Local convergence of adaptively regularized tensor methods

Karl Welzel
Mathematical Institute
University of Oxford

EUROPT 2025, 1 July 2025

Oxford
Mathematics

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

Local convergence of adaptive tensor methods

Collaboration with

Raphael Hauser

Yang Liu

Coralia Cartis

The ARp method

- Unconstrained minimization $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$, $f\colon \mathbb{R}^n \to \mathbb{R}$ smooth, non-convex
- Tensor methods use stronger oracles
  - Gradient descent uses $f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k)$
  - Newton's method uses $f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k), \nabla^2 f(\boldsymbol{x}_k)$
  - pth-order tensor methods use $f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k), \ldots, \nabla^p f(\boldsymbol{x}_k)$
- ARp has optimal worst-case complexity, improves with p [Cartis, Gould, and Toint 2020]
- This talk: local convergence

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 $k = 0, 1, \ldots$ **do**
- Compute objective function and derivatives $f(\boldsymbol{x}_k), \nabla f(\boldsymbol{x}_k), \ldots, \nabla^p f(\boldsymbol{x}_k)$
- 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

Known results

Assumptions:

- $\nabla^p f$ is Lipschitz continuous: $\|\nabla^p f(\boldsymbol{x}) - \nabla^p f(\boldsymbol{y})\| \leq L_p \|\boldsymbol{x} - \boldsymbol{y}\|$
- f is uniformly convex of order q: $f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\mu_q}{q}\|\boldsymbol{y} - \boldsymbol{x}\|^q$

[Doikov and Nesterov 2022]: If $\sigma \geq \frac{pL_p}{(p+1)!}$ constant and $p > q - 1$ then

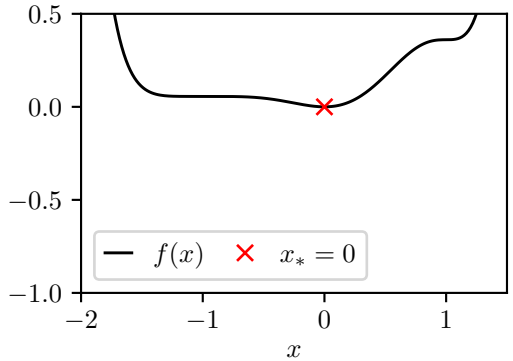
- $f(\boldsymbol{x}_k) \to f^*$ at a $\frac{p}{q-1}$th-order rate and
- $\nabla f(\boldsymbol{x}_k) \to \boldsymbol{0}$ at a $\frac{p}{q-1}$th-order rate.

[Cartis, Gould, and Toint 2022]: Using an adaptive σ_k, for $\nu = \frac{p+1}{p}$, $\mu = \frac{q}{q-1}$ the function values converge as follows (until $f(\boldsymbol{x}) - f^* < \varepsilon$)

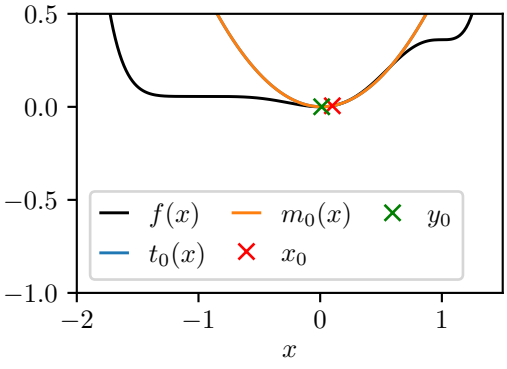
- $p < q - 1$ ($\nu > \mu$): sublinear, $O(\varepsilon^{-\frac{\nu-\mu}{\mu}})$ iterations
- $p = q - 1$ ($\nu = \mu$): linear, $O(\log(\varepsilon^{-1}))$ iterations
- $p > q - 1$ ($\nu < \mu$): superlinear, $O(\log(\log(\varepsilon^{-1})))$ iterations

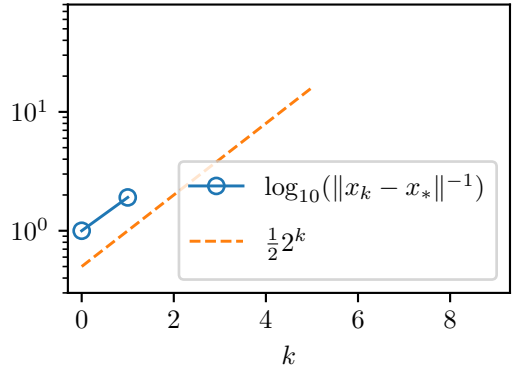
Example with non-degenerate minimizer

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

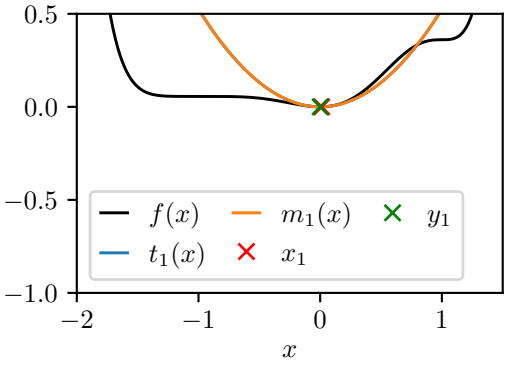


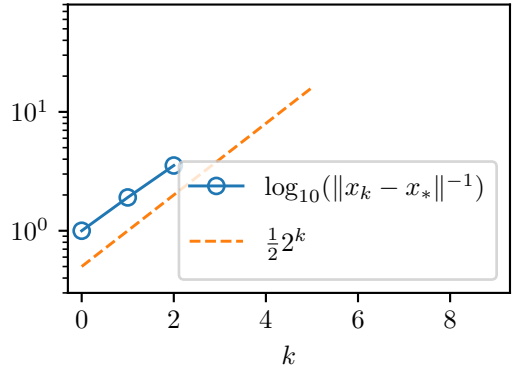
Newton's method ($p = 2$, $\sigma = 0$)



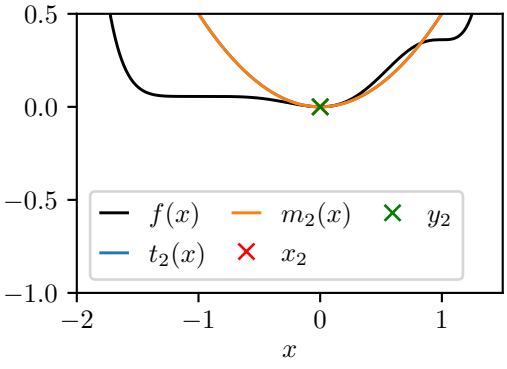


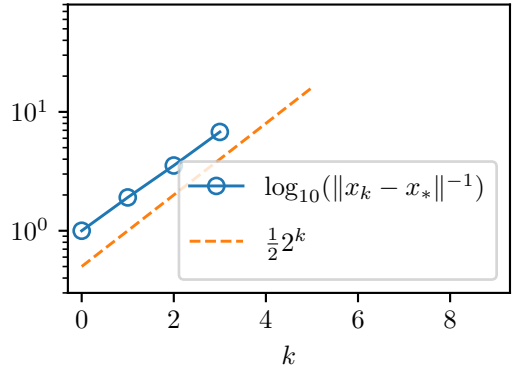
Newton's method ($p = 2$, $\sigma = 0$)



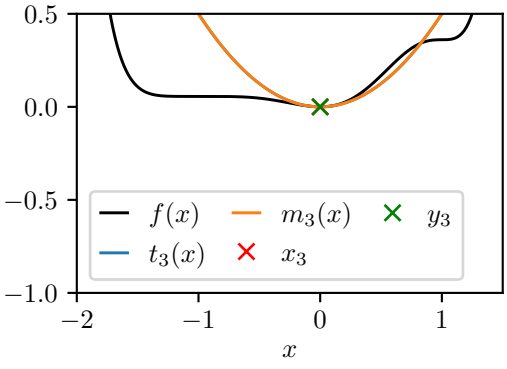


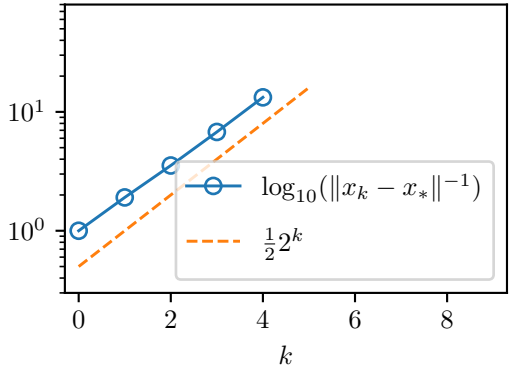
Newton's method ($p = 2$, $\sigma = 0$)



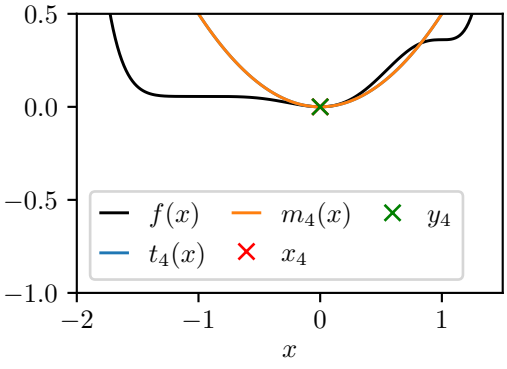


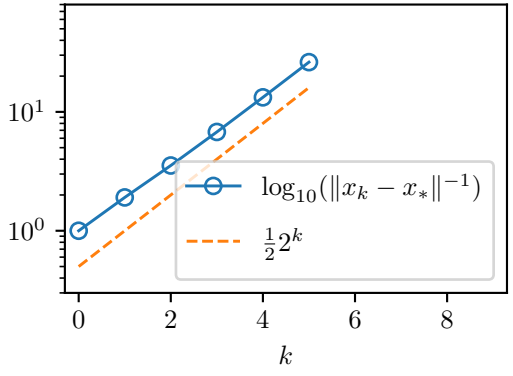
Newton's method ($p = 2$, $\sigma = 0$)



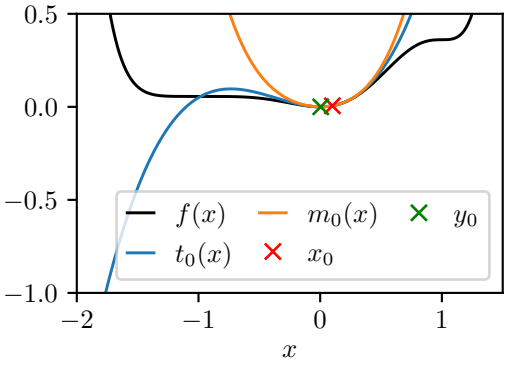


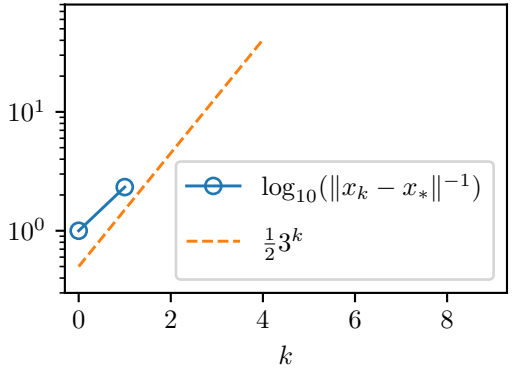
Newton's method ($p = 2$, $\sigma = 0$)



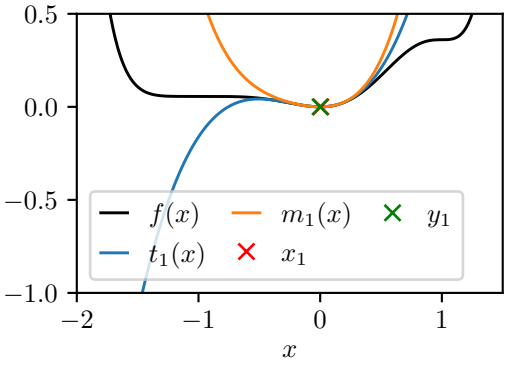


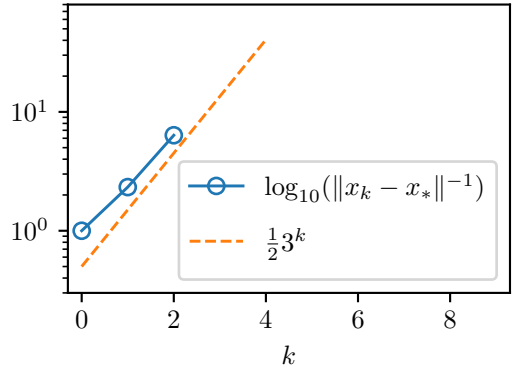
AR3, constant σ_k



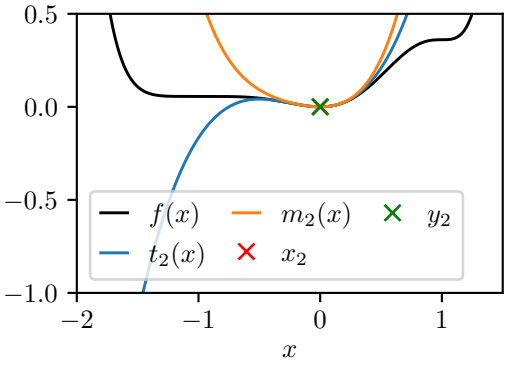


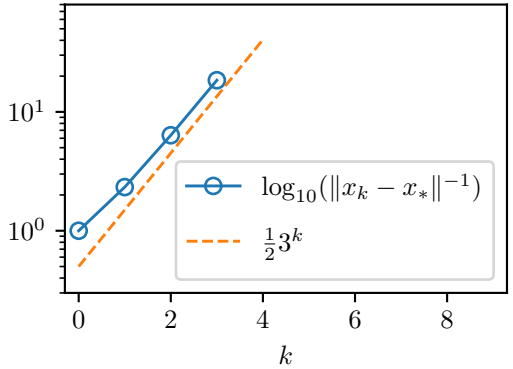
AR3, constant σ_k



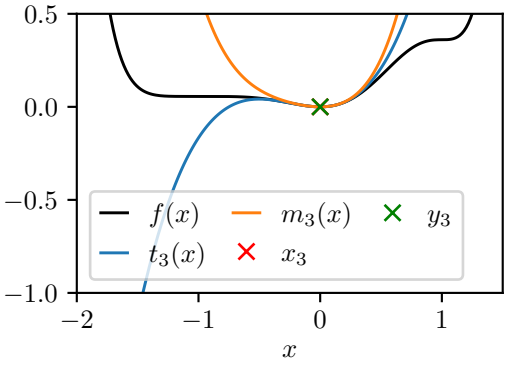


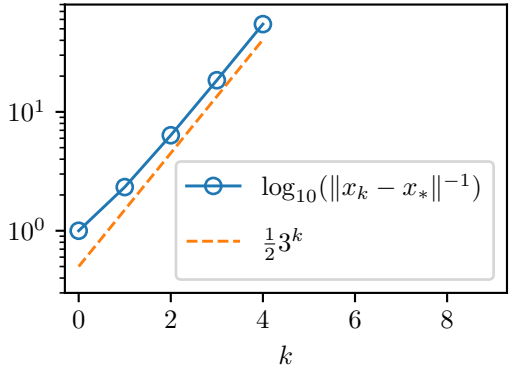
AR3, constant σ_k



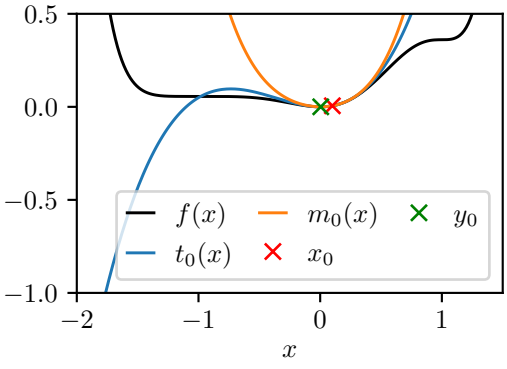


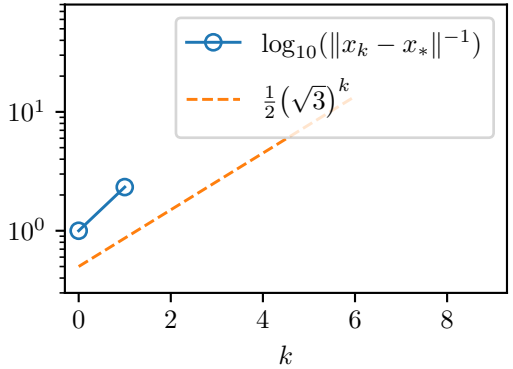
AR3, constant σ_k



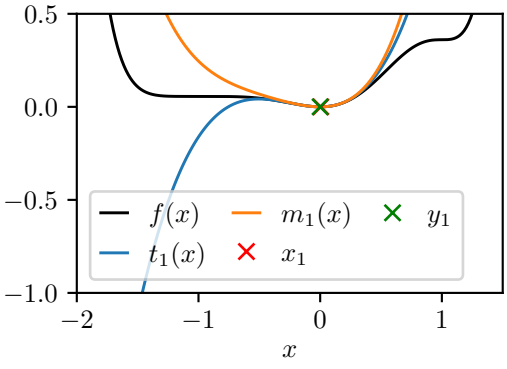


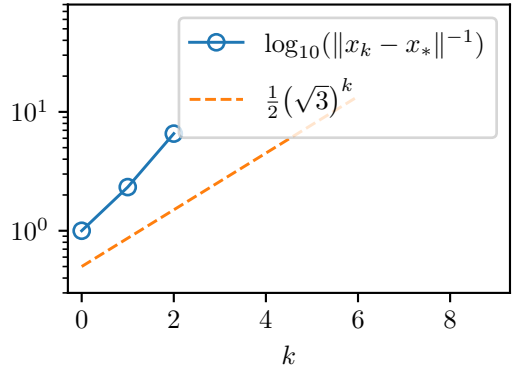
AR3, adaptive σ_k and global model minimizer



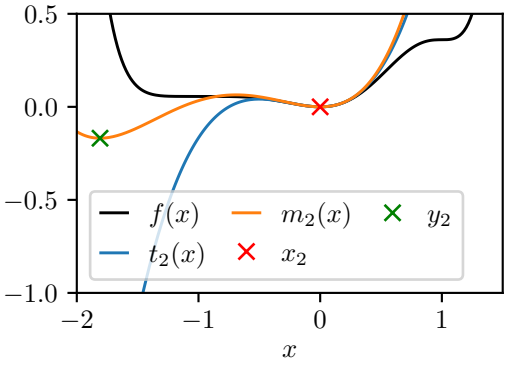


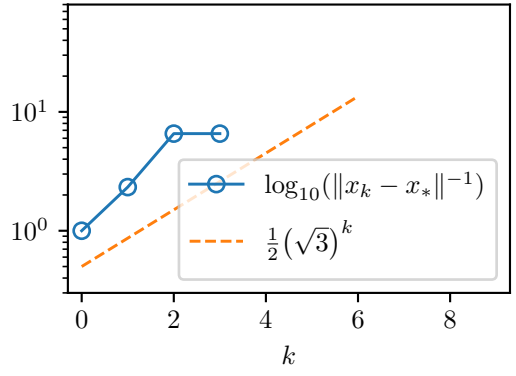
AR3, adaptive σ_k and global model minimizer



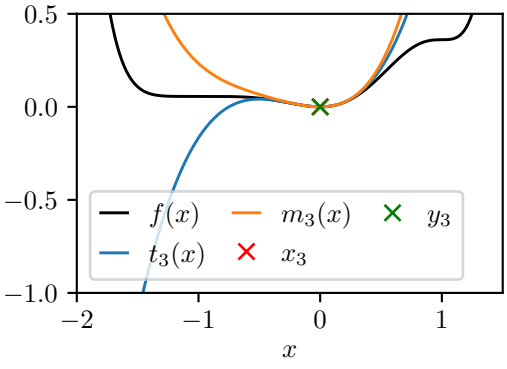


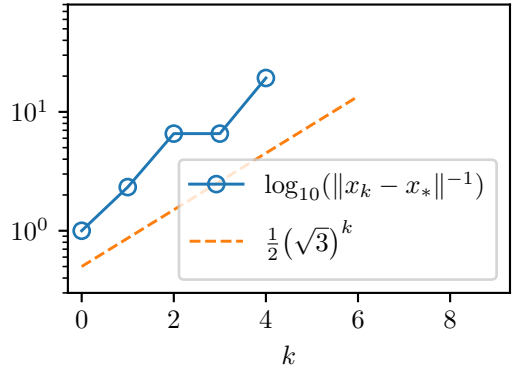
AR3, adaptive σ_k and global model minimizer



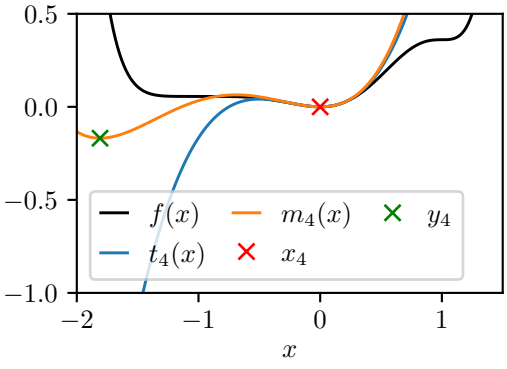


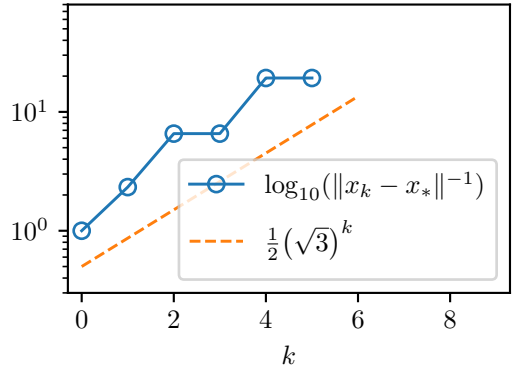
AR3, adaptive σ_k and global model minimizer



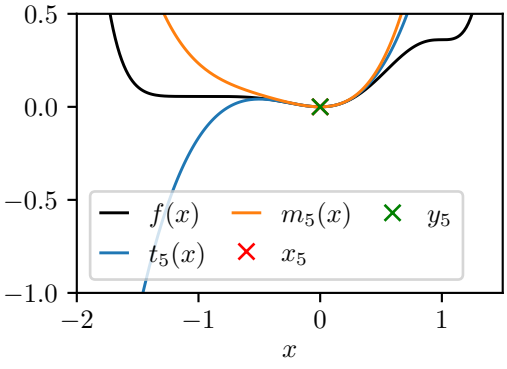


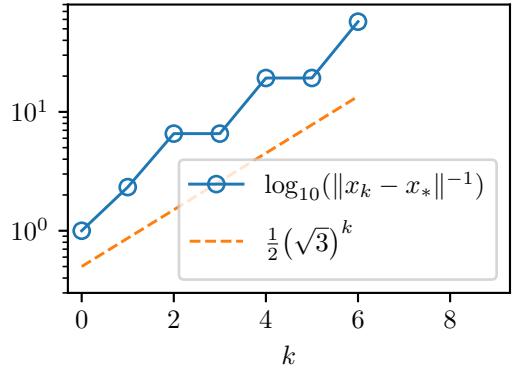
AR3, adaptive σ_k and global model minimizer



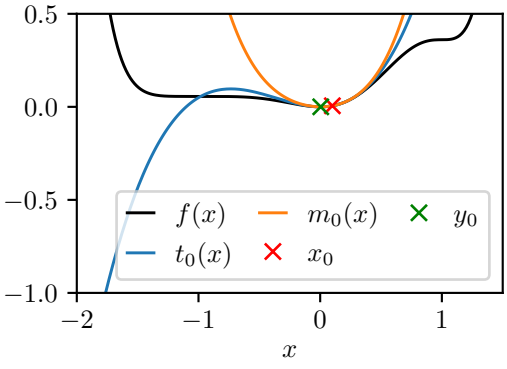


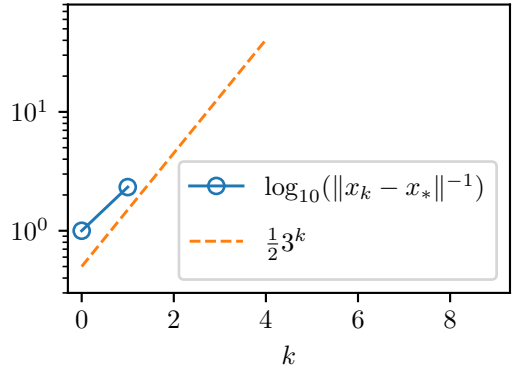
AR3, adaptive σ_k and global model minimizer



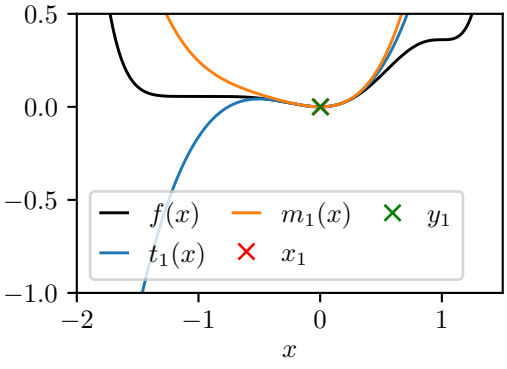


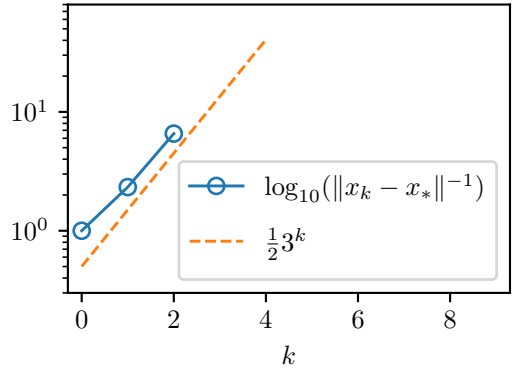
AR3, adaptive σ_k and right model minimizer



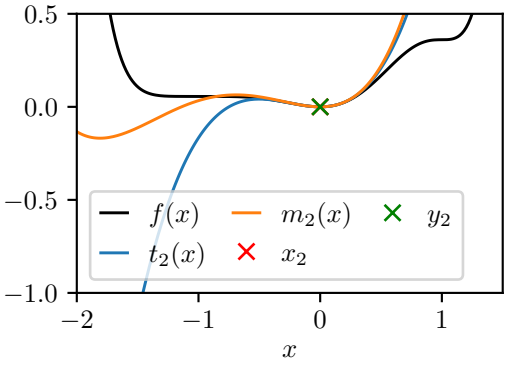


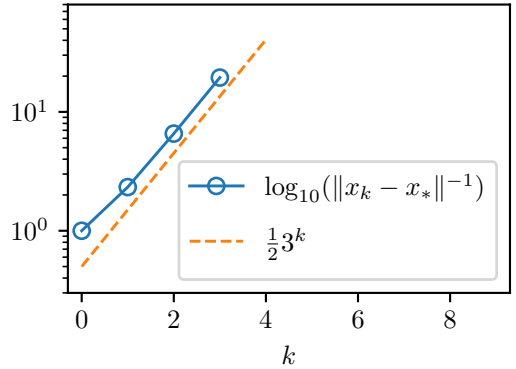
AR3, adaptive σ_k and right model minimizer



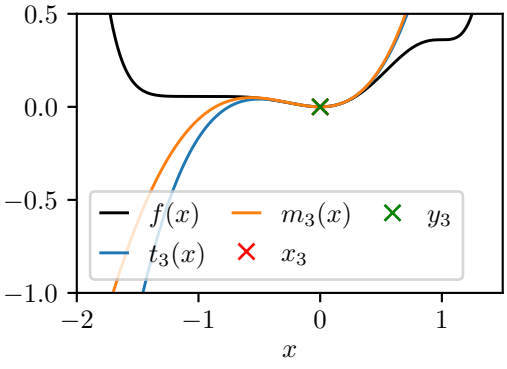


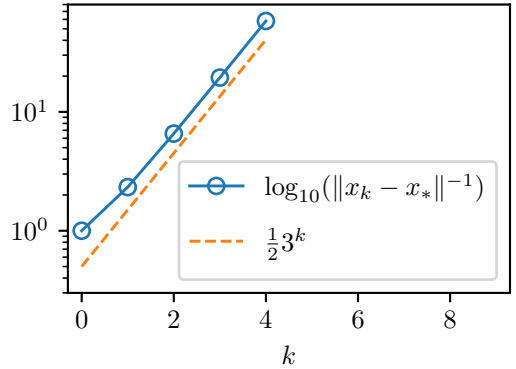
AR3, adaptive σ_k and right model minimizer





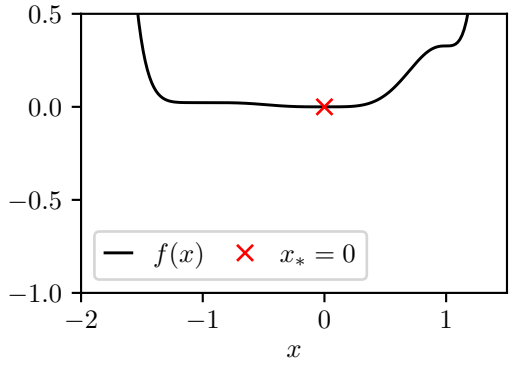
AR3, adaptive σ_k and right model minimizer



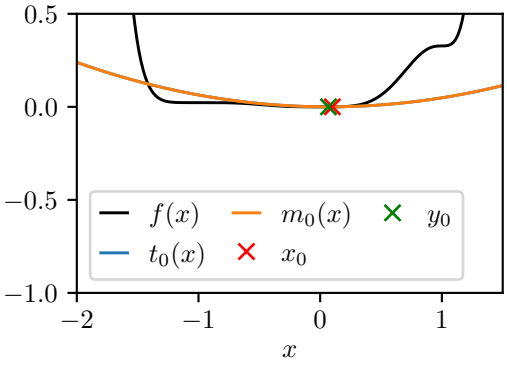


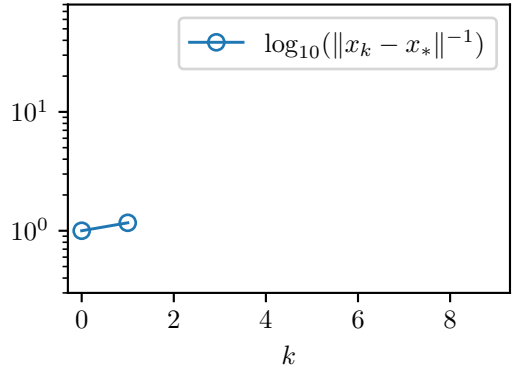
Example with degenerate minimizer

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

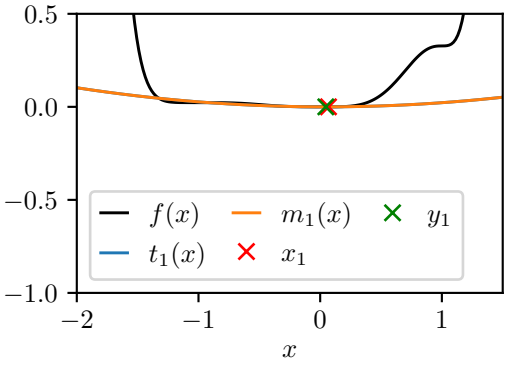


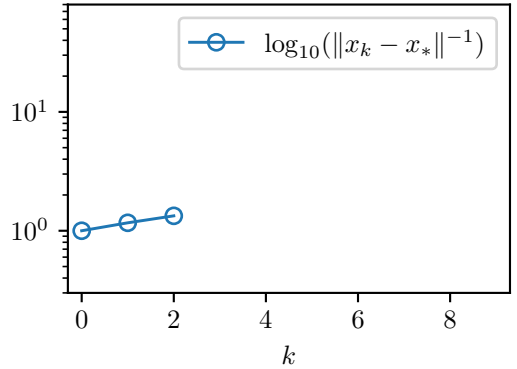
Newton's method ($p = 2$, $\sigma = 0$)



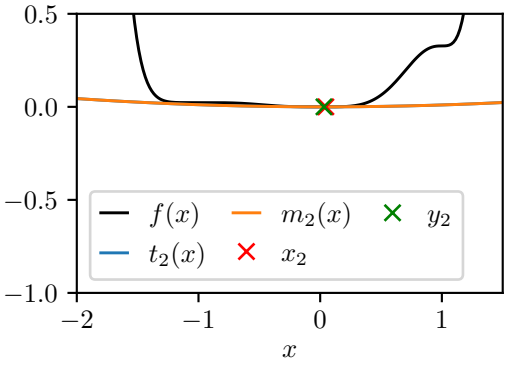


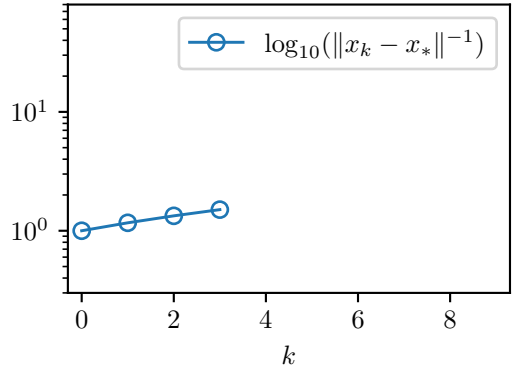
Newton's method ($p = 2$, $\sigma = 0$)



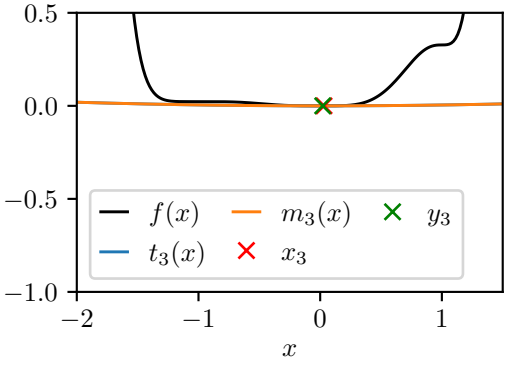


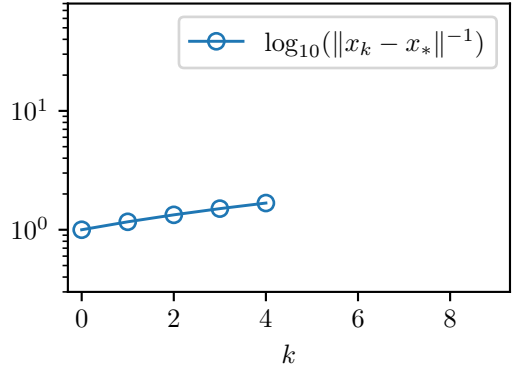
Newton's method ($p = 2$, $\sigma = 0$)



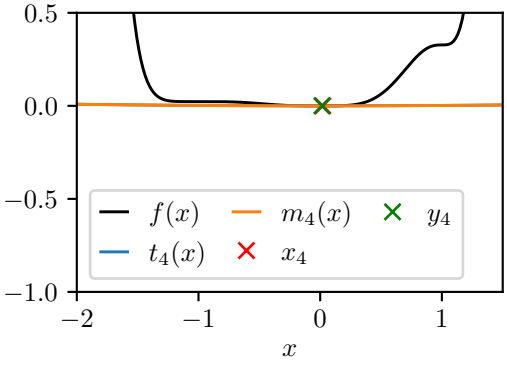


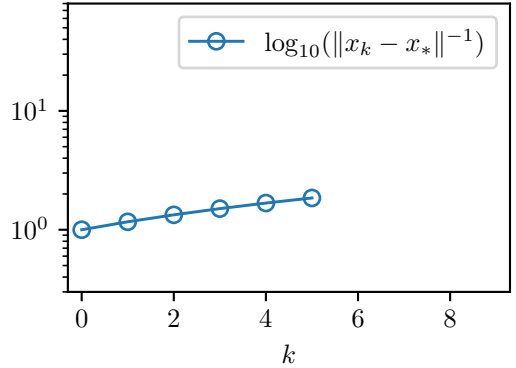
Newton's method ($p = 2$, $\sigma = 0$)



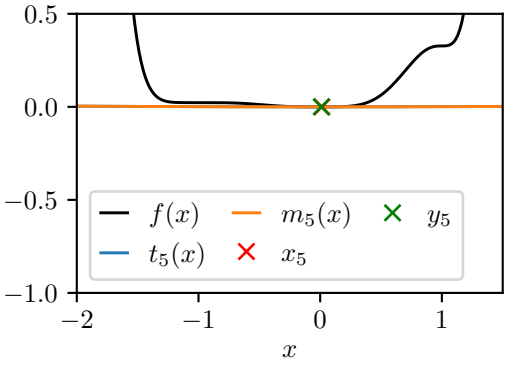


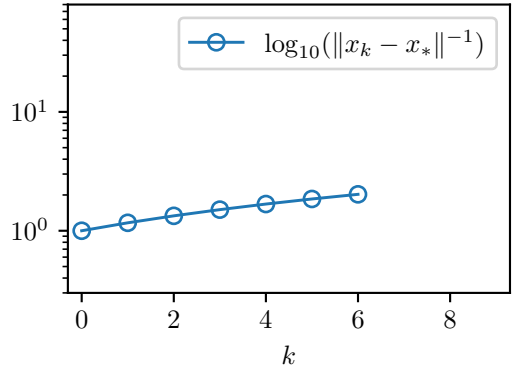
Newton's method ($p = 2$, $\sigma = 0$)



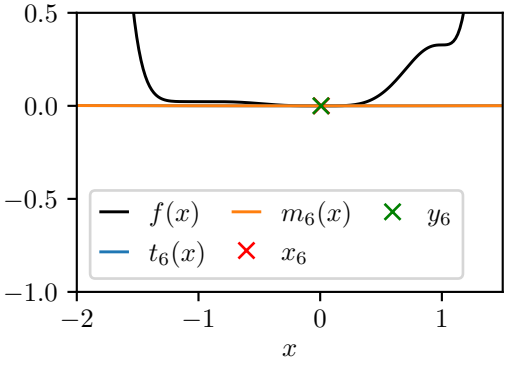


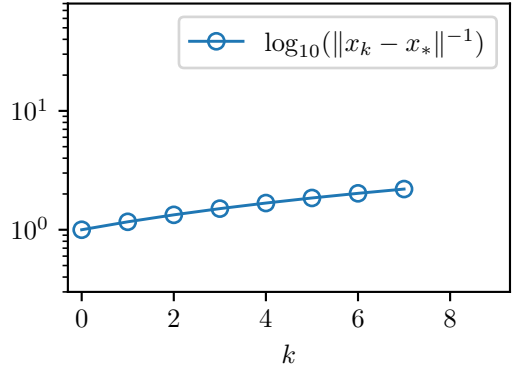
Newton's method ($p = 2$, $\sigma = 0$)



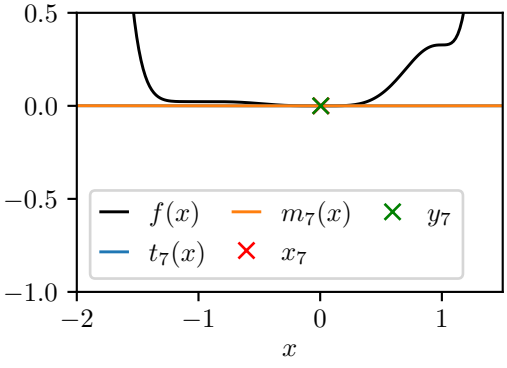


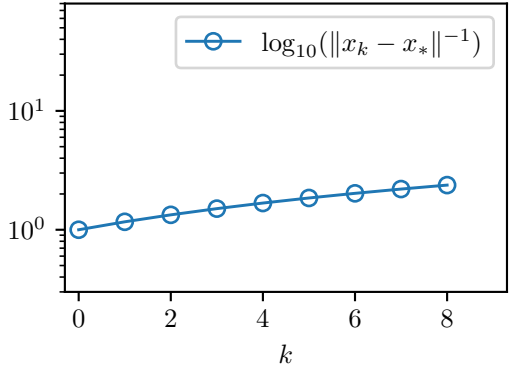
Newton's method ($p = 2$, $\sigma = 0$)



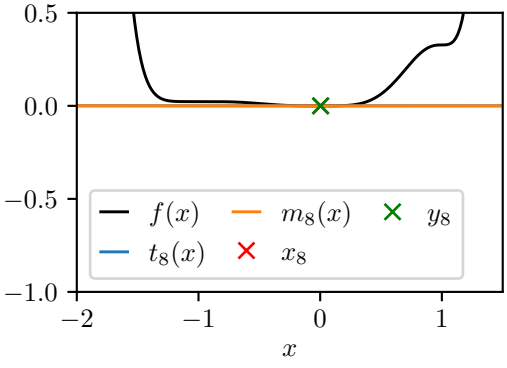


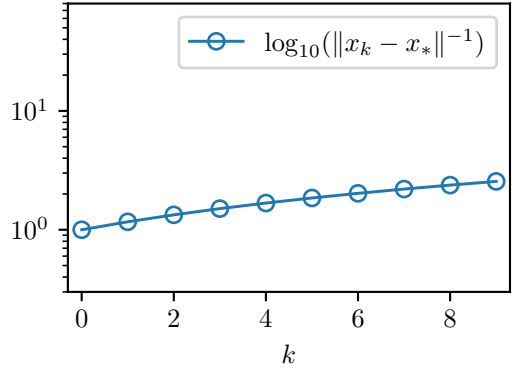
Newton's method ($p = 2$, $\sigma = 0$)



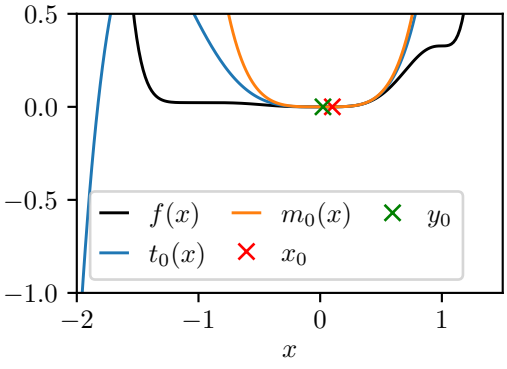


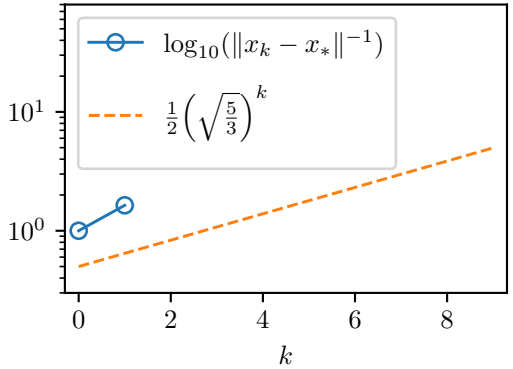
Newton's method ($p = 2$, $\sigma = 0$)



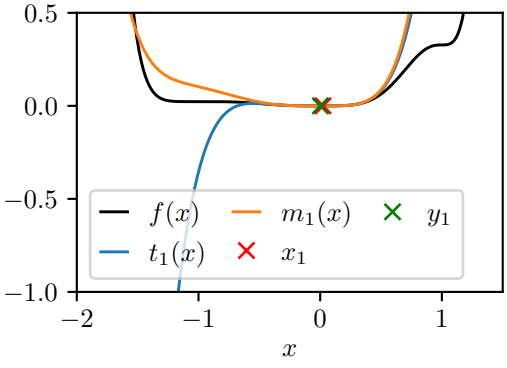


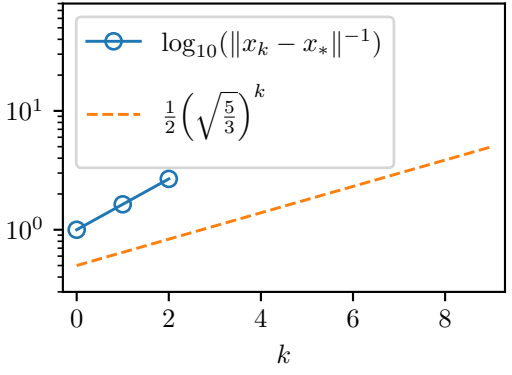
AR5, adaptive σ_k and global model minimizer



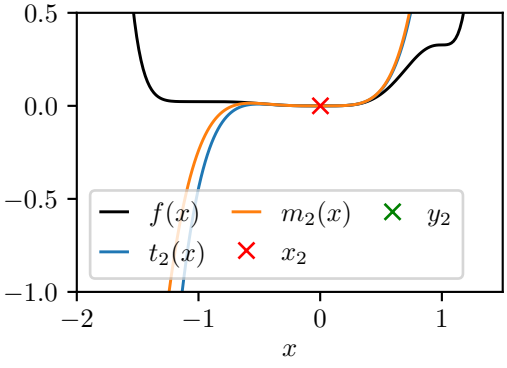


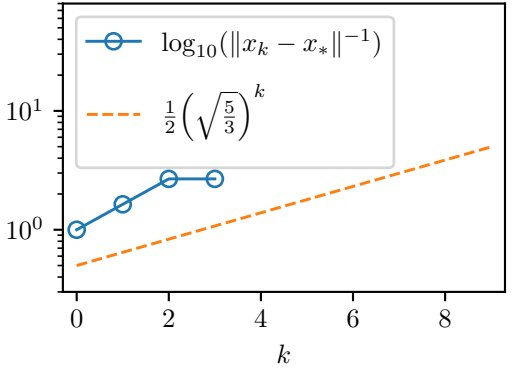
AR5, adaptive σ_k and global model minimizer



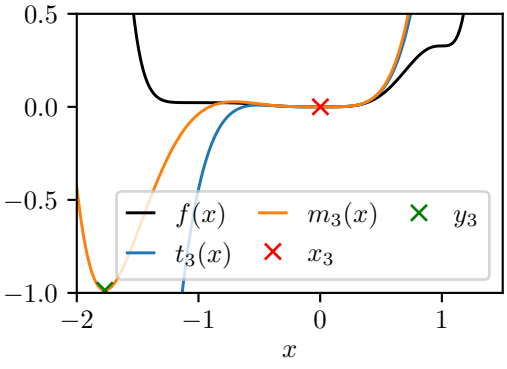


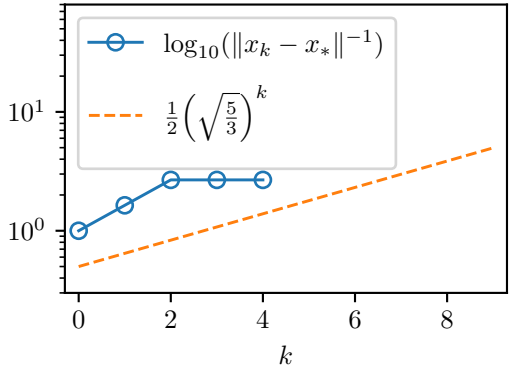
AR5, adaptive σ_k and global model minimizer



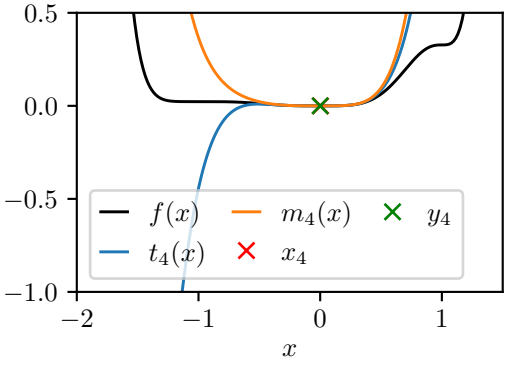


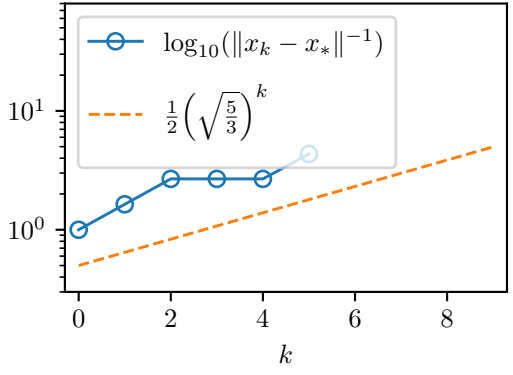
AR5, adaptive σ_k and global model minimizer



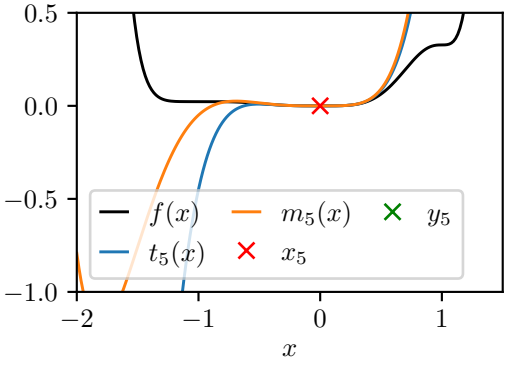


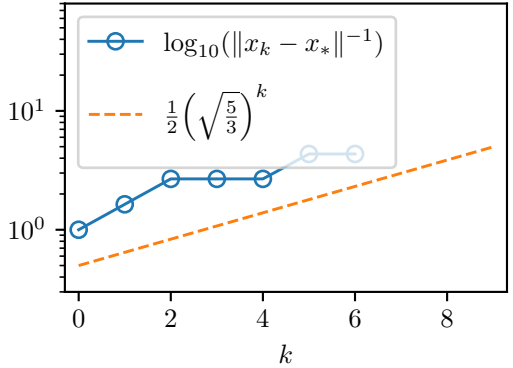
AR5, adaptive σ_k and global model minimizer



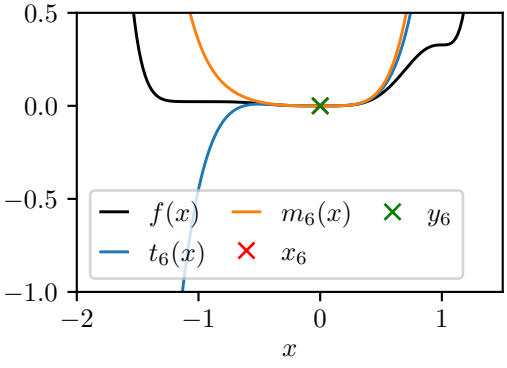


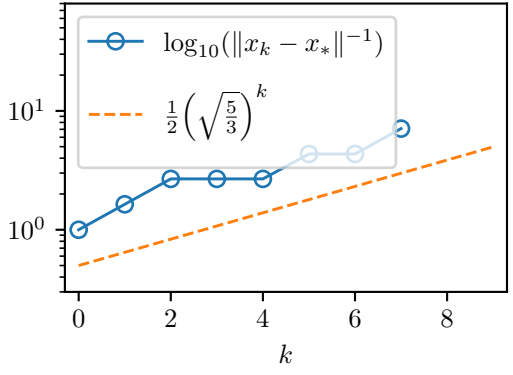
AR5, adaptive σ_k and global model minimizer



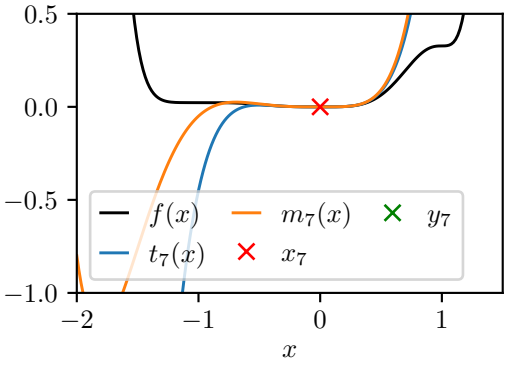


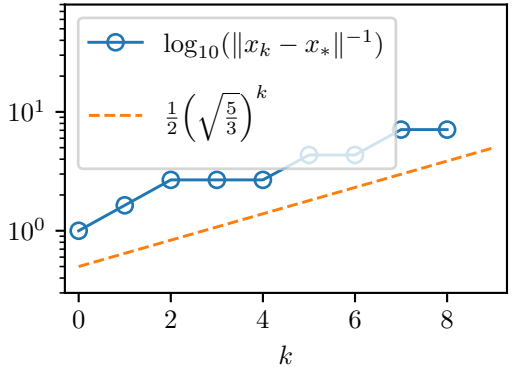
AR5, adaptive σ_k and global model minimizer



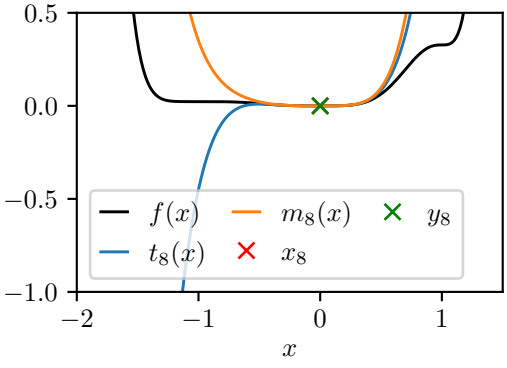


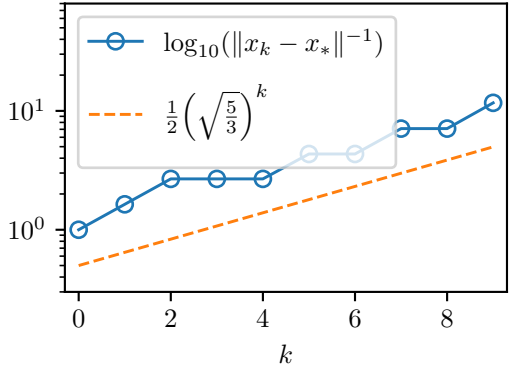
AR5, adaptive σ_k and global model minimizer



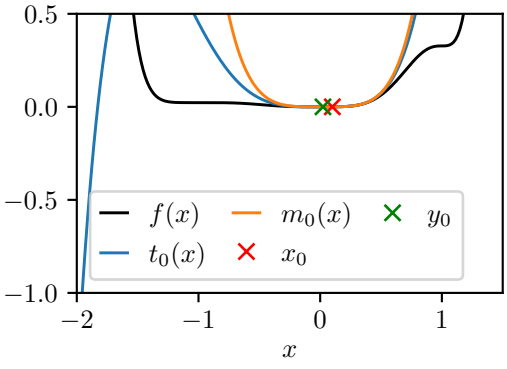


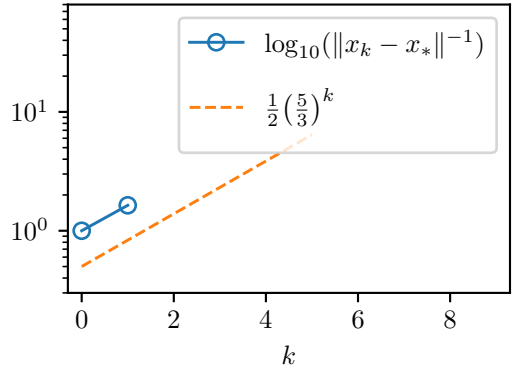
AR5, adaptive σ_k and global model minimizer



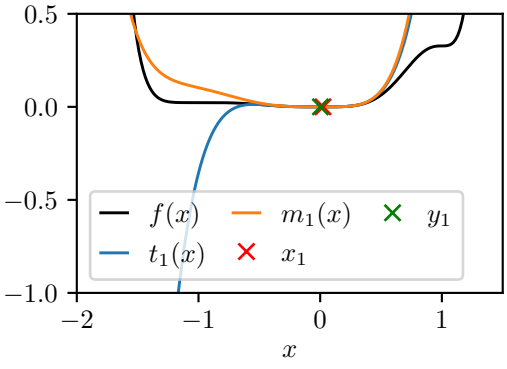


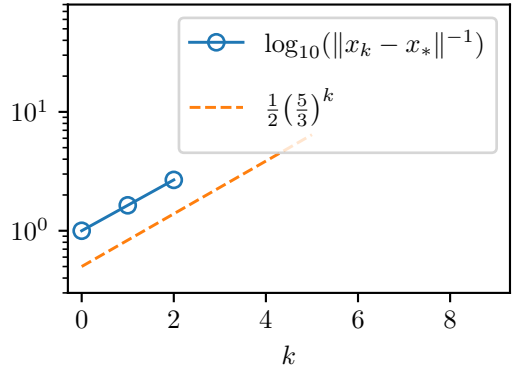
AR5, adaptive σ_k and right model minimizer



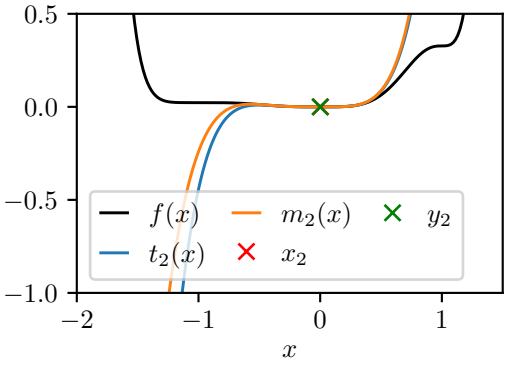


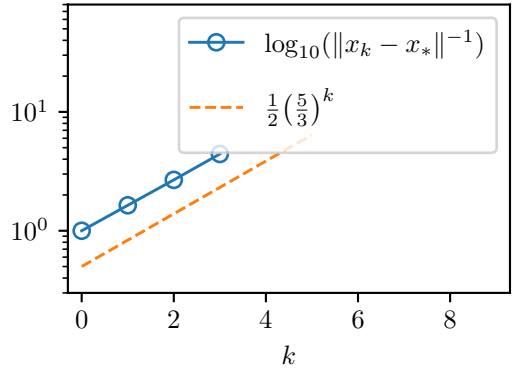
AR5, adaptive σ_k and right model minimizer



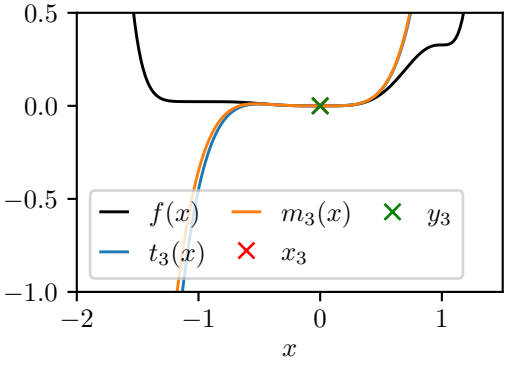


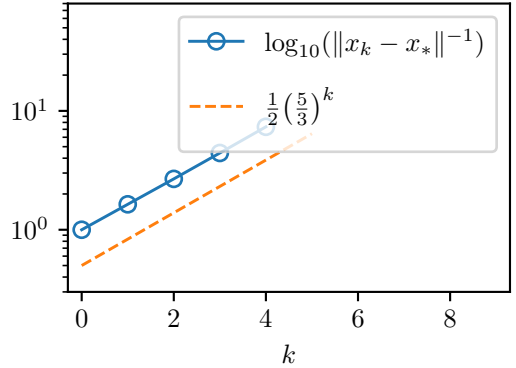
AR5, adaptive σ_k and right model minimizer



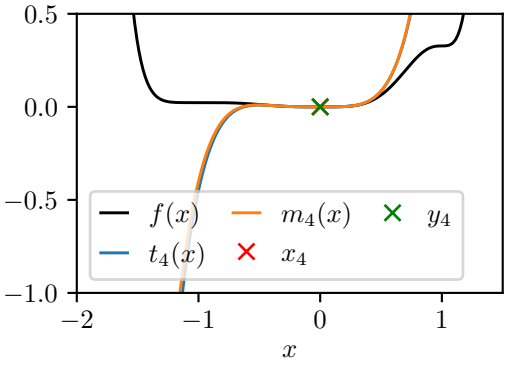


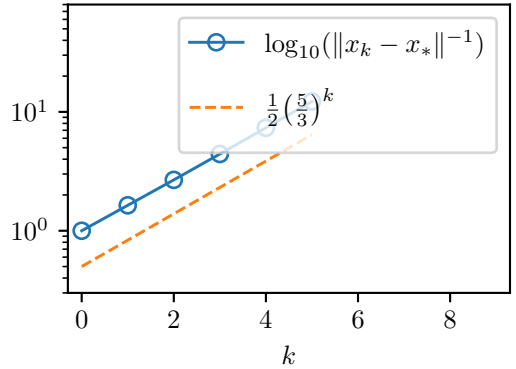
AR5, adaptive σ_k and right model minimizer





AR5, adaptive σ_k and right model minimizer





Assumptions

- $\boldsymbol{x}_k$ and $\boldsymbol{y}_k$ are generated by ARp (Algorithm 1)
- $\boldsymbol{x}_*$ is a local minimizer of f
- $\nabla^p f$ is Lipschitz continuous with constant L_p
- f is uniformly convex of order q inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$
- $p > q - 1$

Convergence for successful iterations

Lemma

If $\boldsymbol{x}_k, \boldsymbol{y}_k \in \mathcal{B}(\boldsymbol{x}_*, r_\mu)$ and 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

If all iterates stay inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$ and $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$, then

$$\|\nabla f(\boldsymbol{x}_{k_i})\| \to 0 \qquad \text{at a Q-}\tfrac{p}{q-1}\text{th-order rate}$$

$$f(\boldsymbol{x}_{k_i}) \to f(\boldsymbol{x}_*) \qquad \text{at a Q-}\tfrac{p}{q-1}\text{th-order rate}$$

$$\boldsymbol{x}_{k_i} \to \boldsymbol{x}_* \qquad \text{at an R-}\tfrac{p}{q-1}\text{th-order rate}$$

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

Convergence for worst-case model minimizer

Theorem

If all iterates stay inside $\mathcal{B}(\boldsymbol{x}_*, r_\mu)$ and $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$, then

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

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

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

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

Convergence for close model minimizer

Lemma

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

Theorem

Assume that $\boldsymbol{y}_k \in \mathcal{B}(\boldsymbol{x}_*, r_y)$ in every iteration. If $\boldsymbol{x}_0$ is close enough to $\boldsymbol{x}_*$, then all iterations are successful and

$$\|\nabla f(\boldsymbol{x}_k)\| \to 0 \qquad \text{at a Q-}\tfrac{p}{q-1}\text{th-order rate}$$

$$f(\boldsymbol{x}_k) \to f(\boldsymbol{x}_*) \qquad \text{at a Q-}\tfrac{p}{q-1}\text{th-order rate}$$

$$\boldsymbol{x}_k \to \boldsymbol{x}_* \qquad \text{at an R-}\tfrac{p}{q-1}\text{th-order rate.}$$

Conclusion

- For tensor methods ($p \geq 3$) 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

References I

- Cartis, Coralia, Nicholas I. M. Gould, and Philippe L. Toint (Jan. 2020). "Sharp Worst-Case Evaluation Complexity Bounds for Arbitrary-Order Nonconvex Optimization with Inexpensive Constraints". In: *SIAM Journal on Optimization* 30.1, pp. 513–541. ISSN: 1052-6234, 1095-7189. DOI: 10.1137/17M1144854.
- — (Jan. 2022). *Evaluation Complexity of Algorithms for Nonconvex Optimization: Theory, Computation and Perspectives*. Philadelphia, PA: Society for Industrial and Applied Mathematics. ISBN: 978-1-61197-698-4 978-1-61197-699-1. DOI: 10.1137/1.9781611976991.
- Doikov, Nikita and Yurii Nesterov (May 2022). "Local Convergence of Tensor Methods". In: *Mathematical Programming* 193.1, pp. 315–336. ISSN: 1436-4646. DOI: 10.1007/s10107-020-01606-x.