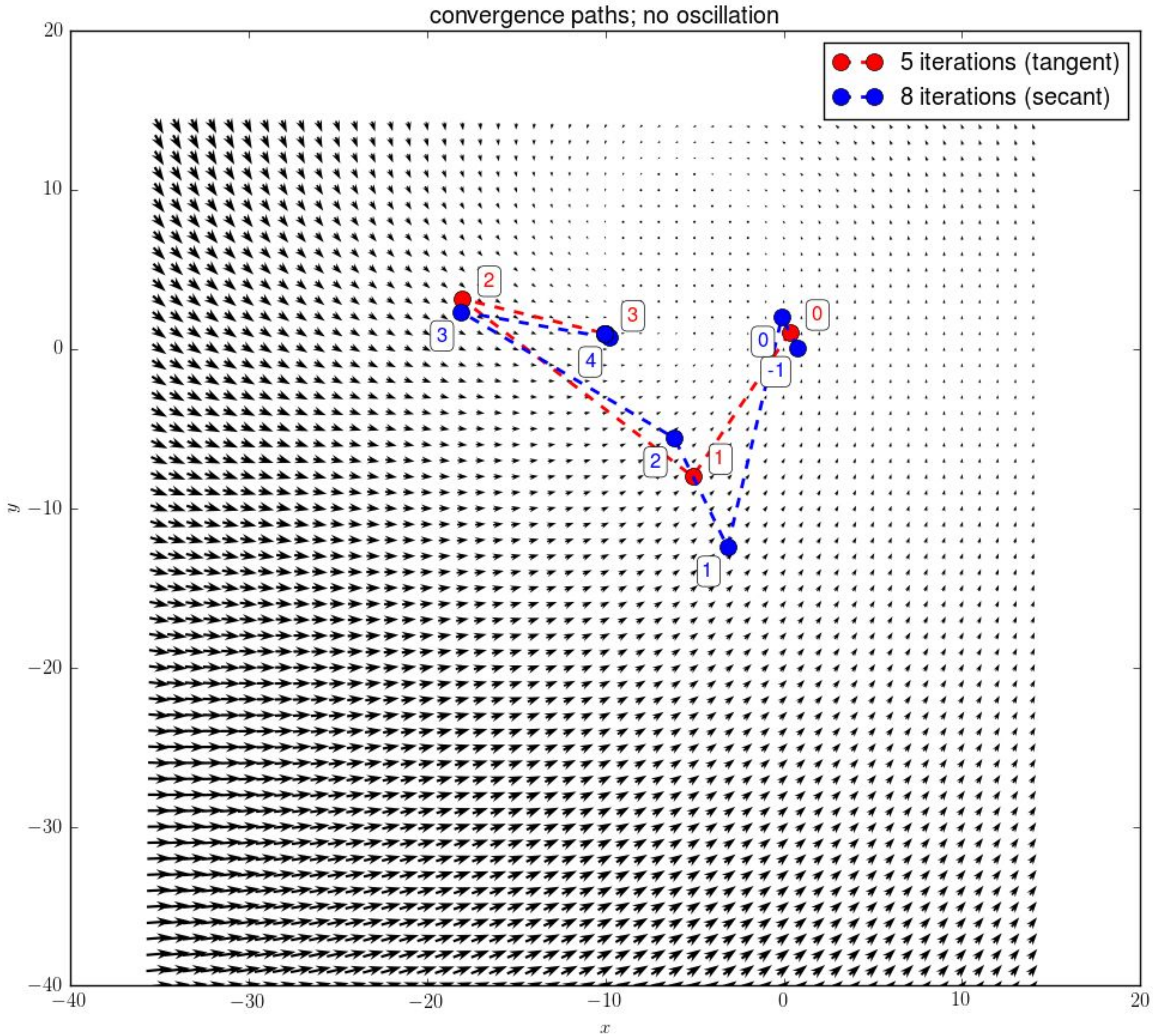


A 2D oscillating test example – homogen. part

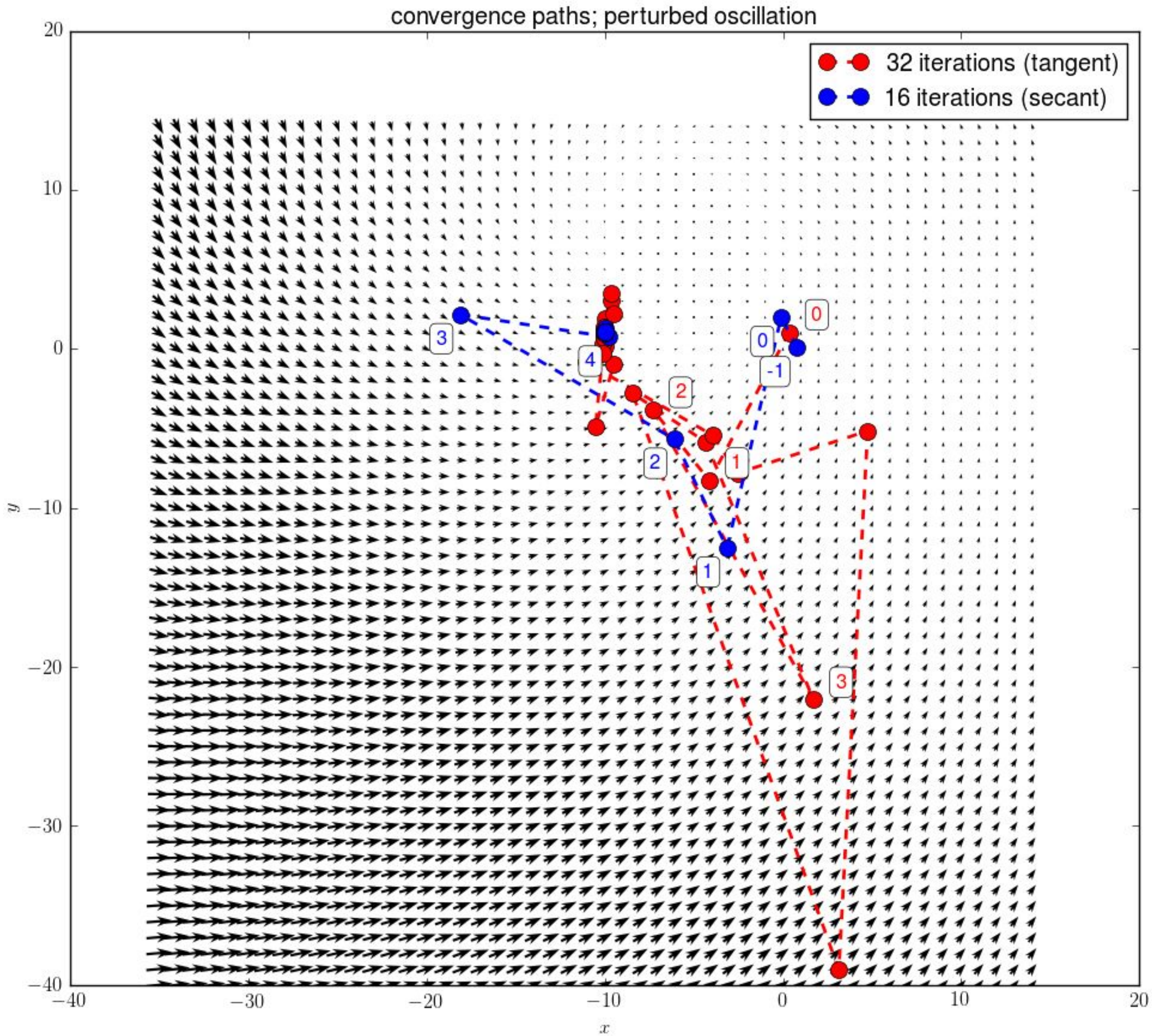
- upper half of $\mathbb{R}^2$ is stretched (blue)
- lower half is compressed (red)
- is bijective and Lipschitz continuous
- the line $\{x \geq 0\}$ is kept fixed
- almost everywhere differentiable, but not at origin



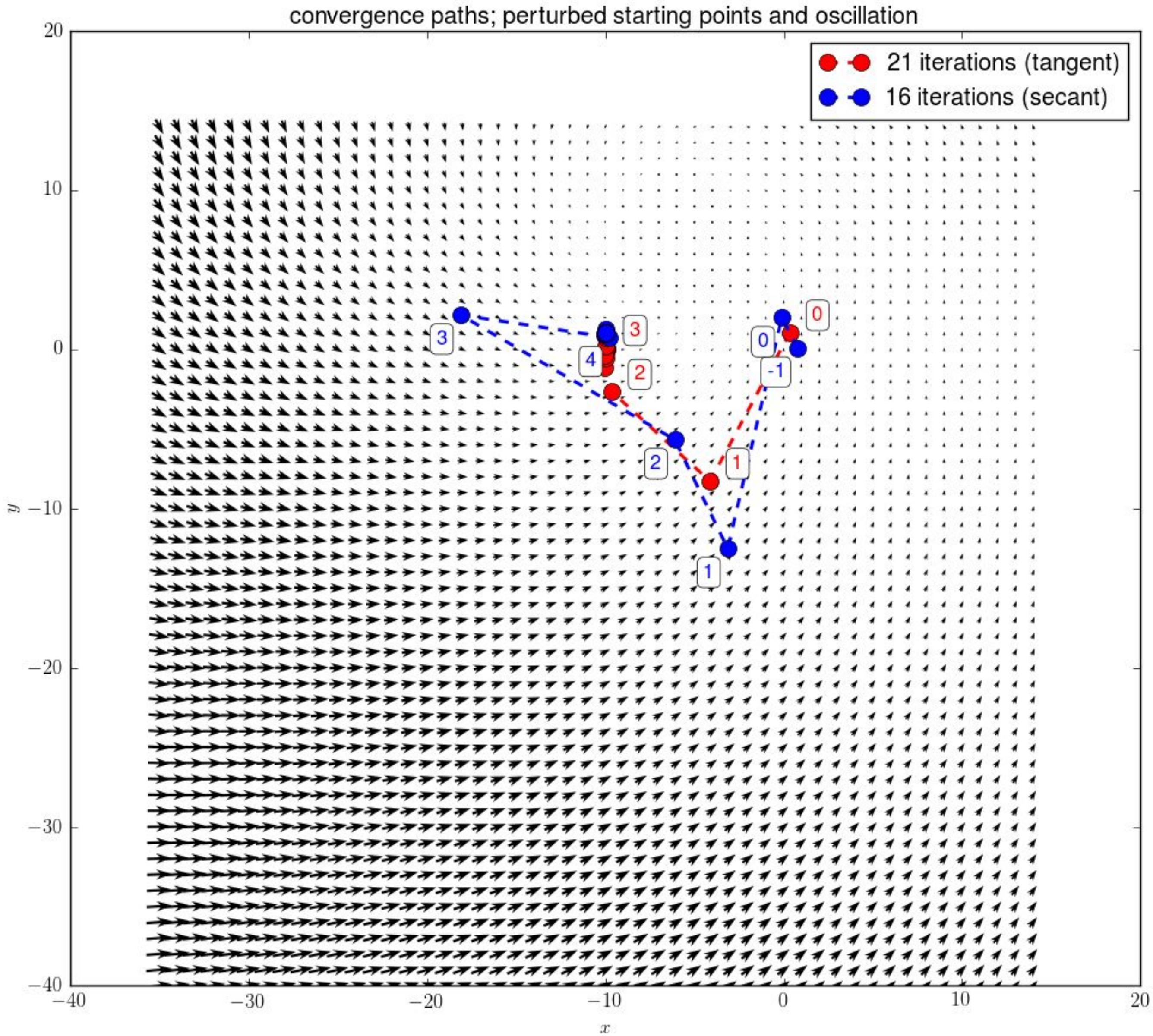
A 2D oscillating test example



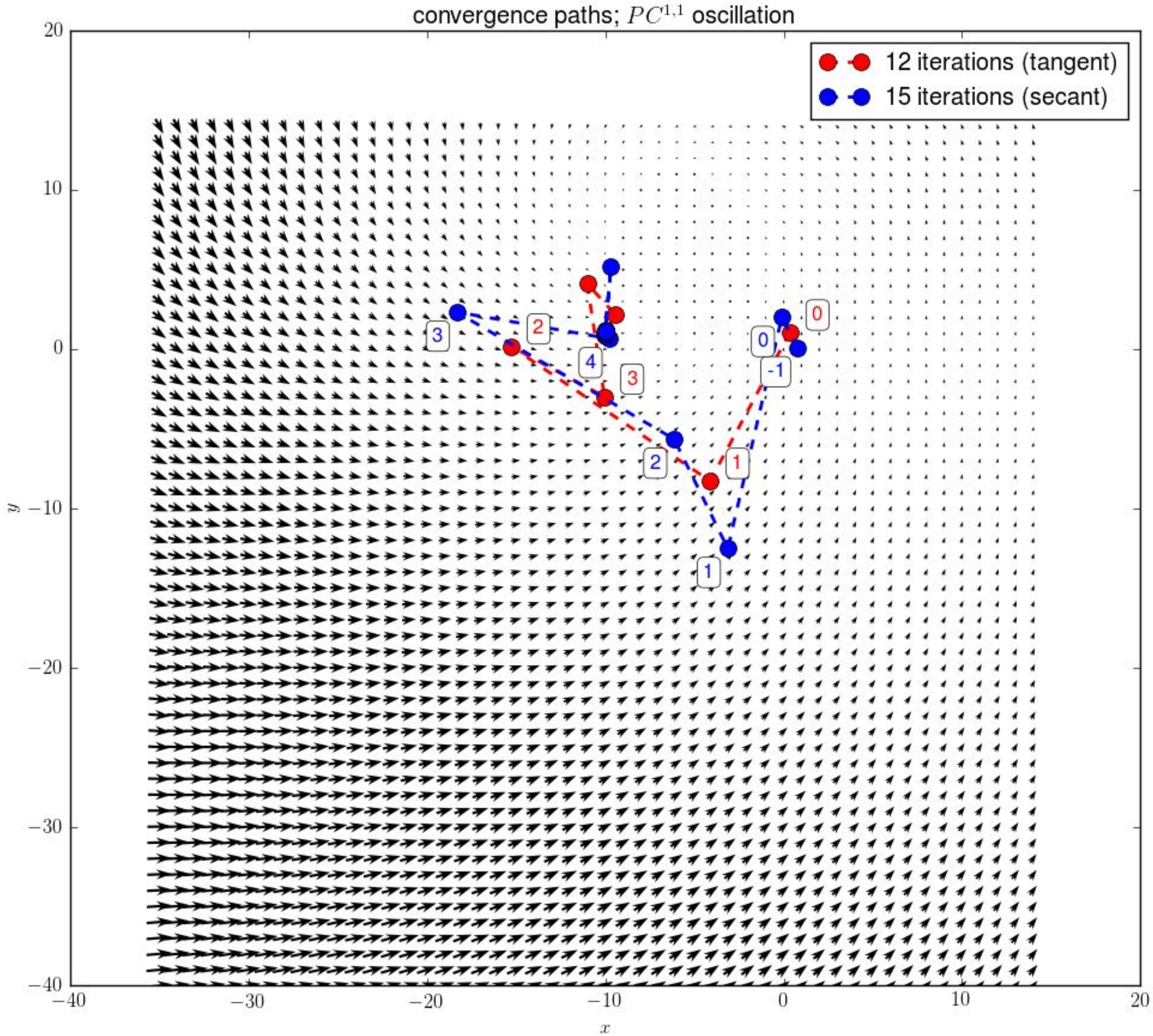
A 2D oscillating test example



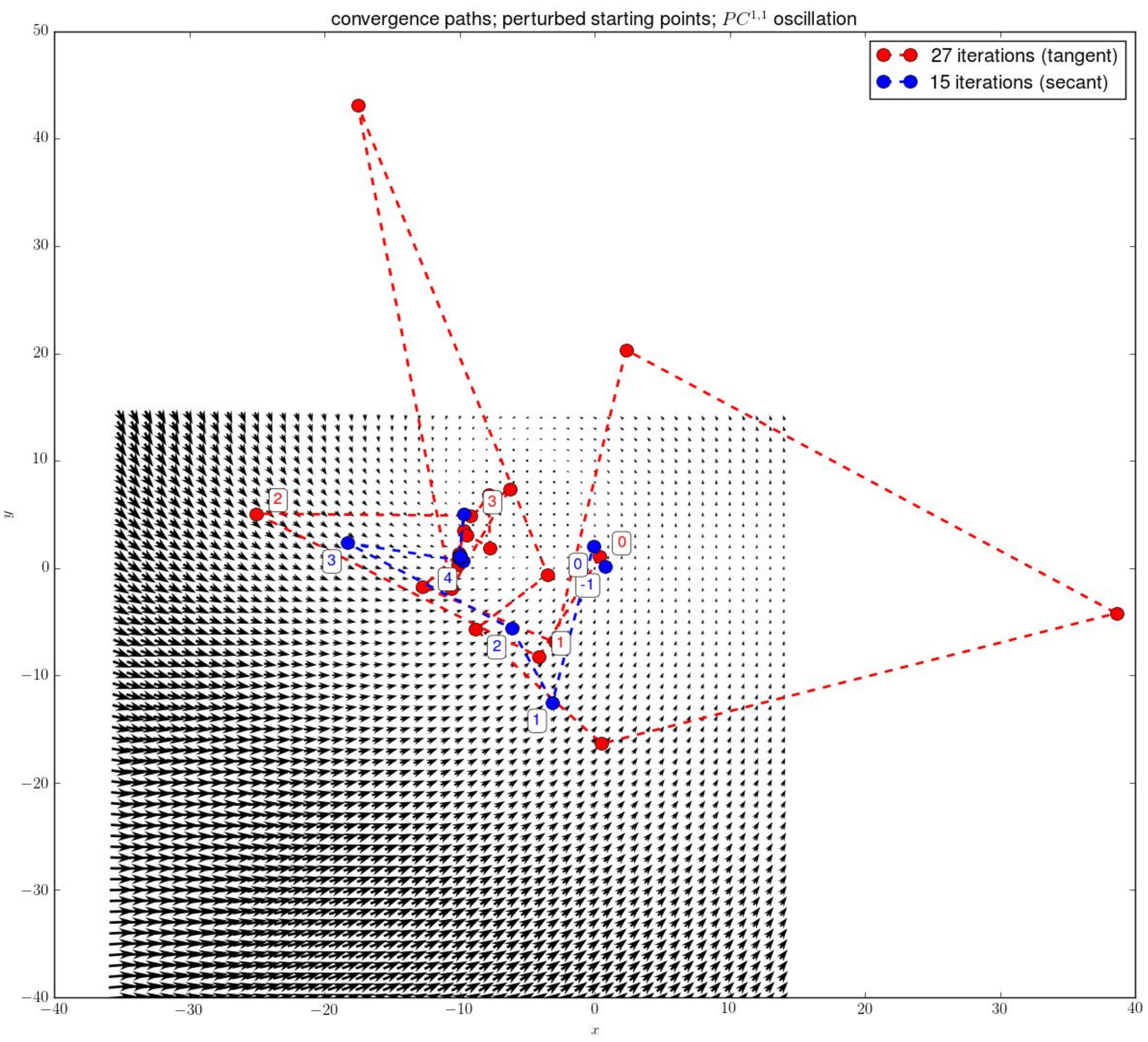
A 2D oscillating test example



A 2D oscillating test example



A 2D oscillating test example



Piecewise linear subproblem I

Definition: Abs-normal Form $F : \mathbb{R}^n \to \mathbb{R}^n$ PL

$$\begin{bmatrix} z \\ F(x) \end{bmatrix} = \begin{bmatrix} c \\ b \end{bmatrix} + \begin{bmatrix} Z & L \\ J & Y \end{bmatrix} \cdot \begin{bmatrix} x \\ |z| \end{bmatrix}$$

$$Z \in \mathbb{R}^{s\times n},\ L \in \mathbb{R}^{s\times s},\ J \in \mathbb{R}^{n\times n},\ Y \in \mathbb{R}^{n\times s} \qquad c \in \mathbb{R}^s,\ b \in \mathbb{R}^n$$

- **Any piecewise linear function can be represented this way**
- the matrix $L \in \mathbb{R}^{s\times s}$ is of strict lower triangular form
  thus $z \in \mathbb{R}^s$ can be evaluated explicitly and element wise
- the abs-normal form is numerically stable $\longrightarrow$ use as **data structure**
- the signature of $z \in \mathbb{R}^s$ is defined as follows

$$\sigma(z) \equiv \{\sigma \in \{-1, 0, 1\}^s \mid \forall i : \sigma_i z_i = |z_i|\}$$
$$\Sigma(z) \equiv \{\mathrm{diag}(\sigma) \mid \sigma \in \sigma(z)\} \qquad \longrightarrow \qquad |z| = \Sigma z$$

  each one corresponds to a polyhedron from the polyhedral subdivision of F

- **Task**: search a root $x^* \in \mathbb{R}^n$ such that $F(x^*) \equiv 0$

Piecewise linear subproblem II

- one can simplify the polyhedral structure of a given problem

$$\text{Find } x^* \in \mathbb{R}^n : \quad \begin{bmatrix} z \\ 0 \end{bmatrix} = \begin{bmatrix} c \\ b \end{bmatrix} + \begin{bmatrix} Z & L\Sigma \\ J & Y\Sigma \end{bmatrix} \cdot \begin{bmatrix} x^* \\ z \end{bmatrix}, \qquad \Sigma \in \Sigma(z(x^*))$$

$$Z \in \mathbb{R}^{s\times n},\ L \in \mathbb{R}^{s\times s},\ J \in \mathbb{R}^{n\times n},\ Y \in \mathbb{R}^{n\times s} \qquad c \in \mathbb{R}^s,\ b \in \mathbb{R}^n$$

(we refer this as original piecewise linear problem or short **OPL**)

- evaluate Schur-complement $S \equiv L - ZJ^{-1}Y$ of J and define

$$\text{Find } z^* \in \mathbb{R}^n : \quad H(z^*) = (I - S\Sigma)z^* - \hat{c} = 0, \qquad \Sigma \in \Sigma(z^*)$$

(we refer this as complementary piecewise linear problem or short **CPL**)

- **CPL**'s and **LCP**'s are equivalent formulations via Möbius transformation
- there is a one-to-one solution correspondence between **OPL** and **CPL**

Full step Newton method I

By the one to one solution correspondence search a root of one of the two systems

(OPL)

$$\begin{bmatrix} z \\ F(x) \end{bmatrix} = \begin{bmatrix} c \\ b \end{bmatrix} + \begin{bmatrix} Z & L\Sigma \\ J & Y\Sigma \end{bmatrix} \cdot \begin{bmatrix} x \\ z \end{bmatrix}$$

$$\Sigma \in \Sigma(z(x))$$

$$x_+ = x - J_\sigma^{-1} F(x)$$

where $\sigma \in \sigma(z)$ essential

$$J_\sigma \equiv J + Y\Sigma(I - L\Sigma)^{-1} Z$$

(CPL)

$$H(z) = (I - S\Sigma)z - \hat{c}$$

$$\Sigma \in \Sigma(z)$$

$$S = L - ZJ^{-1}Y$$

$$z_+ = z - J_\sigma^{-1} H(z)$$

where $\sigma \in \sigma(z) \cap \{-1, 1\}^s$

$$J_\sigma \equiv I - S\Sigma$$

- both are generalized Newton methods in the sense of Qi and Sun
- But we seek global rather than local convergence criteria
- Converges from every starting point towards a solution if either

$$\|I - J_{\tilde{\sigma}} J_\sigma^{-1}\| < 1, \qquad \forall \sigma, \tilde{\sigma} \text{ essential}$$

or

$$\|I - J_\sigma^{-1} J_{\tilde{\sigma}}\| < 1, \qquad \forall \sigma, \tilde{\sigma} \text{ essential}$$

is satisfied and the root is unique

- J_σ for a essential signature is always a limiting Jacobian of the underlying PL function

Full step Newton method II

conditions for contractivity

$$\|I - J_{\tilde{\sigma}} J_{\sigma}^{-1}\| < 1, \qquad \forall \sigma, \tilde{\sigma} \text{ essential} \qquad \textbf{or}$$
$$\|I - J_{\sigma}^{-1} J_{\tilde{\sigma}}\| < 1, \qquad \forall \sigma, \tilde{\sigma} \text{ essential}$$

Verify the conditions is NP-hard but one can find sufficient conditions:

- **OPL**: Assume J from the abs-normal Form to be regular then if

$$\hat{\rho} \equiv \|J^{-1}Y\|_p \|Z\|_p < 1 - \|L\|_p$$
$$\bar{\rho} \equiv \frac{2\hat{\rho}}{(1 - \hat{\rho} - \|L\|_p)(1 - \|L\|_p)} < 1$$

  both conditions are satisfied.

- **CPL**: both conditions are satisfied if

$$\|S\|_p < \frac{1}{3} \qquad \text{or} \qquad \rho(|S|) < \frac{1}{2}$$

Restricted Newton method

Under the assumption of coherent orientation (c.o.):

Piecewise-Newton

(OPL)	$x_+ = x - \tau J_\sigma^{-1} F(x), \quad \sigma \in \sigma(z(x))$ essential
(CPL)	$z_+ = z - \tau J_\sigma^{-1} H(z), \quad \sigma \in \sigma(z) \cap \{-1, 1\}^s$

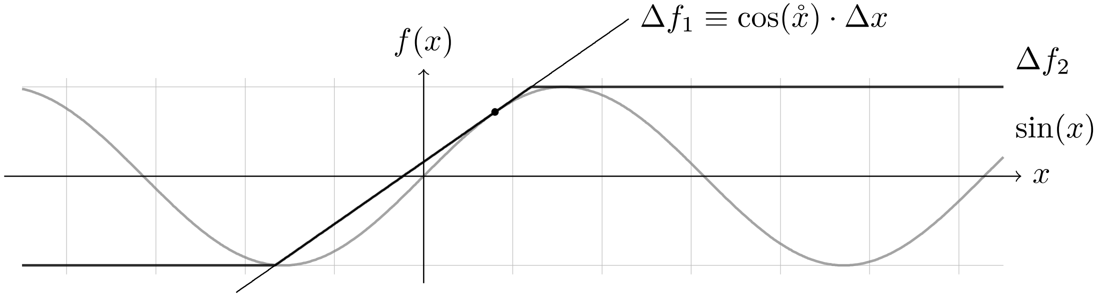
- here $\tau \leq 1$ is called critical multiplier and maximal s.t. the Newton step doesn't leave the closure of the polyhedron corresponding to the chosen essential Signature σ
- the step is shrunk by non smoothness arising on its direction
- the paths
$$[x_0] \equiv \{x \in \mathbb{R}^n \mid F(x) = \lambda F(x_0),\ \lambda \in [0, 1]\}$$
are bifurcation free for almost all starting points $x_0 \in \mathrm{dom}(F)$ and also for the **CPL**
- if the Problem is c.o. then the piecewise Newton converges from everywhere to a root

Outlook

- proof open Newton conjecture
- further develop PL Algebra Package Plan-C (C++)

  → method optimization and comparison
- Branin's modification for PL-Newton on PL equation systems

  (for non open problems)
- use clipped Models to preserve global properties (i.e. symmetric, bounded)



- extension to euclidean norm or $\mathrm{sign}(\cdot) \implies$ algebraic inclusion

References

- A.G.: On stable piecewise linearization and generalized algorithmic differentiation. *Optimization Methods and Software*, (2013).
- P. Boeck, B. Gompil, A. Griewank, R. Hasenfelder, N. Strogies: Experiments with Generalized Midpoint and Trapezoidal Rules on two Nonsmooth ODE's, *Mongolian Journal of Mathematics*, (2014).
- A.G., M. Radons, T. Streubel, J.U. Bernd: On solving piecewise linear systems in abs-normal form. *Linear Algebra and Applications*, (2014).
- A.G., A. Walther, S. Fiege, T. Bosse: On Lipschitz optimization based on gray-box piecewise linearization, *Mathematical Programming*, (2015).
- A.G. et al: Plan-C, Piecewise Linear functions in Abs Normal form, software under development.
- M. Radons, A.G., T. Streubel, J.U. Bernd: IFIP TC 7 / 2013 System Modelling and Optimization September 8-13, (2013),

  Proceeding paper published in *System Modeling and Optimization*, *Springer*, (2014)

Arriving at Yachay Tech University, Ecuador