Local convergence of adaptive tensor methods

Karl Welzel
Mathematical Institute
University of Oxford

NA Internal Seminar, 21 November 2024

Mathematical Institute

Oxford Mathematics

Local convergence of adaptive tensor methods

Work in progress with

Raphael Hauser

Yang Liu

Coralia Cartis

Outline

Outline

Why higher-order methods?

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司

University of Oxford Mathematical Institute

- Goal: Find a local minimizer $\boldsymbol{x}_*$ of a smooth function $f\colon \mathbb{R}^n \to \mathbb{R}$
- We compare rates of convergence when $\boldsymbol{x}_k$ is close to $\boldsymbol{x}_*$ (local convergence)
- Newton's method converges quadratically when $\nabla^2 f(\boldsymbol{x}_*) \succ \mathbf{0}$
- Newton's method converges only linearly for $f(x) = x^4$ for example
- Access to higher derivatives $\to$ superlinear convergence for singular $\nabla^2 f(\boldsymbol{x}_*)$

The ARp method

Algorithm 1.1: Adaptive regularization algorithm using up to pth derivatives

Parameters: $\sigma_0 > 0$, $0 < \eta < 1$, $0 < \gamma_1 \leq 1 < \gamma_2$

for $k = 0, 1, \dots$ **do**
- Compute objective function and derivatives $f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k), \dots, \nabla^p f(\boldsymbol{x}_k)$
- Construct local Taylor expansion as $t_k(\boldsymbol{y}) = \sum_{j=0}^{p} \frac{1}{j!} \nabla^j f(\boldsymbol{x}_k)[\boldsymbol{y} - \boldsymbol{x}_k]^j$
- Construct local model as $m_k(\boldsymbol{y}) = t_k(\boldsymbol{y}) + \sigma_k \|\boldsymbol{y} - \boldsymbol{x}_k\|^{p+1}$
- Find a local minimizer $\boldsymbol{y}_k \in \mathbb{R}^n$ of m_k that satisfies $m_k(\boldsymbol{y}_k) < m_k(\boldsymbol{x}_k)$
- **if** $f(\boldsymbol{x}_k) - f(\boldsymbol{y}_k) \geq \eta(t_k(\boldsymbol{x}_k) - t_k(\boldsymbol{y}_k))$ **then**
  - Set $\boldsymbol{x}_{k+1} = \boldsymbol{y}_k$ and $\sigma_{k+1} = \gamma_1 \sigma_k$ `// successful iteration`
- **else**
  - Set $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$ and $\sigma_{k+1} = \gamma_2 \sigma_k$ `// unsuccessful iteration`
- **end**

end

The ARp method

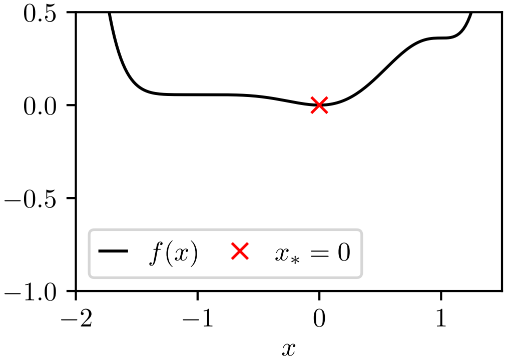
- Assume $\nabla^p f$ is Lipschitz continuous: $\|\nabla^p f(\boldsymbol{x}) - \nabla^p f(\boldsymbol{y})\| \leq L_p \|\boldsymbol{x} - \boldsymbol{y}\|$
- "Ideal" regularization is $\sigma = \frac{L_p}{(p+1)!}$ but L_p is unknown
- Adaptive method is less dependent on user choice σ_0

Outline

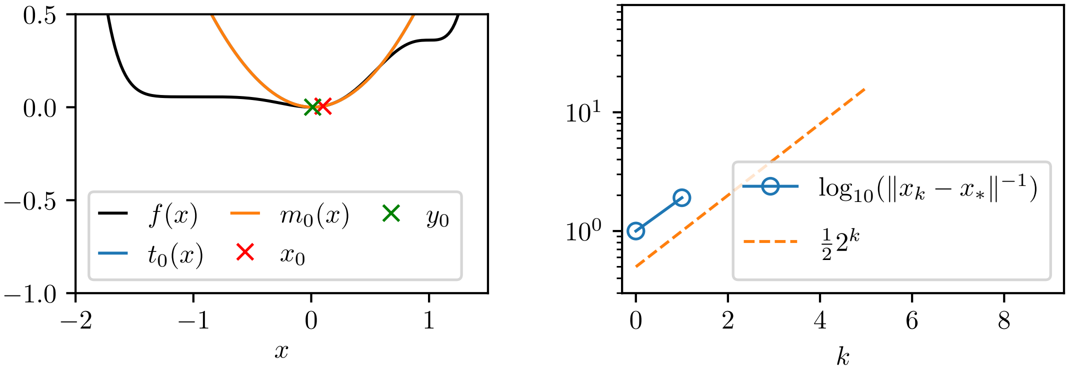
1. Motivation
   - Why higher-order methods?
   - The ARp method
2. Numerical Illustrations
   - Example with non-degenerate minimizer
   - Example with degenerate minimizer
3. Theoretical Results
4. Conclusion

Example with non-degenerate minimizer

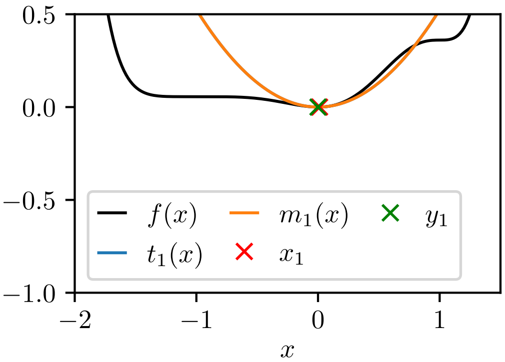
- $f(x) = \int_0^x (t+1)^4 (t-1)^2 t \,\mathrm{d}t$
- f has one local (and global) minimizer at $x_* = 0$
- f is not globally convex, but locally strongly convex: $f''(x_*) = 1$

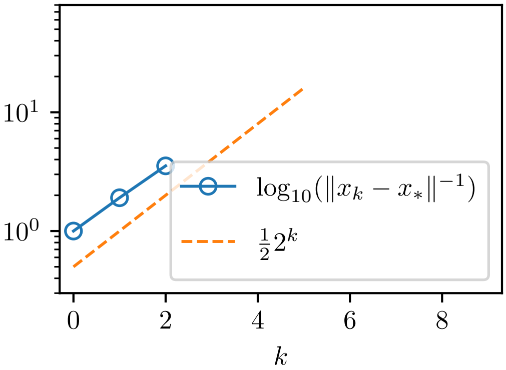


Newton's method ($p = 2$, $\sigma = 0$)

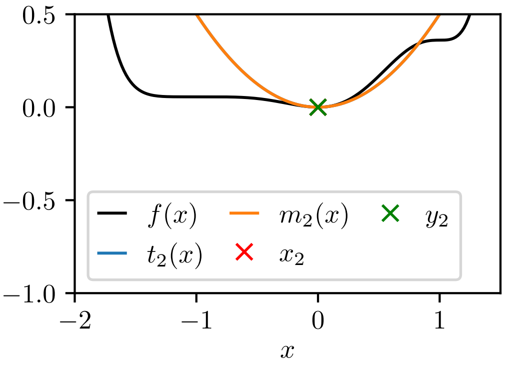


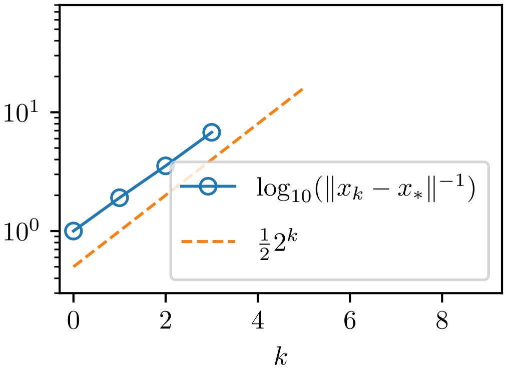
Newton's method ($p = 2$, $\sigma = 0$)



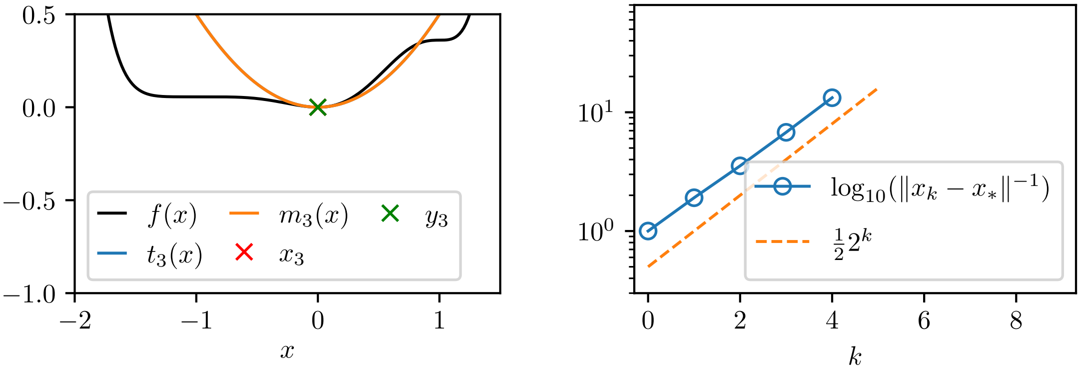


Newton's method ($p = 2$, $\sigma = 0$)

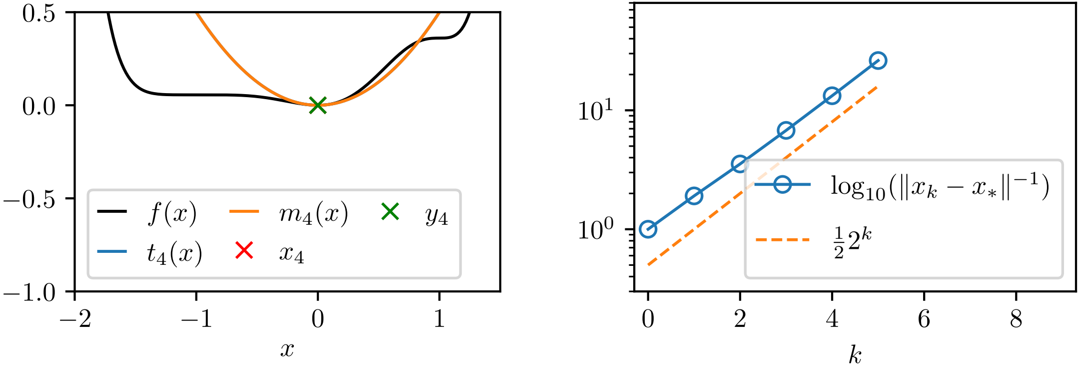




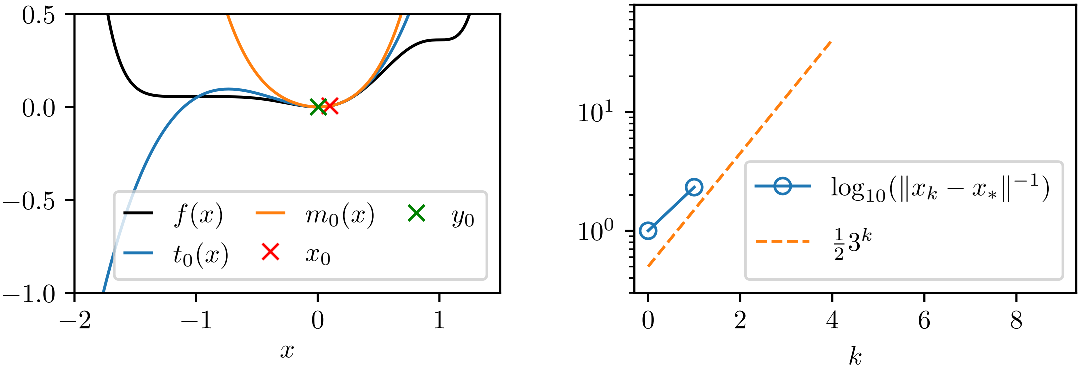
Newton's method ($p = 2,\ \sigma = 0$)



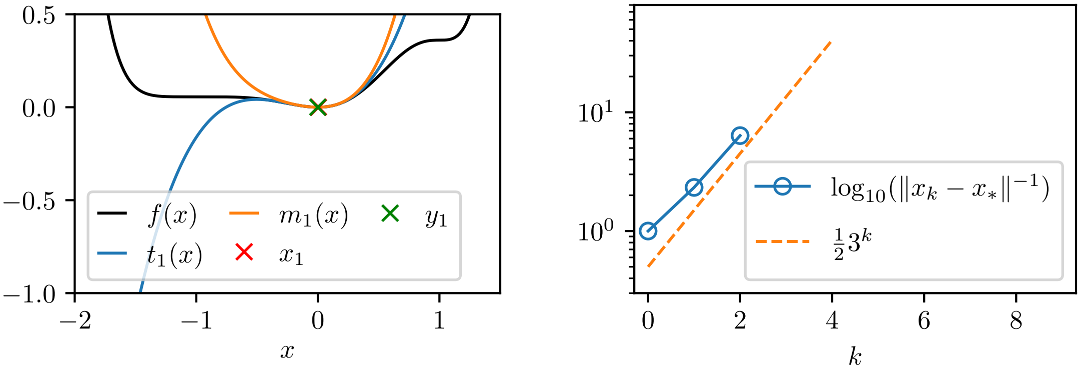
Newton's method ($p = 2$, $\sigma = 0$)



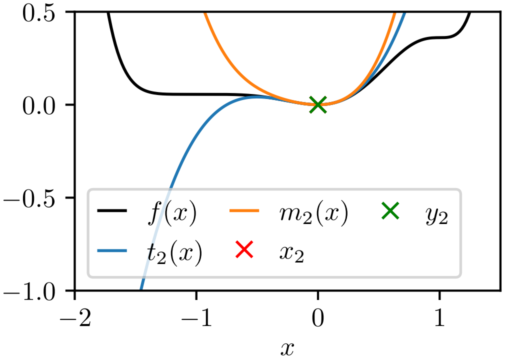
AR3, semi-adaptive σ_k

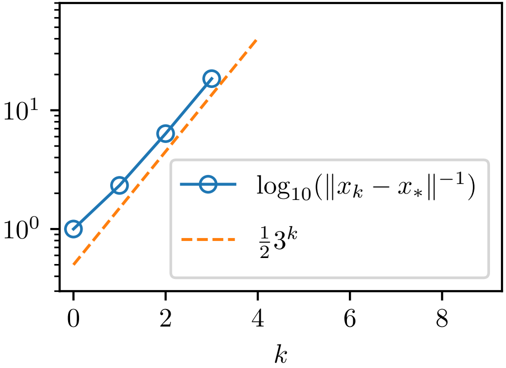


AR3, semi-adaptive σ_k

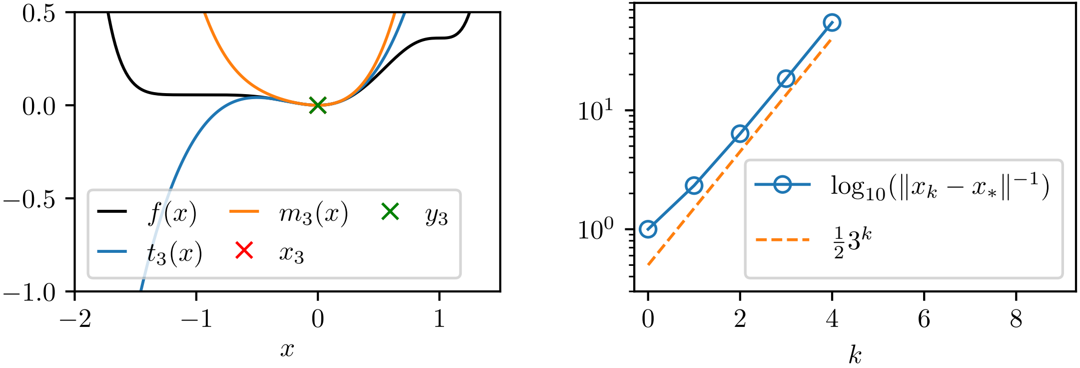


AR3, semi-adaptive σ_k

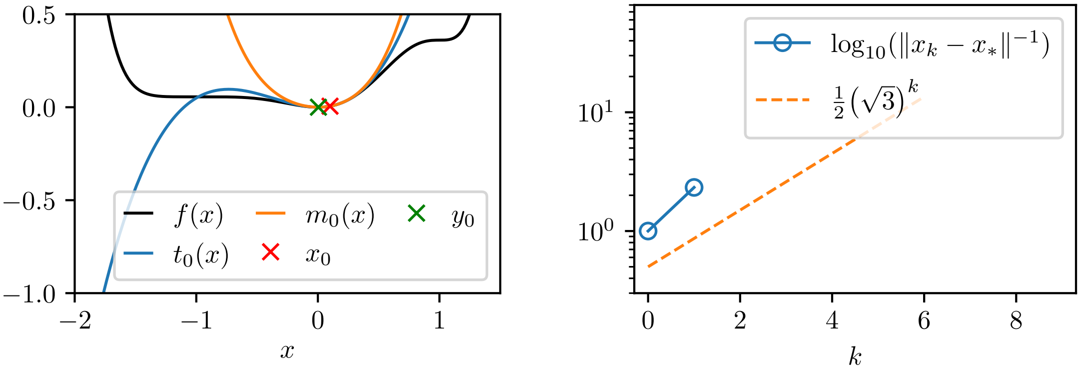




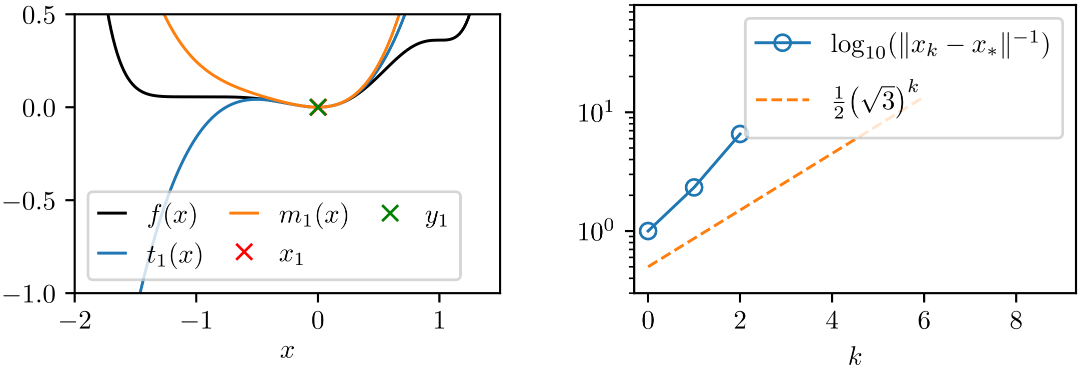
AR3, semi-adaptive σ_k



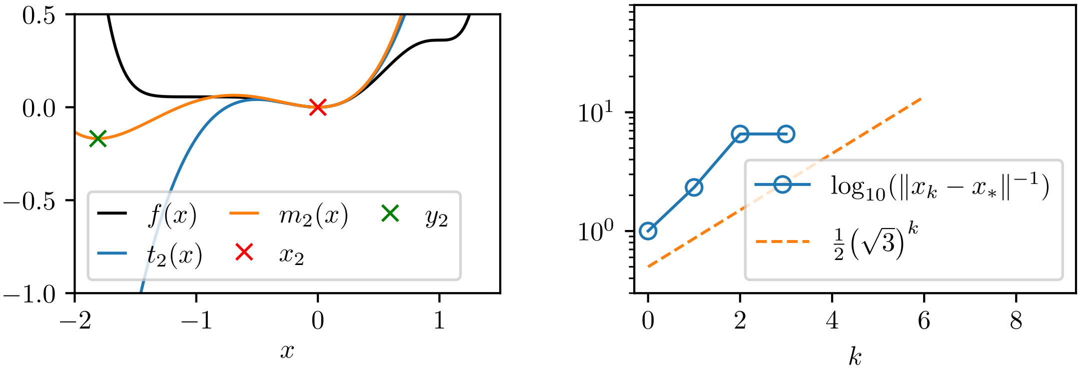
AR3, adaptive σ_k and global model minimizer



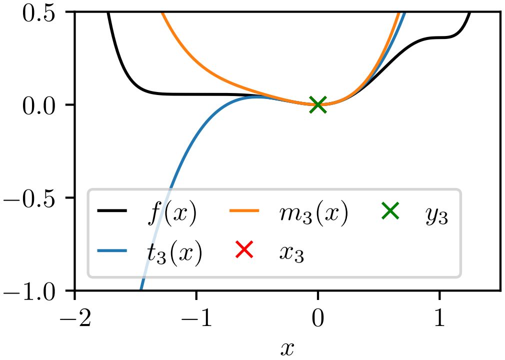
AR3, adaptive σ_k and global model minimizer

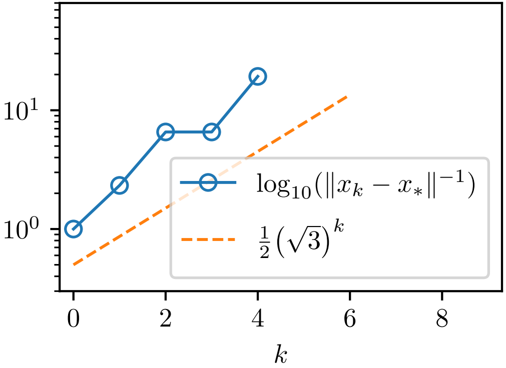


AR3, adaptive σ_k and global model minimizer

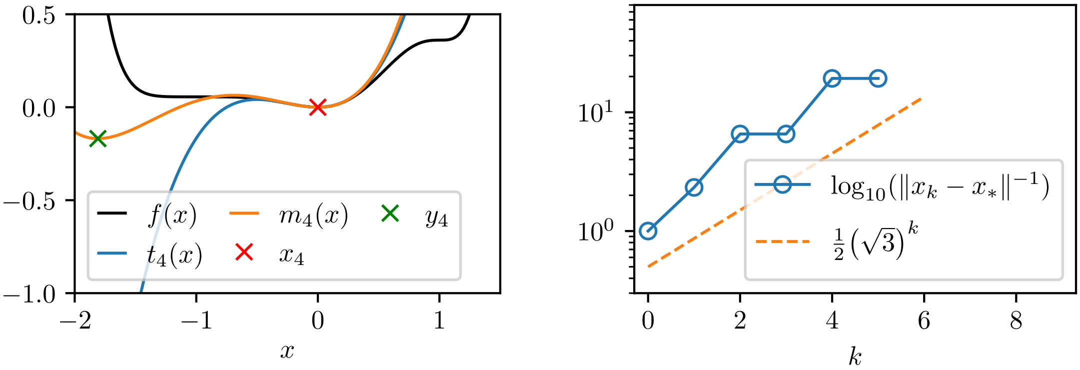


AR3, adaptive σ_k and global model minimizer

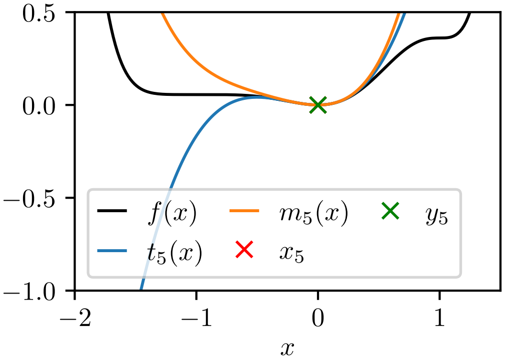


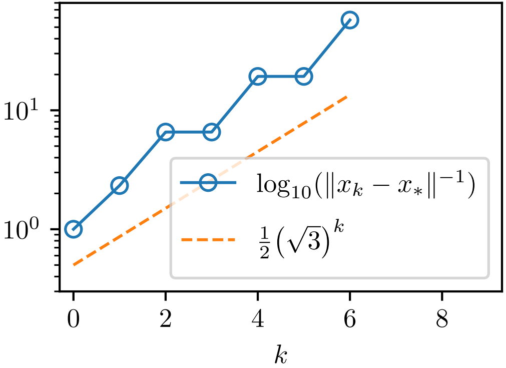


AR3, adaptive σ_k and global model minimizer

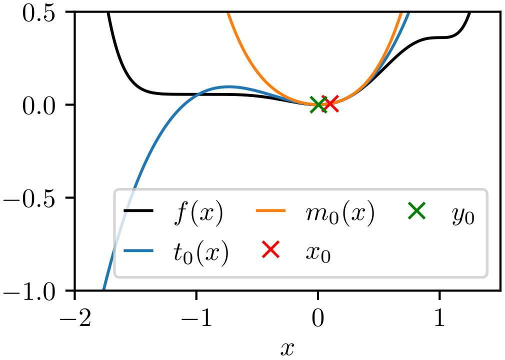


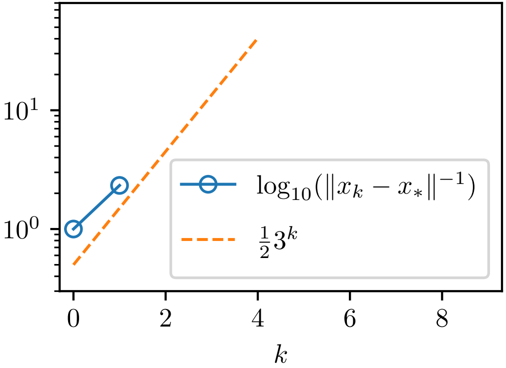
AR3, adaptive σ_k and global model minimizer



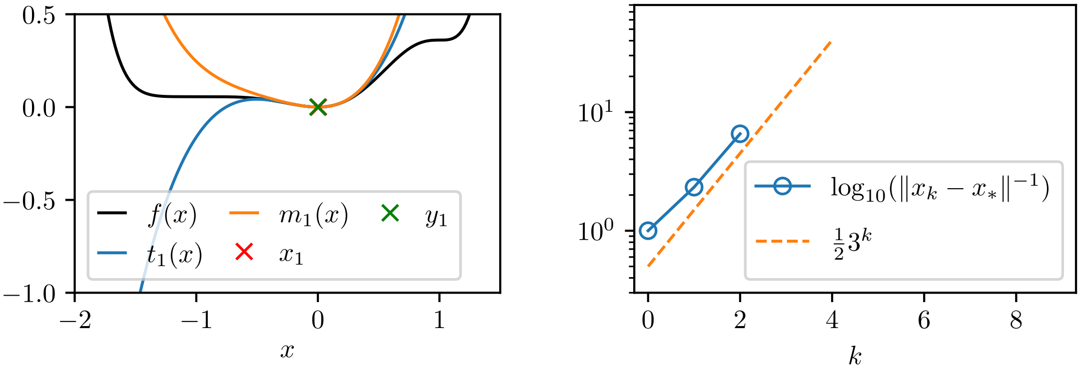


AR3, adaptive σ_k and right model minimizer

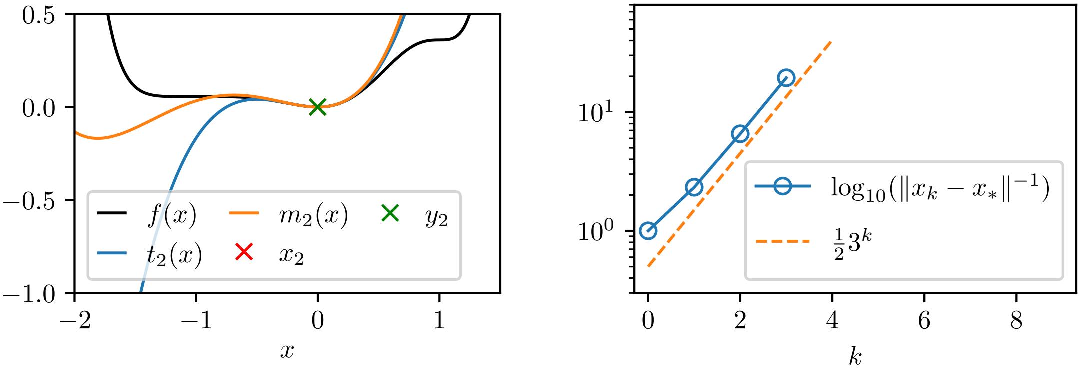




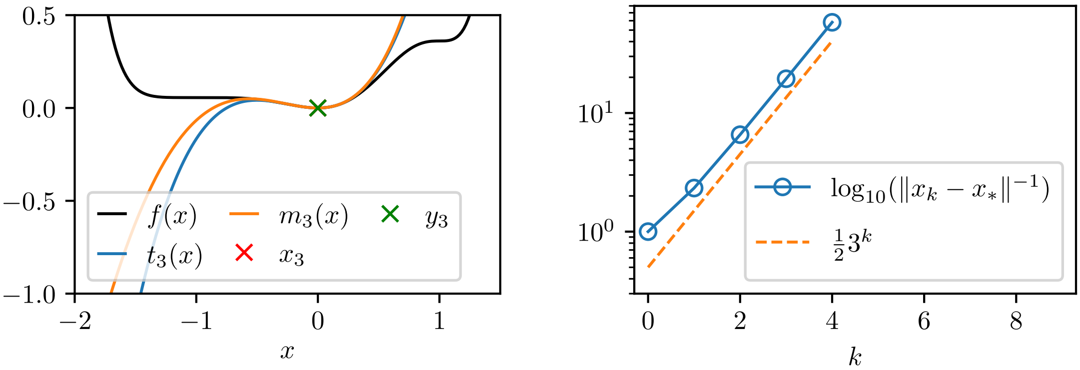
AR3, adaptive σ_k and right model minimizer



AR3, adaptive σ_k and right model minimizer



AR3, adaptive σ_k and right model minimizer



Example with degenerate minimizer

Definition

A function $f : \mathbb{R}^n \to \mathbb{R}$ is uniformly convex of order $q \geq 2$ inside $\Omega \subset \mathbb{R}^n$ if

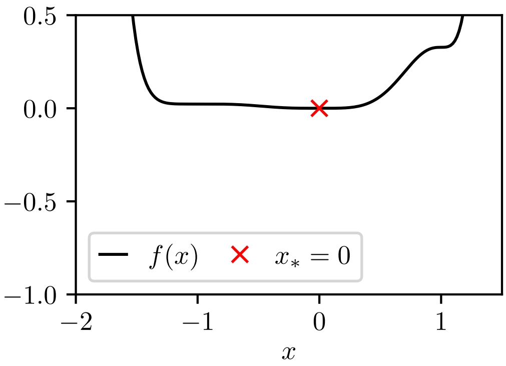
$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu_q}{q}\|\boldsymbol{y} - \boldsymbol{x}\|^q \quad \forall \boldsymbol{x}, \boldsymbol{y} \in \Omega$$

for some $\mu_q > 0$.

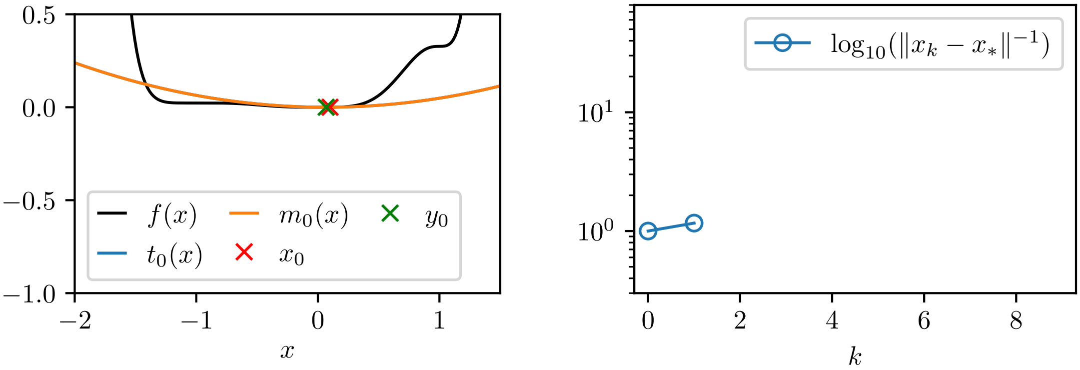
- Uniformly convex of order q around $\boldsymbol{x}_* \implies f(\boldsymbol{x}) - f(\boldsymbol{x}_*) \geq \frac{\mu_q}{q}\|\boldsymbol{x} - \boldsymbol{x}_*\|^q$
- Uniformly convex of order q around $\boldsymbol{x}_* \implies \boldsymbol{x}_*$ is an isolated minimizer
- Uniformly convex of order $2 \iff$ strongly convex

Example with degenerate minimizer

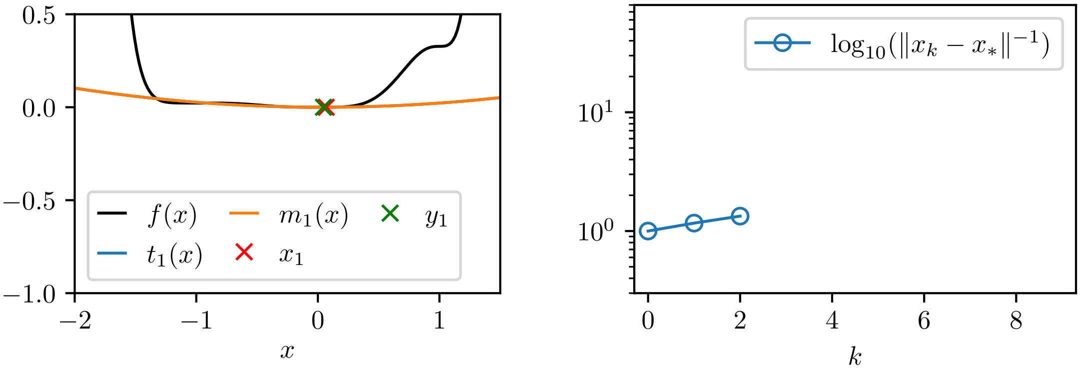
- $f(x) = \int_0^x 3(t+1)^4(t-1)^2 t^3 \, \mathrm{d}t$
- f has one local (and global) minimizer at $x_* = 0$
- f is locally uniformly convex around x_* with $q = 4$:
  - $f''(x_*) = 0$
  - $f'''(x_*) = 0$
  - $f''''(x_*) = 18$



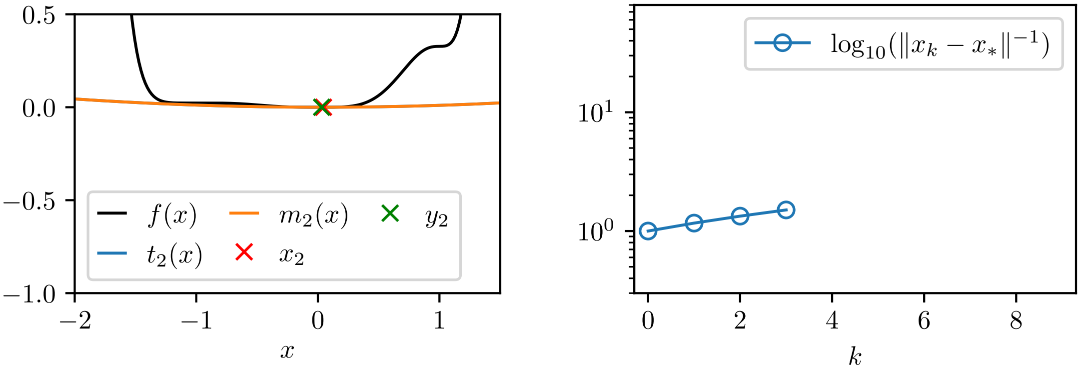
Newton's method ($p = 2$, $\sigma = 0$)



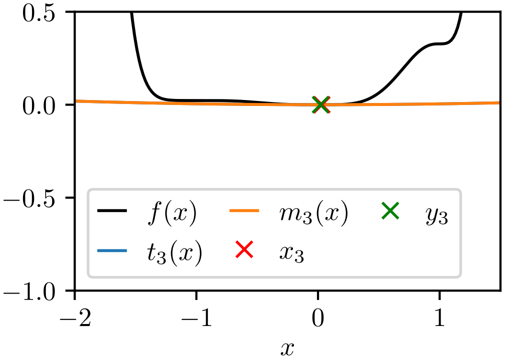
Newton's method ($p = 2,\ \sigma = 0$)

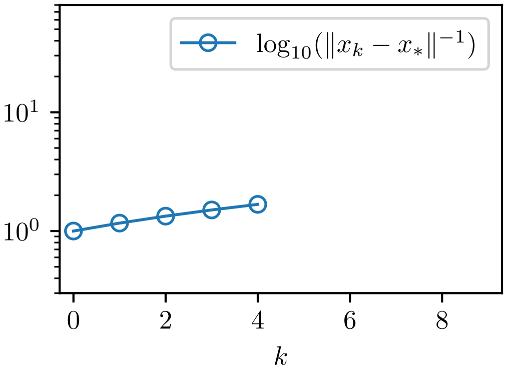


Newton's method ($p = 2,\ \sigma = 0$)

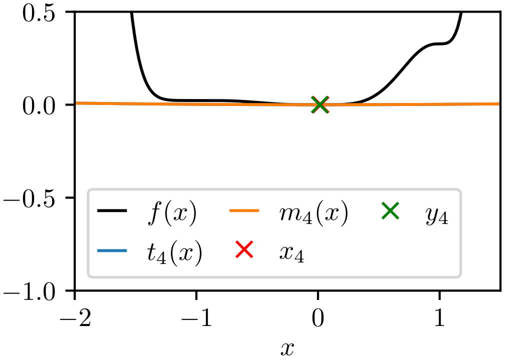


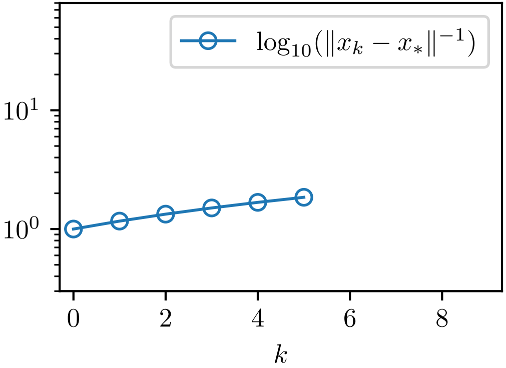
Newton's method ($p = 2,\ \sigma = 0$)



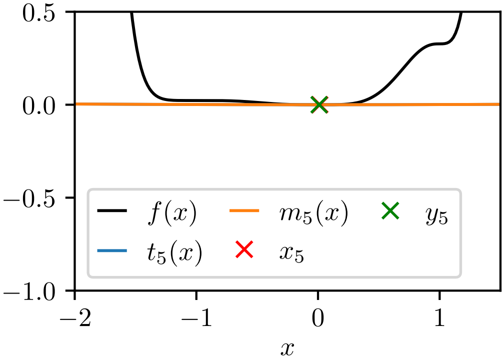


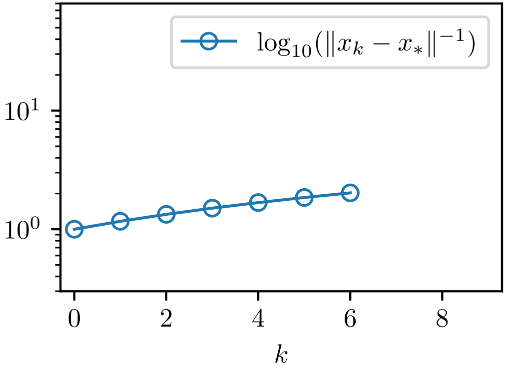
Newton's method ($p = 2,\ \sigma = 0$)



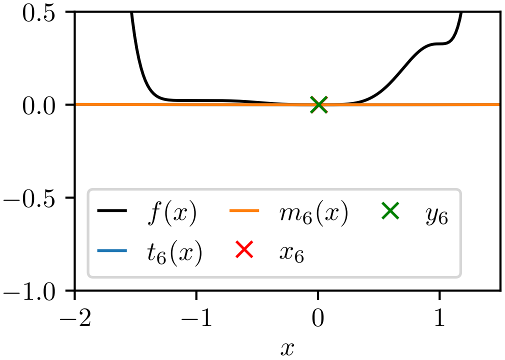


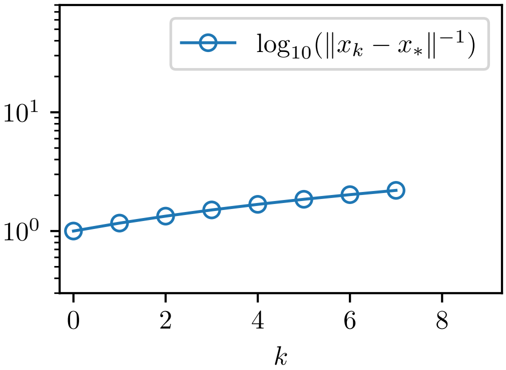
Newton's method ($p = 2$, $\sigma = 0$)



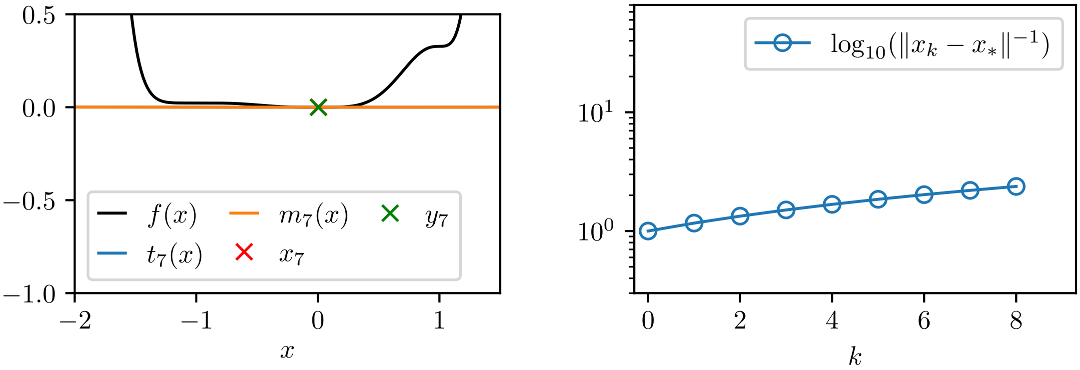


Newton's method ($p = 2$, $\sigma = 0$)

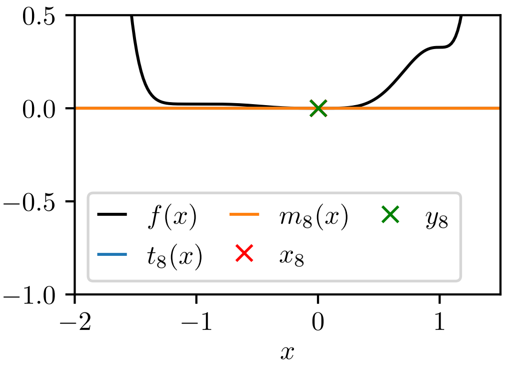


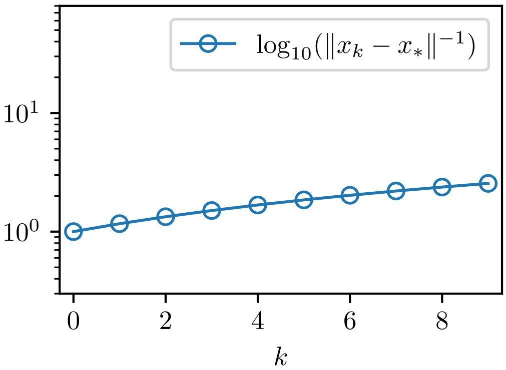


Newton's method ($p = 2$, $\sigma = 0$)

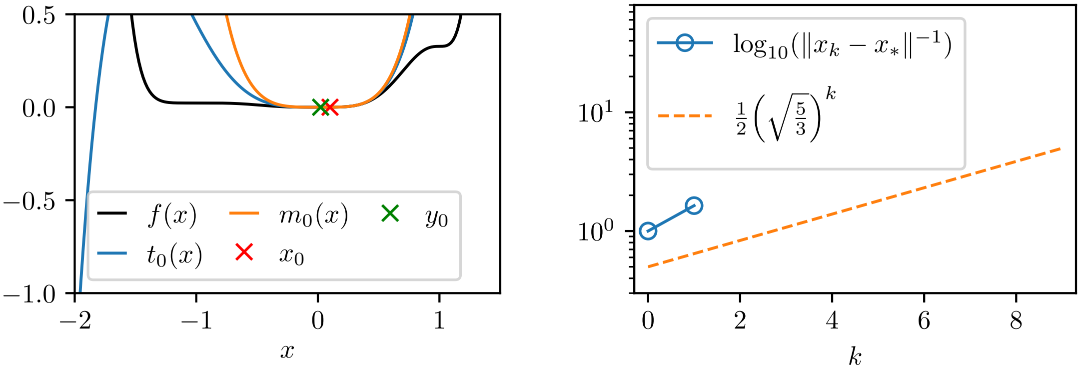


Newton's method ($p = 2$, $\sigma = 0$)

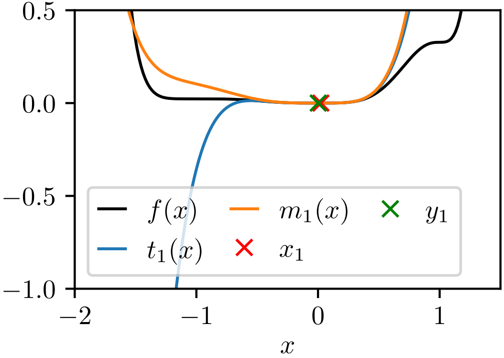


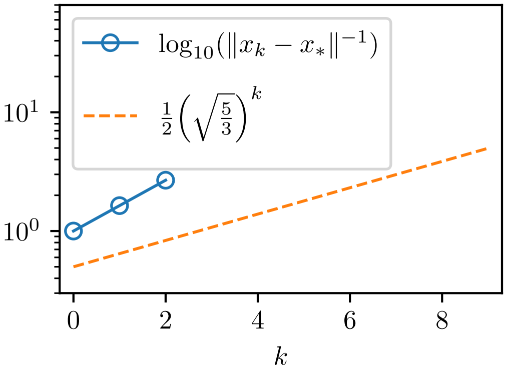


AR5, adaptive σ_k and global model minimizer

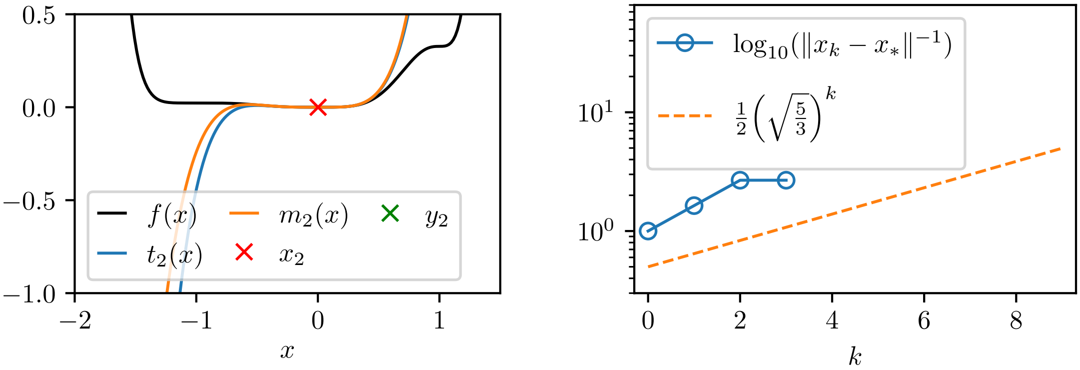


AR5, adaptive σ_k and global model minimizer

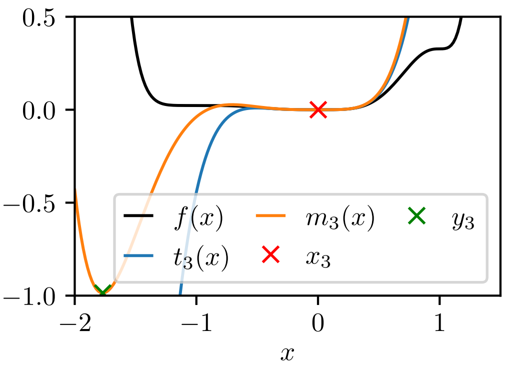


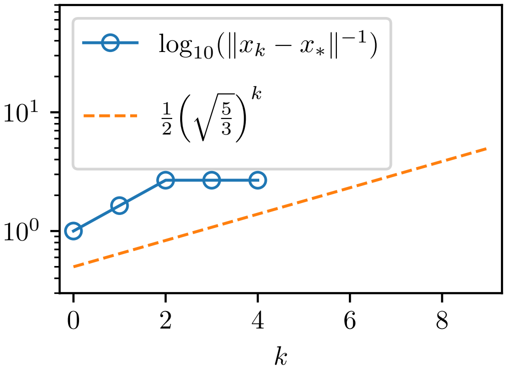


AR5, adaptive σ_k and global model minimizer

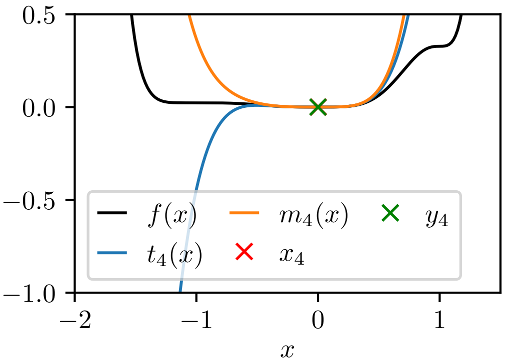


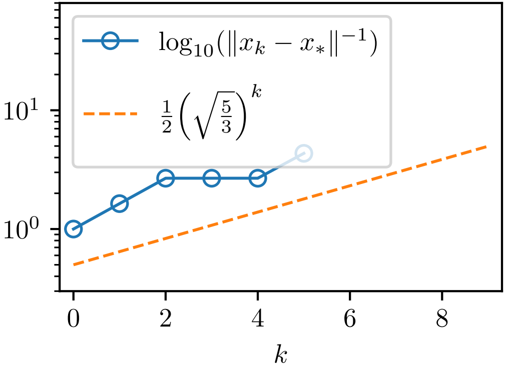
AR5, adaptive σ_k and global model minimizer



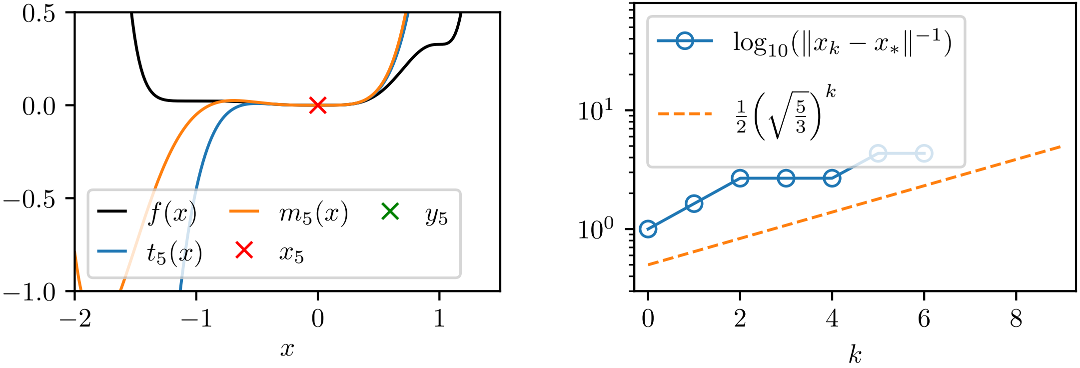


AR5, adaptive σ_k and global model minimizer

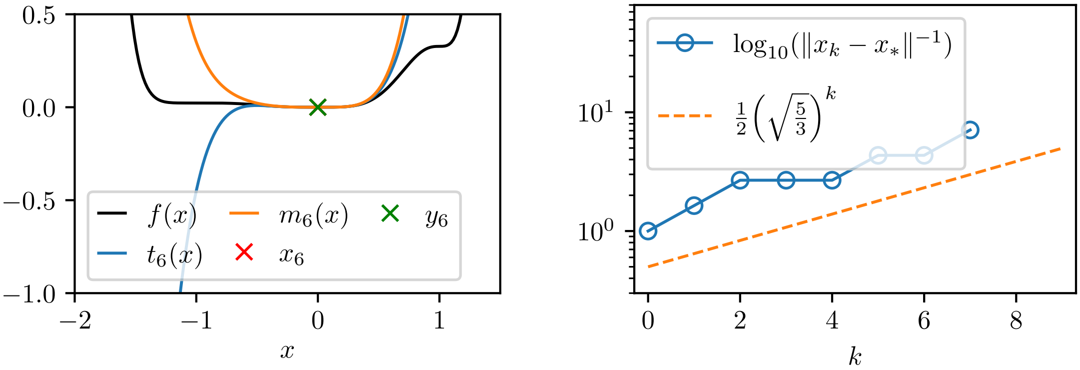




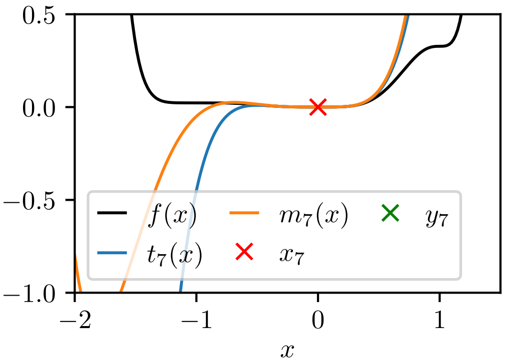
AR5, adaptive σ_k and global model minimizer

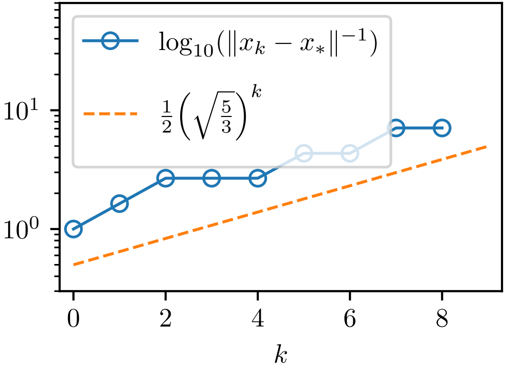


AR5, adaptive σ_k and global model minimizer

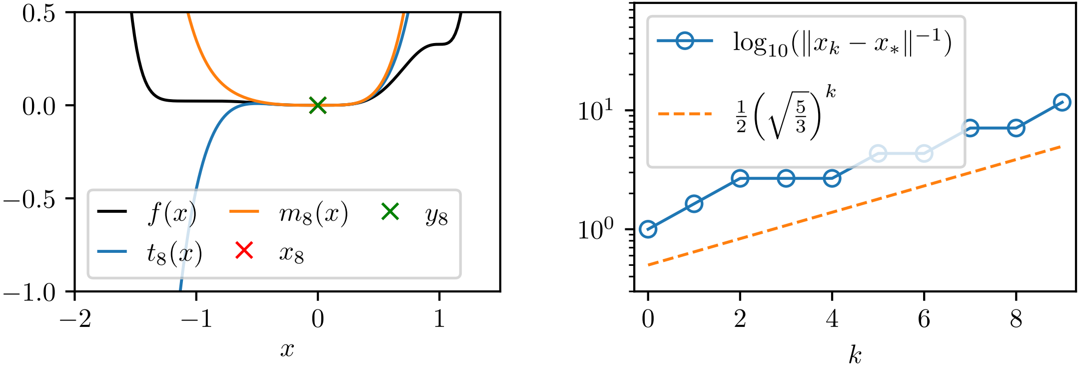


AR5, adaptive σ_k and global model minimizer

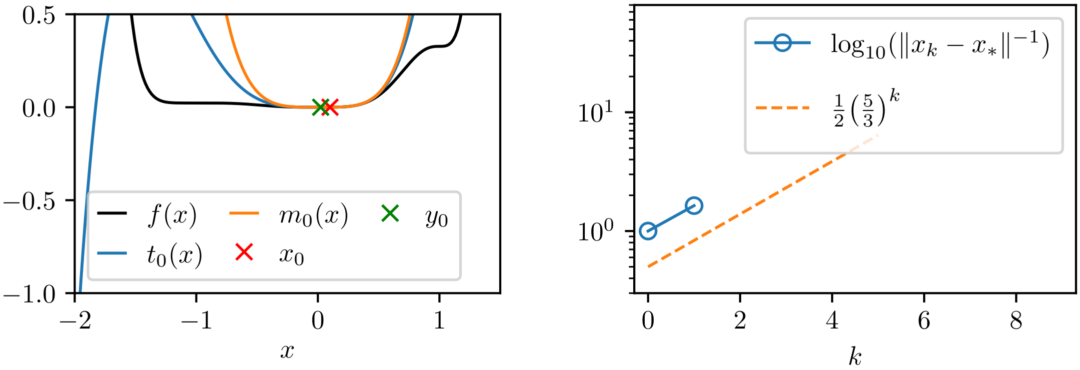




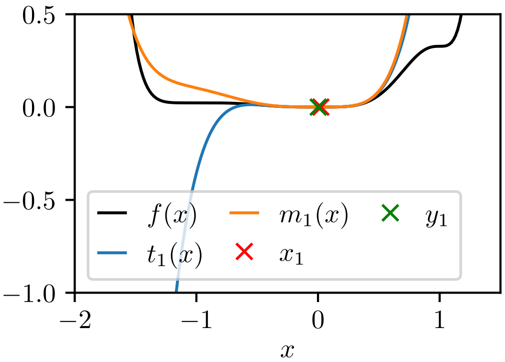
AR5, adaptive σ_k and global model minimizer

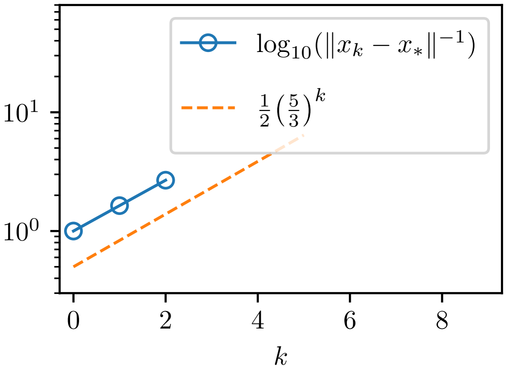


AR5, adaptive σ_k and right model minimizer

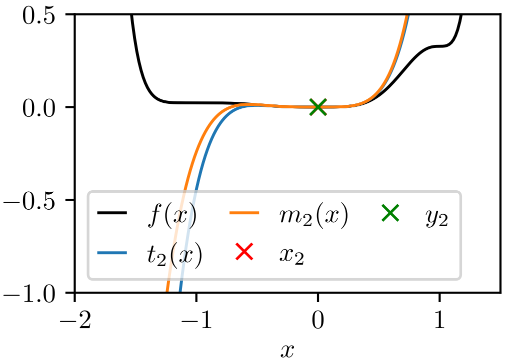


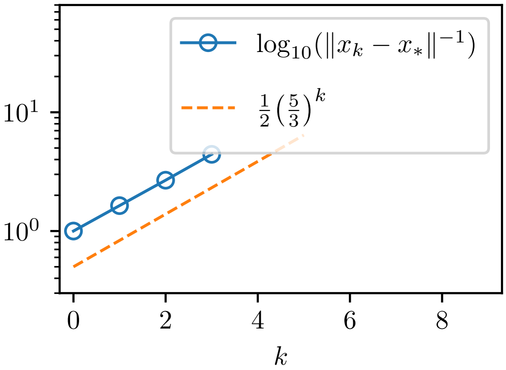
AR5, adaptive σ_k and right model minimizer



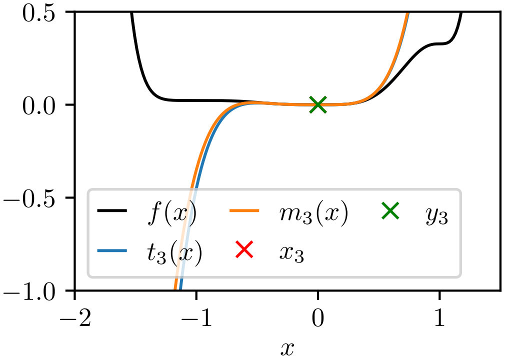


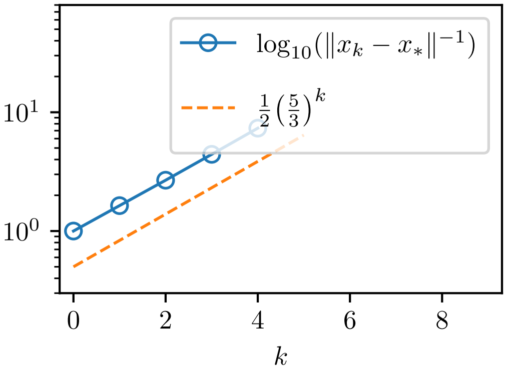
AR5, adaptive σ_k and right model minimizer



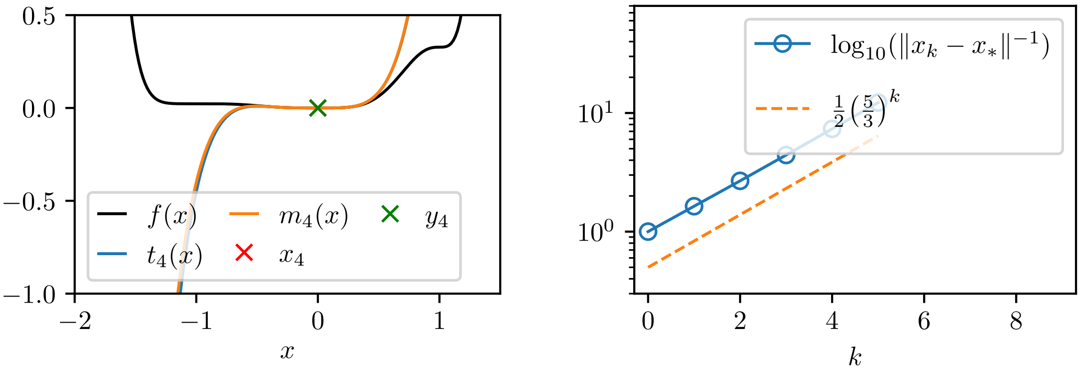


AR5, adaptive σ_k and right model minimizer





AR5, adaptive σ_k and right model minimizer



Outline

Convergence for successful iterations

Lemma

Let $\boldsymbol{x}_*$ be a local minimizer of f, let $\nabla^p f$ be Lipschitz continuous and let f be uniformly convex of order q inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$. Assume that $\boldsymbol{x}_k, \boldsymbol{y}_k \in \mathcal{B}(\boldsymbol{x}_*, r_\mu)$ and that iteration k is successful, then the gradients satisfy

$$\|\nabla f(\boldsymbol{y}_k)\| \leq (L_p/p! + (p+1)\sigma_k)\left(\frac{q}{\mu_q}\right)^{\frac{p}{q-1}} \|\nabla f(\boldsymbol{x}_k)\|^{\frac{p}{q-1}}.$$

Convergence for successful iterations

Theorem

Let $\boldsymbol{x}_*$ be a local minimizer of f, let $\nabla^p f$ be Lipschitz continuous and let f be uniformly convex of order q inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$. Assume that all iterates stay inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$, that $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$ and that $p > q - 1$, then

$$
\begin{aligned}
\|\nabla f(\boldsymbol{x}_{k_i})\| &\to 0 && \text{at a } Q\text{-}\tfrac{p}{q-1}\text{th-order rate} \\
f(\boldsymbol{x}_{k_i}) &\to f(\boldsymbol{x}_*) && \text{at an } R\text{-}\tfrac{p}{q-1}\text{th-order rate} \\
\boldsymbol{x}_{k_i} &\to \boldsymbol{x}_* && \text{at an } R\text{-}\tfrac{p}{q-1}\text{th-order rate}
\end{aligned}
$$

where $\{k_1, k_2, \ldots\}$ are the successful iterations.

Convergence for worst-case model minimizer

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司

University of Oxford Mathematical Institute

Theorem

Let $\boldsymbol{x}_*$ be a local minimizer of f, let $\nabla^p f$ be Lipschitz continuous and let f be uniformly convex of order q inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$. Assume that all iterates stay inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$, that $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$ and that $p > q - 1$, then

$\|\nabla f(\boldsymbol{x}_k)\| \to 0$ *at an R-* $\sqrt[\alpha+1]{\frac{p}{q-1}}$ *th-order rate*

$f(\boldsymbol{x}_k) \to f(\boldsymbol{x}_*)$ *at an R-* $\sqrt[\alpha+1]{\frac{p}{q-1}}$ *th-order rate*

$\boldsymbol{x}_k \to \boldsymbol{x}_*$ *at an R-* $\sqrt[\alpha+1]{\frac{p}{q-1}}$ *th-order rate*

where $\gamma_1 = \gamma_2^{-\alpha}$. ($\alpha = 0$ and $\alpha = 1$ in experiments)

Convergence for close model minimizer

Lemma

Let $\boldsymbol{x}_*$ *be a local minimizer of* f*, let* $\nabla^p f$ *be Lipschitz continuous and let* f *be uniformly convex of order* q *inside* $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$*. If* $p > q - 1$*, there exists radii* $r_x, r_y > 0$ *with* $r_x \leq r_y \leq r_\mu$ *such that for any* $\boldsymbol{x}_k \in \mathcal{B}(\boldsymbol{x}_*, r_x)$ *there exists a local minimizer of* m_k *inside* $\mathcal{B}(\boldsymbol{x}_*, r_y)$ *and any such minimizer will give a successful iteration.*

Convergence for close model minimizer

Centre for Intelligent Multidimensional Data Analysis Limited
智能多維數據分析研究中心有限公司

University of Oxford Mathematical Institute

Theorem

Let $\boldsymbol{x}_*$ be a local minimizer of f, let $\nabla^p f$ be Lipschitz continuous, let f be uniformly convex of order q inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$ and let r_x and r_y be chosen according to the previous lemma. Assume that $\boldsymbol{y}_k \in \mathcal{B}(\boldsymbol{x}_*, r_y)$ in every iteration, that $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$ and that $p > q - 1$, then all iterations are successful and

$$
\begin{aligned}
\|\nabla f(\boldsymbol{x}_k)\| &\to 0 && \text{at a Q-}\tfrac{p}{q-1}\text{th-order rate} \\
f(\boldsymbol{x}_k) &\to f(\boldsymbol{x}_*) && \text{at an R-}\tfrac{p}{q-1}\text{th-order rate} \\
\boldsymbol{x}_k &\to \boldsymbol{x}_* && \text{at an R-}\tfrac{p}{q-1}\text{th-order rate.}
\end{aligned}
$$

Comparison with Doikov and Nesterov, 2022

Doikov, N., Nesterov, Y. Local convergence of tensor methods. Math. Program. 193, 315–336 (2022). https://doi.org/10.1007/s10107-020-01606-x

- Assumptions:
  - $\nabla^p f$ Lipschitz continuous
  - f uniformly convex of order q
  - $p > q - 1$
  - σ_k constant and $\geq \frac{pL_p}{(p+1)!}$
- $\|\nabla f(\boldsymbol{x}_k)\|$ and $f(\boldsymbol{x}_k)$ converge with $\frac{p}{q-1}$th-order rate
- Also covers composite optimization case with convex non-differentiable component

- Assumptions
  - $\nabla^p f$ Lipschitz continuous
  - f locally uniformly convex of order q
  - $p > q - 1$
  - σ_k adaptively chosen
- $\|\nabla f(\boldsymbol{x}_k)\|$, $f(\boldsymbol{x}_k)$ and $\boldsymbol{x}_k$ converge with $\sqrt[1+\alpha]{\frac{p}{q-1}}$th-order rate
- $\|\nabla f(\boldsymbol{x}_k)\|$, $f(\boldsymbol{x}_k)$ and $\boldsymbol{x}_k$ converge with $\frac{p}{q-1}$th-order rate when choosing the "right" minimizer

Outline

1 Motivation
- Why higher-order methods?
- The ARp method

2 Numerical Illustrations
- Example with non-degenerate minimizer
- Example with degenerate minimizer

3 Theoretical Results

4 Conclusion

Conclusion

- For tensor methods ($p > 3$) choosing the right model minimizer is crucial
- Tensor methods achieve $\frac{p}{q-1}$th-order rates in theory and experiments
- Tensor methods achieve superlinear convergence even for degenerate minimizers