

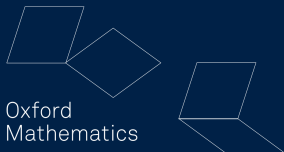
Local convergence of adaptively regularized tensor methods



Mathematical
Institute

KARL WELZEL
Mathematical Institute
University of Oxford

ICCOPT 2025, 22 July 2025



Collaboration with



Raphael Hauser



Yang Liu



Coralia Cartis

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

Algorithm 1.1: Adaptive regularization algorithm using up to p th derivatives

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

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

Assumptions:

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

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

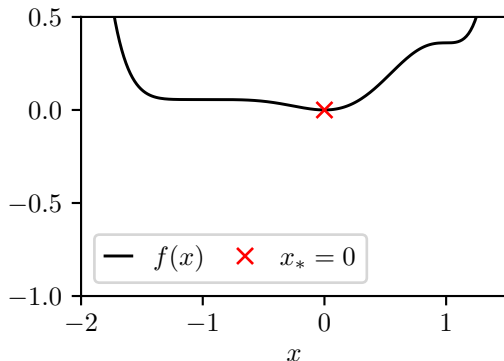
- $f(\mathbf{x}_k) \rightarrow f^*$ at a $\frac{p}{q-1}$ th-order rate and
- $\nabla f(\mathbf{x}_k) \rightarrow \mathbf{0}$ at a $\frac{p}{q-1}$ th-order rate.

[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 $f(\mathbf{x}) - f^* < \varepsilon$)

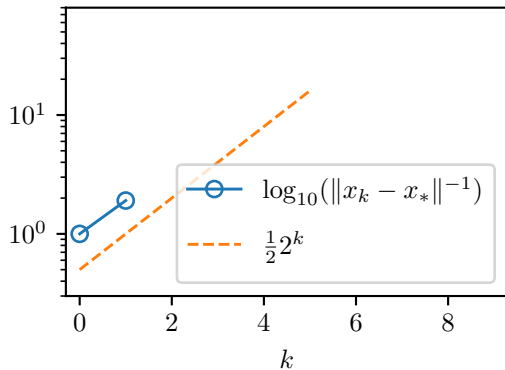
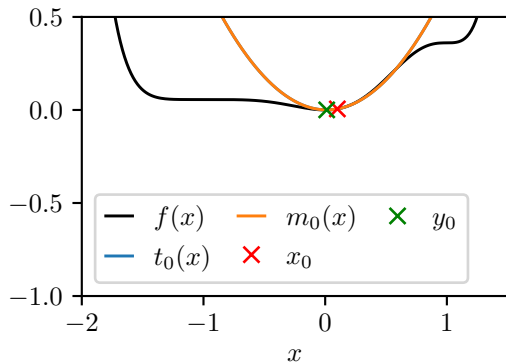
- $p < q - 1$ ($\nu > \mu$): sublinear, $O(\varepsilon^{-\frac{\nu-\mu}{\mu}})$ iterations
- $p = q - 1$ ($\nu = \mu$): linear, $O(\log(\varepsilon^{-1}))$ iterations
- $p > q - 1$ ($\nu < \mu$): superlinear, $O(\log(\log(\varepsilon^{-1})))$ iterations

Example with non-degenerate minimizer

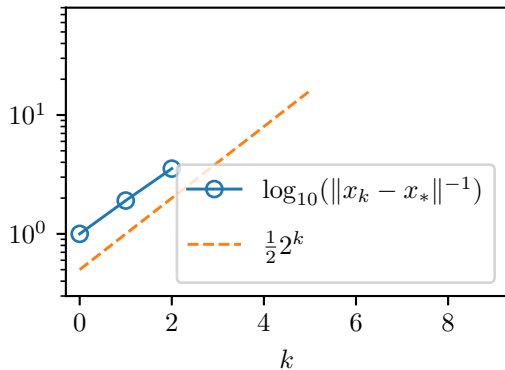
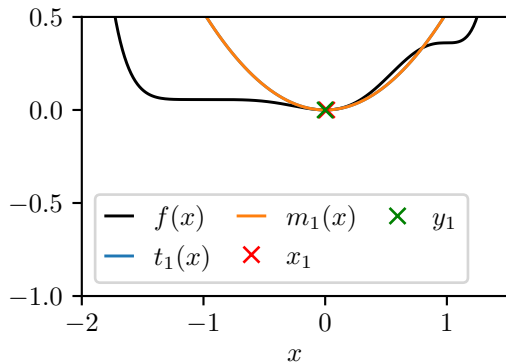
- $f(x) = \int_0^x (t+1)^4(t-1)^2 t \, dt$
- f has one local (and global) minimizer at $x_* = 0$
- f is not globally convex, but locally strongly convex: $f''(x_*) = 1$



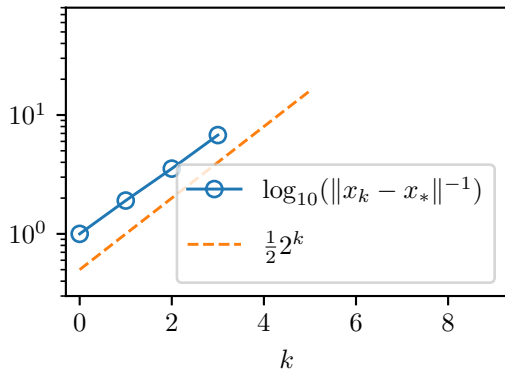
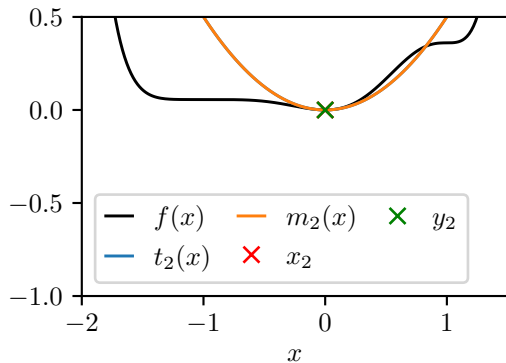
Newton's method ($p = 2, \sigma = 0$)



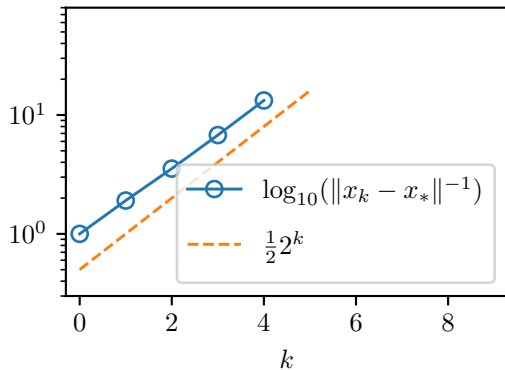
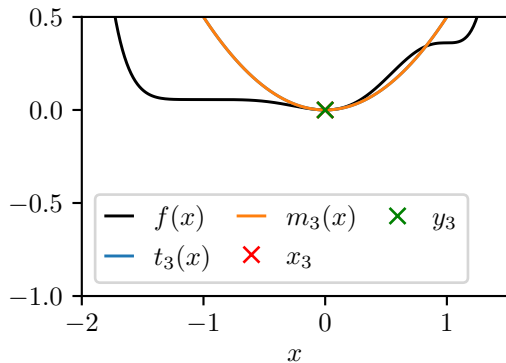
Newton's method ($p = 2, \sigma = 0$)



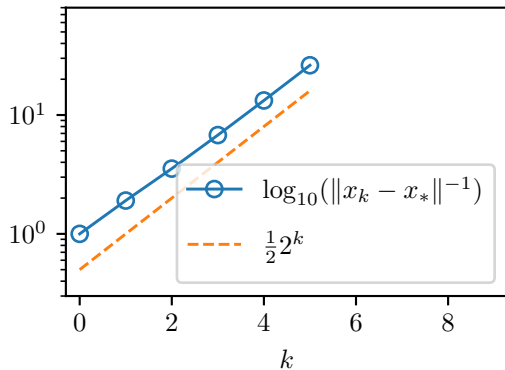
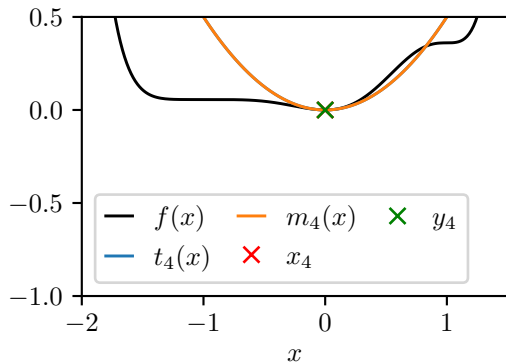
Newton's method ($p = 2, \sigma = 0$)



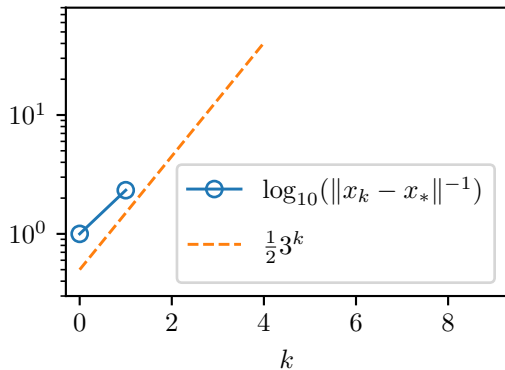
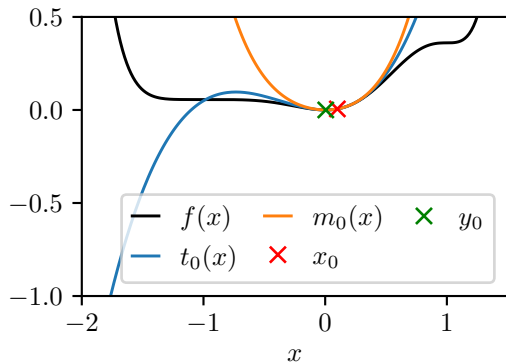
Newton's method ($p = 2, \sigma = 0$)

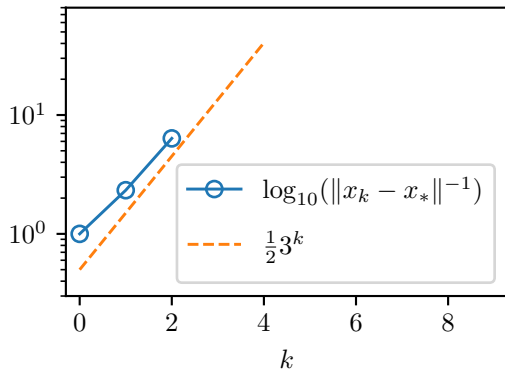
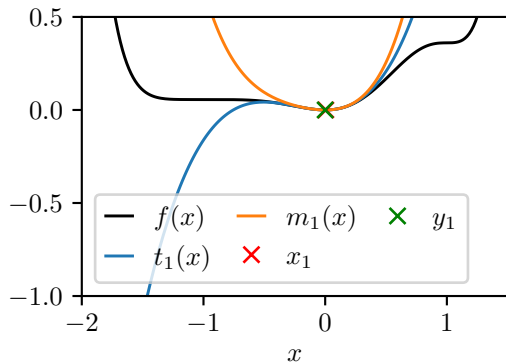


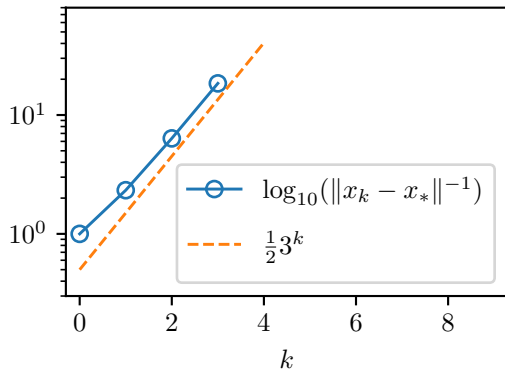
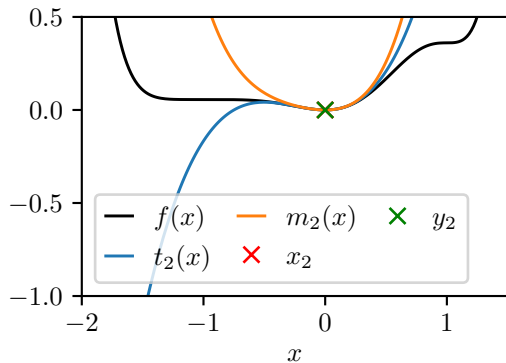
Newton's method ($p = 2, \sigma = 0$)

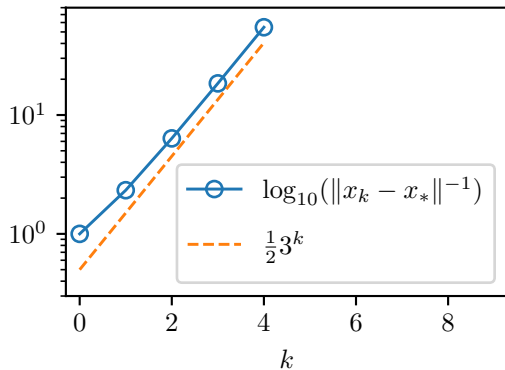
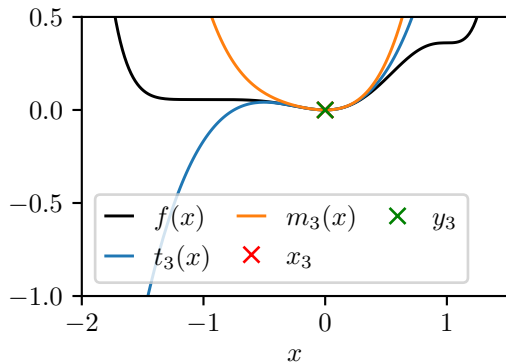


AR3, constant σ_k

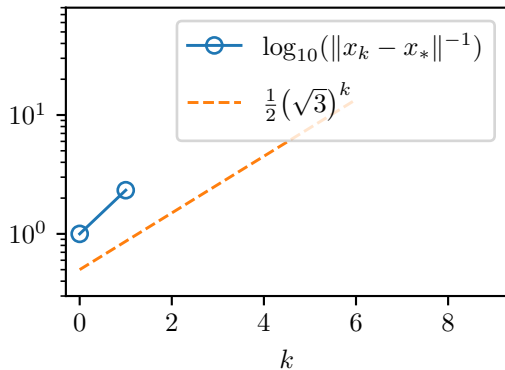
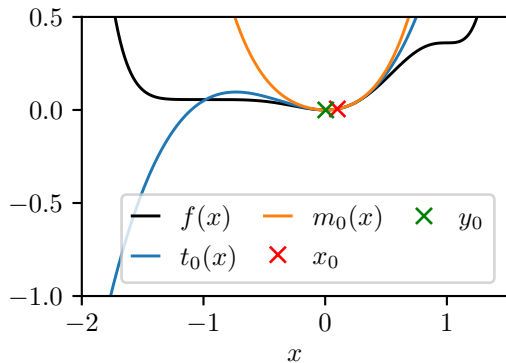




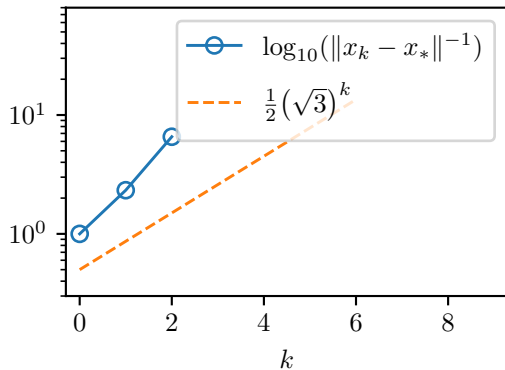
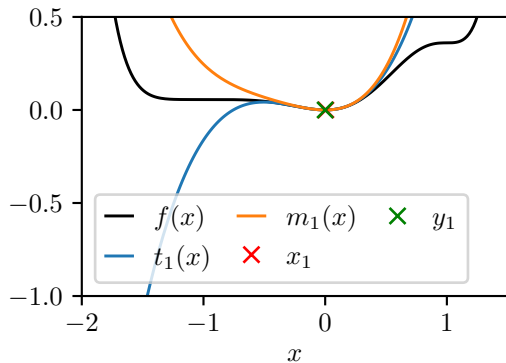




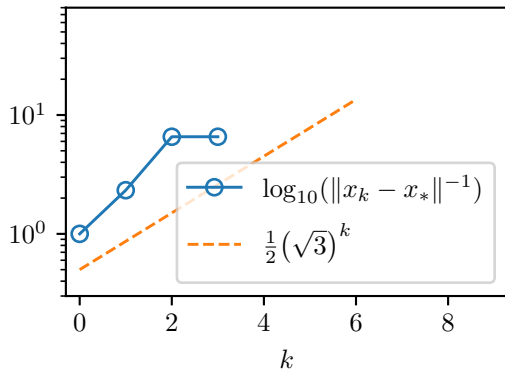
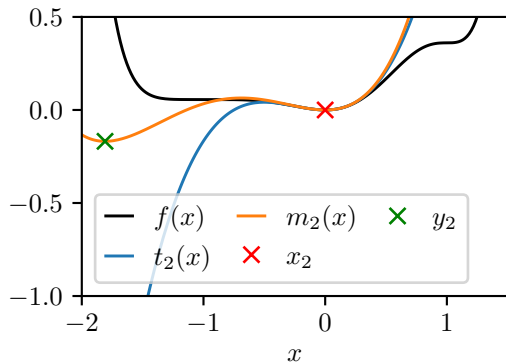
AR3, adaptive σ_k and global model minimizer



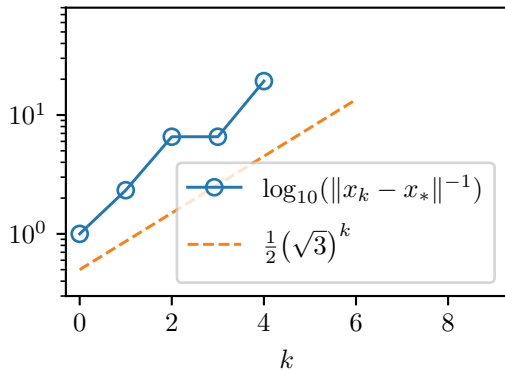
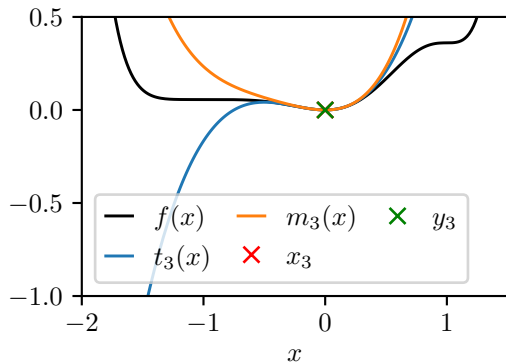
AR3, adaptive σ_k and global model minimizer



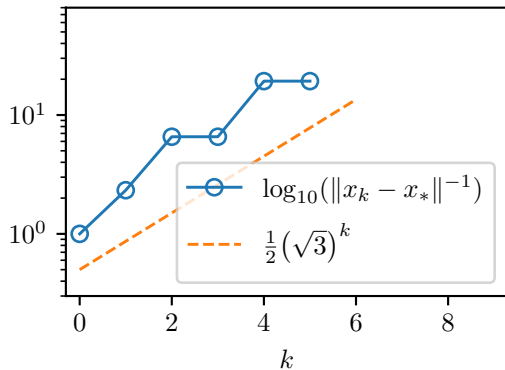
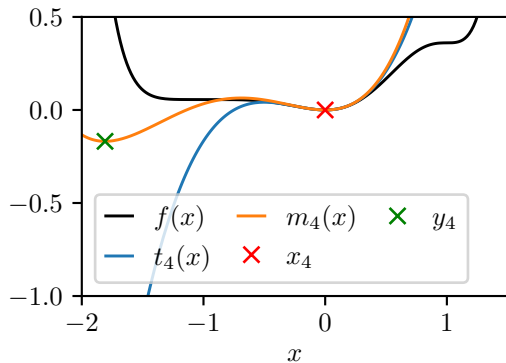
AR3, adaptive σ_k and global model minimizer



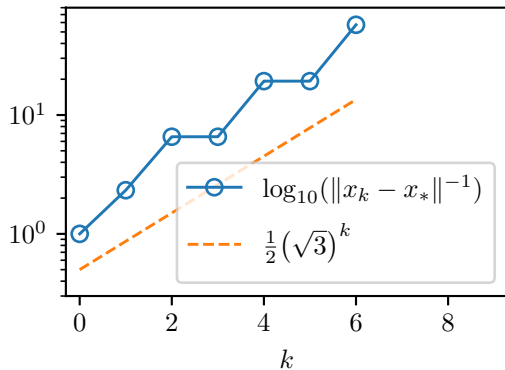
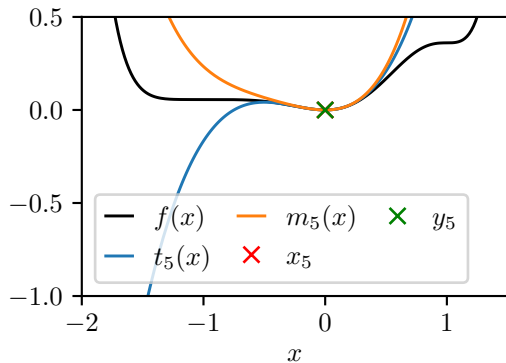
AR3, adaptive σ_k and global model minimizer



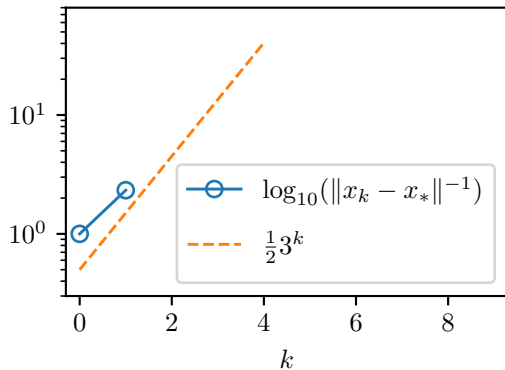
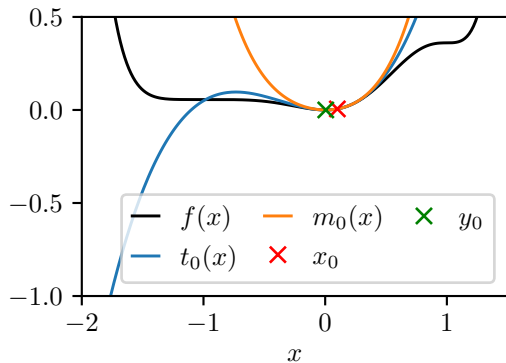
AR3, adaptive σ_k and global model minimizer



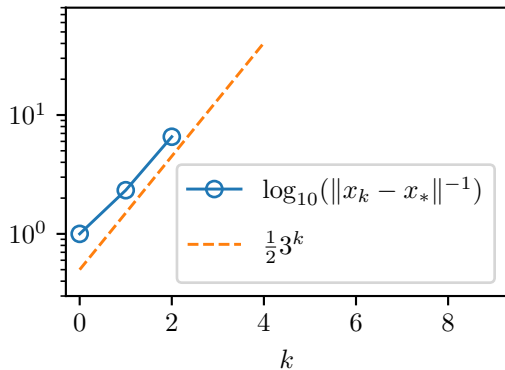
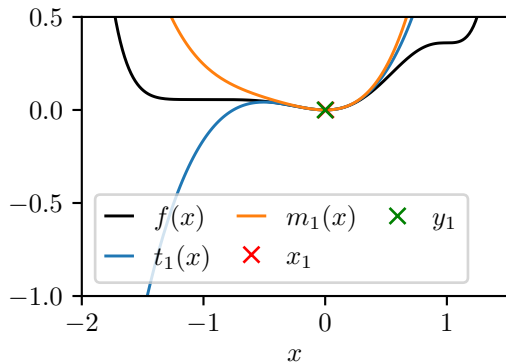
AR3, adaptive σ_k and global model minimizer



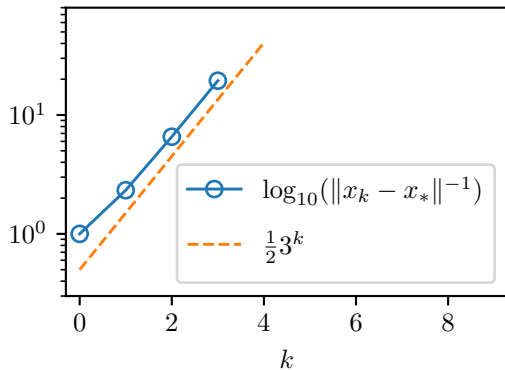
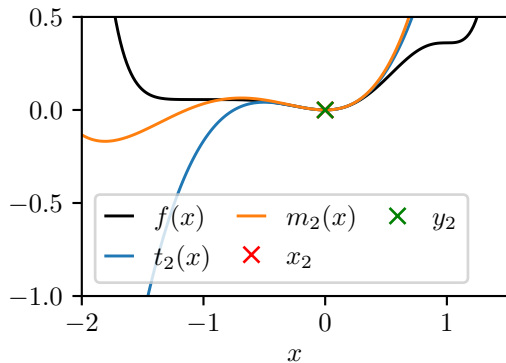
AR3, adaptive σ_k and right model minimizer



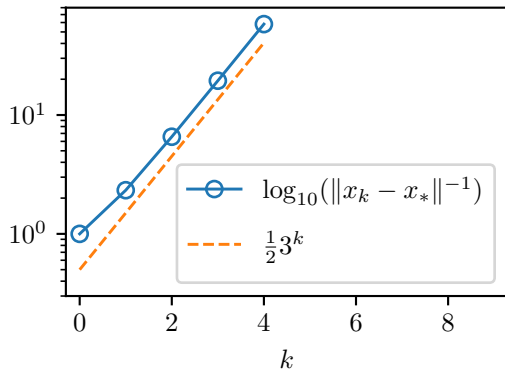
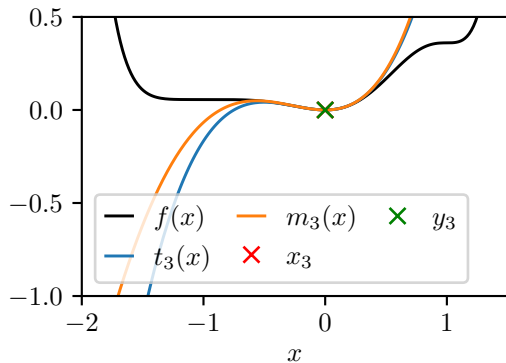
AR3, adaptive σ_k and right model minimizer



AR3, adaptive σ_k and right model minimizer

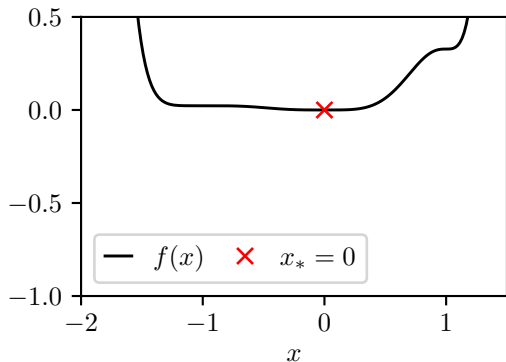


AR3, adaptive σ_k and right model minimizer

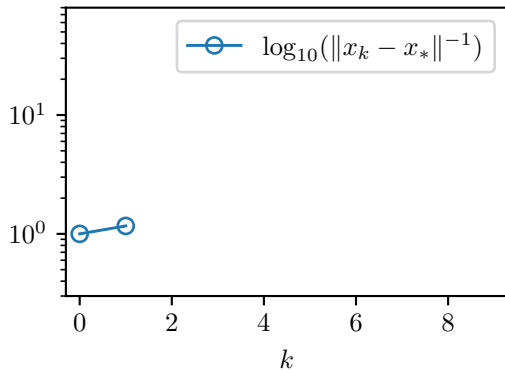
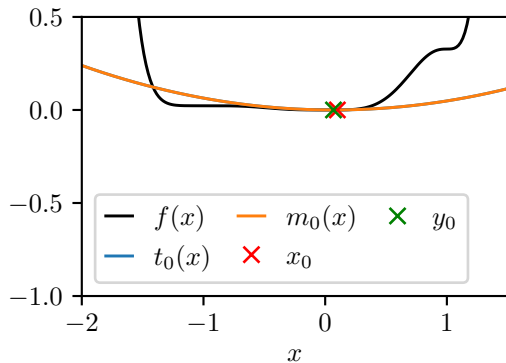


Example with degenerate minimizer

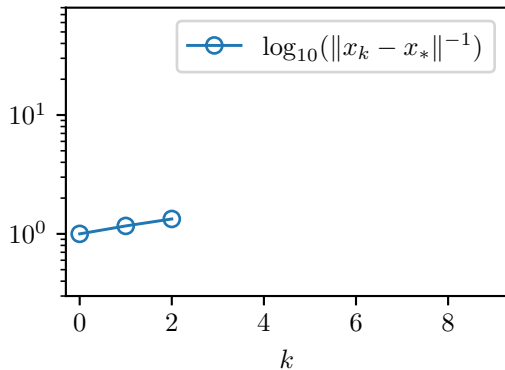
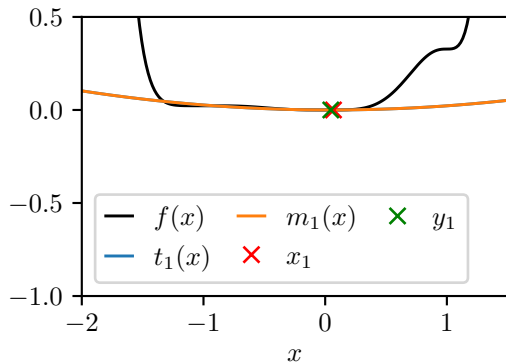
- $f(x) = \int_0^x 3(t+1)^4(t-1)^2 t^3 dt$
- f has one local (and global) minimizer at $x_* = 0$
- f is locally uniformly convex around x_* with $q = 4$:
 - $f''(x_*) = 0$
 - $f'''(x_*) = 0$
 - $f^{(4)}(x_*) = 18$



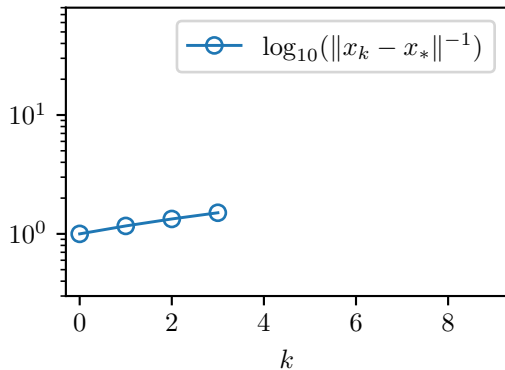
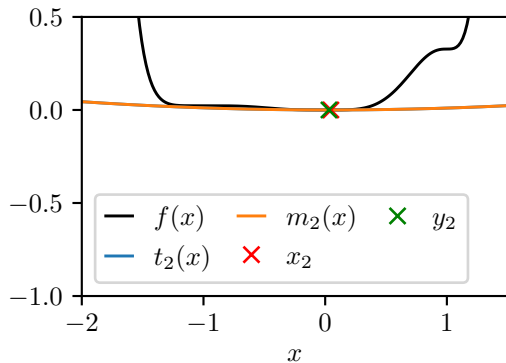
Newton's method ($p = 2, \sigma = 0$)



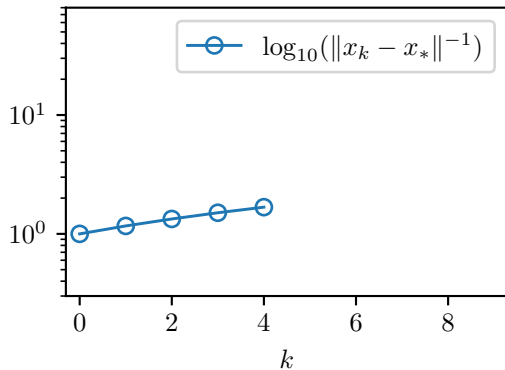
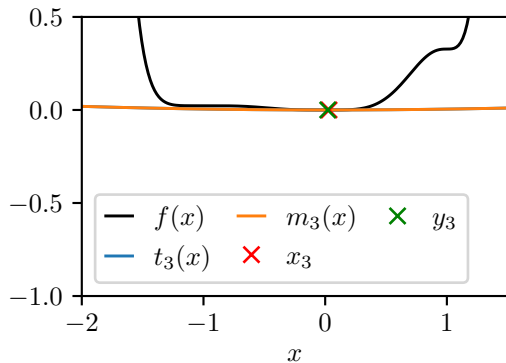
Newton's method ($p = 2, \sigma = 0$)



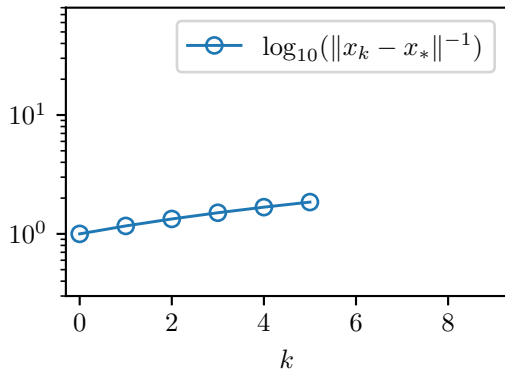
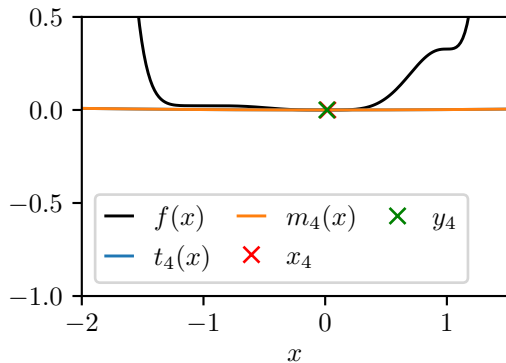
Newton's method ($p = 2, \sigma = 0$)



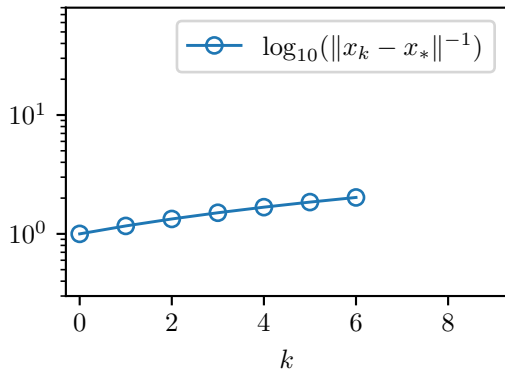
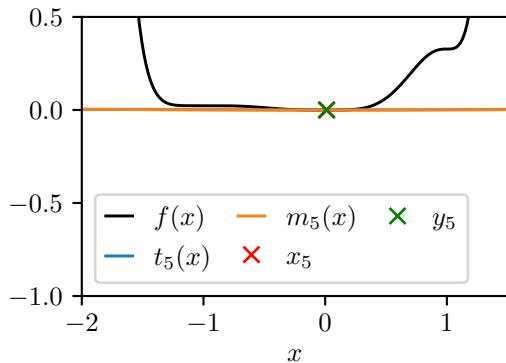
Newton's method ($p = 2, \sigma = 0$)



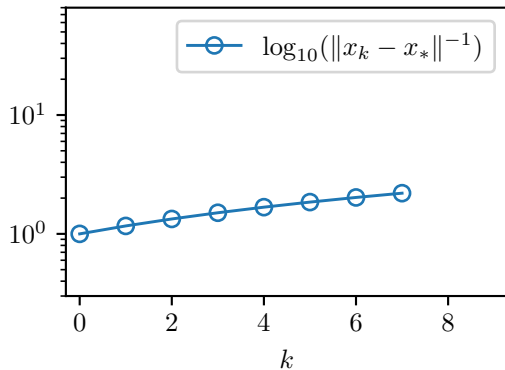
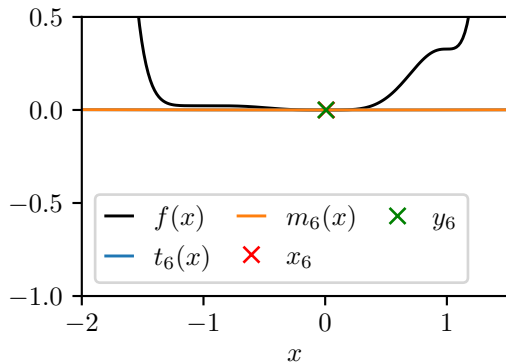
Newton's method ($p = 2, \sigma = 0$)



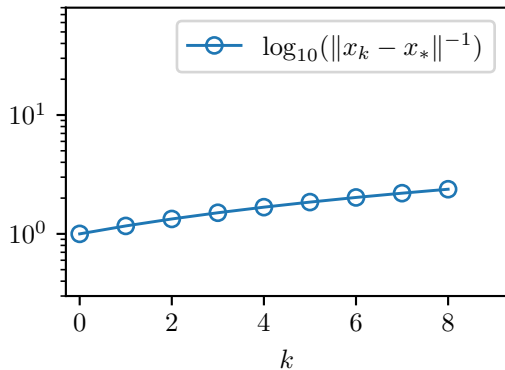
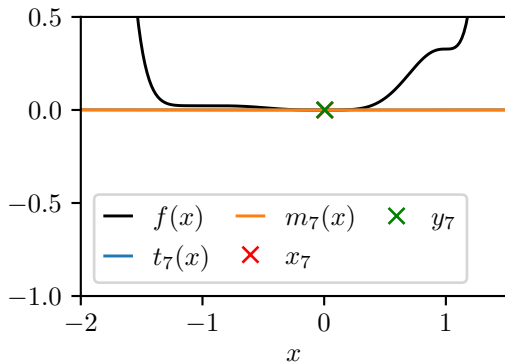
Newton's method ($p = 2, \sigma = 0$)



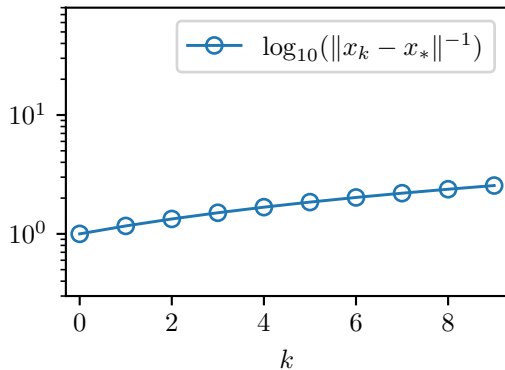
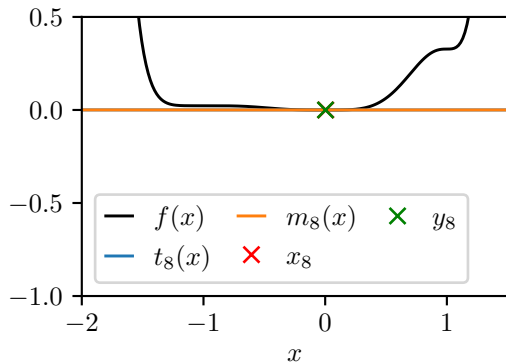
Newton's method ($p = 2, \sigma = 0$)



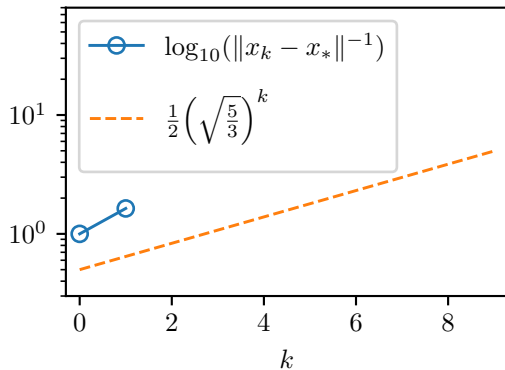
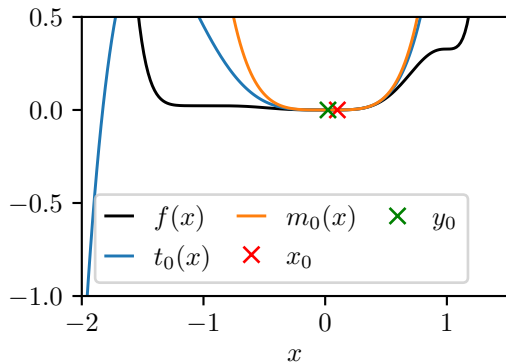
Newton's method ($p = 2, \sigma = 0$)



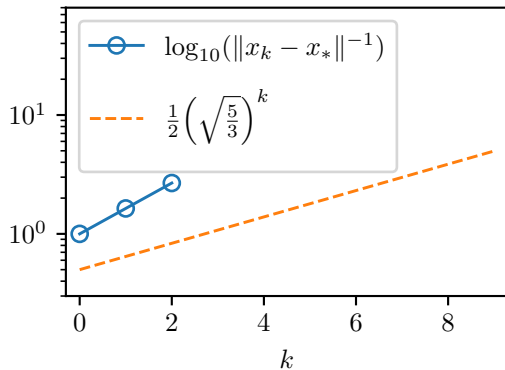
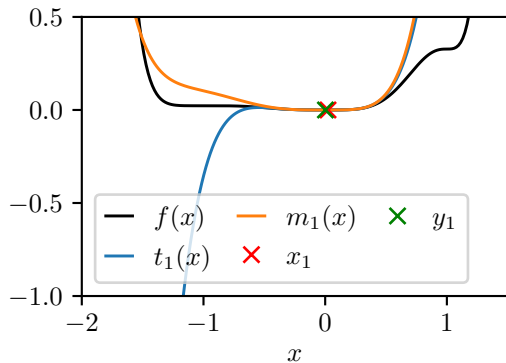
Newton's method ($p = 2, \sigma = 0$)



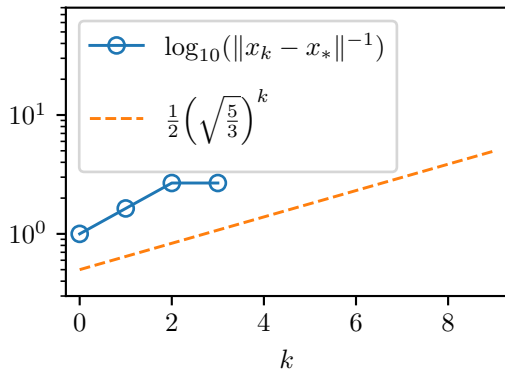
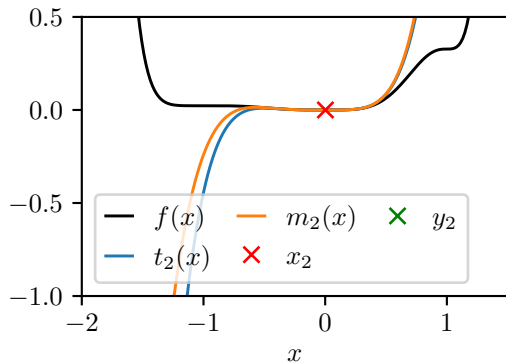
AR5, adaptive σ_k and global model minimizer



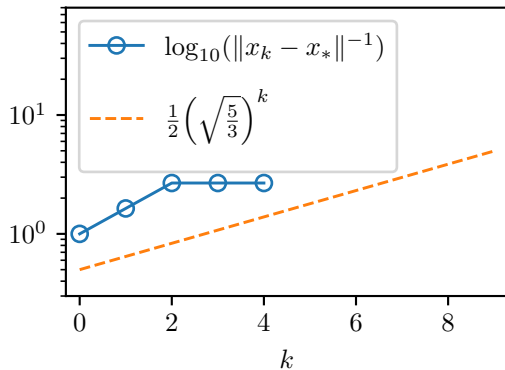
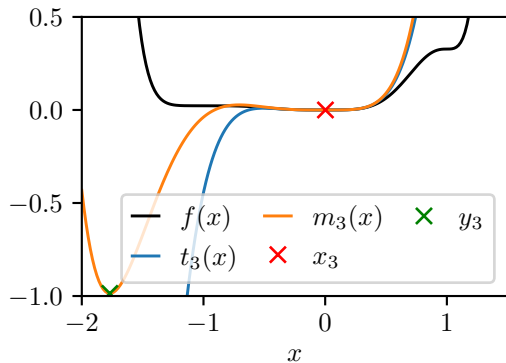
AR5, adaptive σ_k and global model minimizer



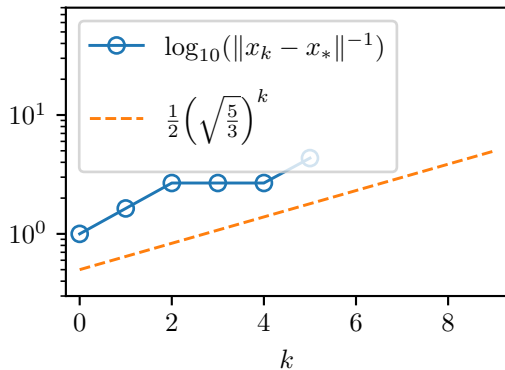
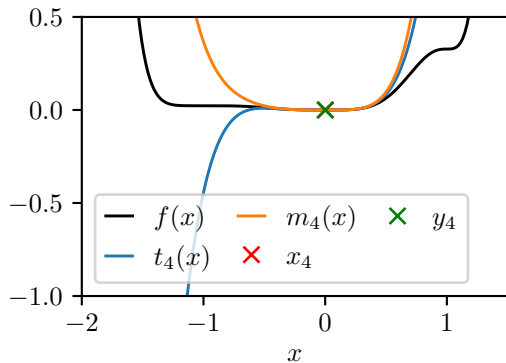
AR5, adaptive σ_k and global model minimizer



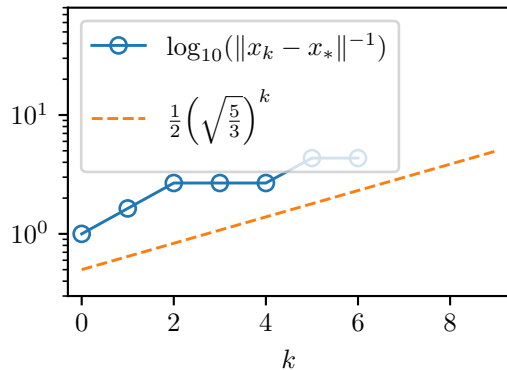
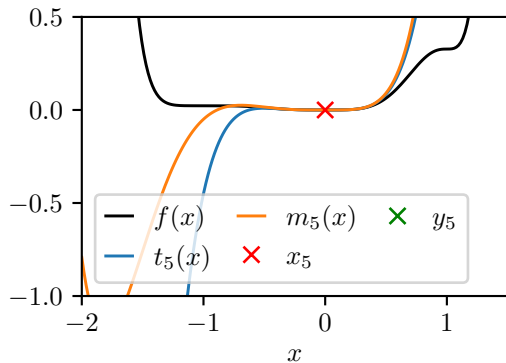
AR5, adaptive σ_k and global model minimizer



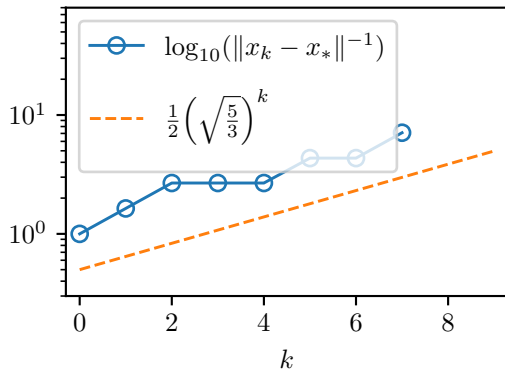
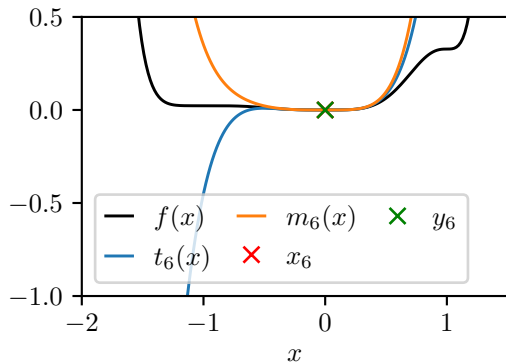
AR5, adaptive σ_k and global model minimizer



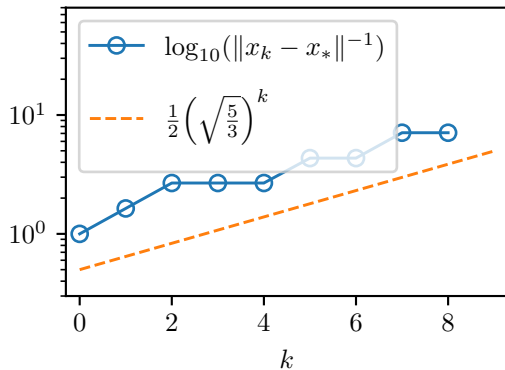
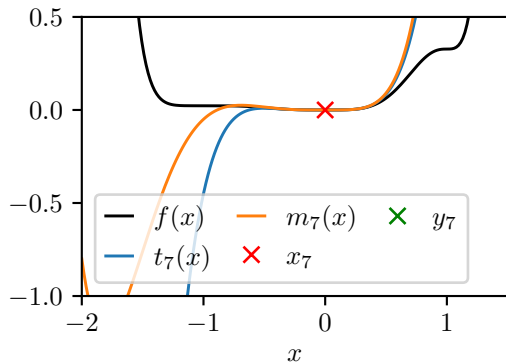
AR5, adaptive σ_k and global model minimizer



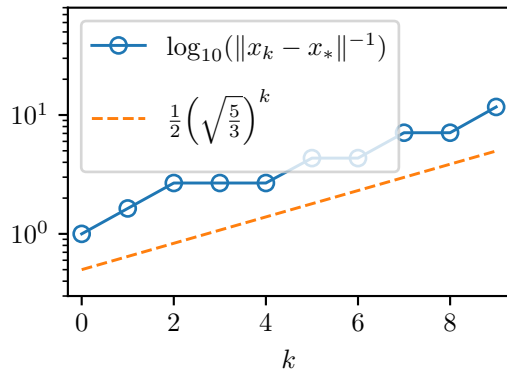
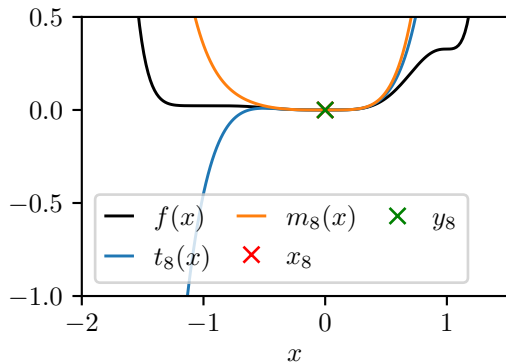
AR5, adaptive σ_k and global model minimizer



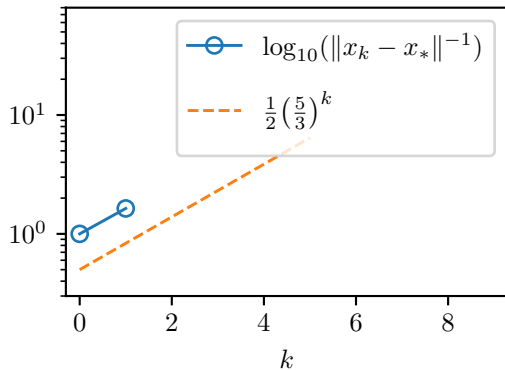
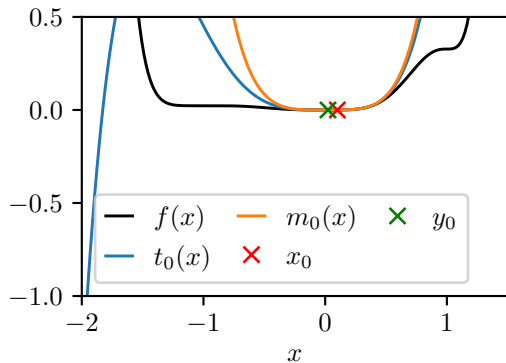
AR5, adaptive σ_k and global model minimizer



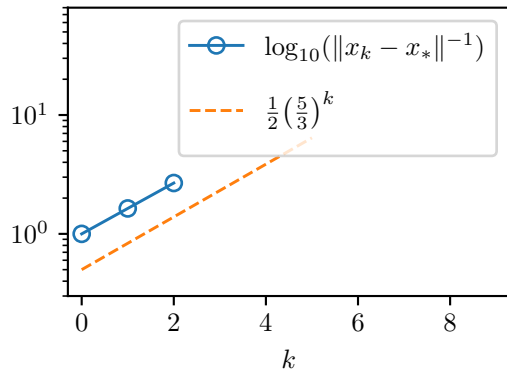
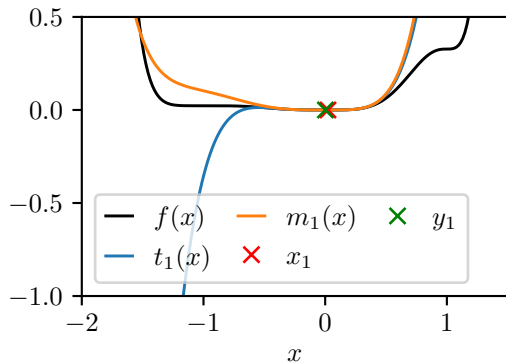
AR5, adaptive σ_k and global model minimizer



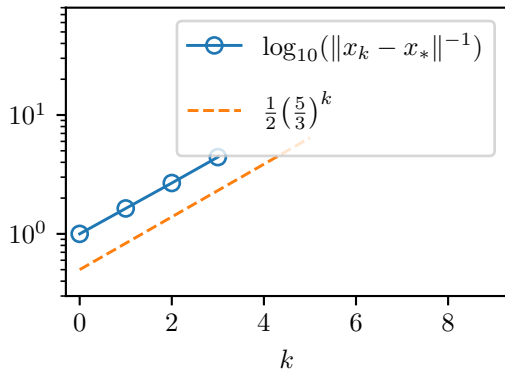
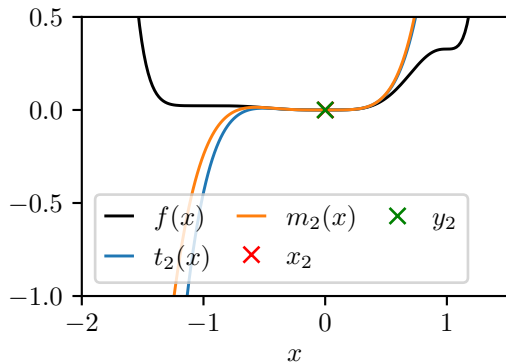
AR5, adaptive σ_k and right model minimizer



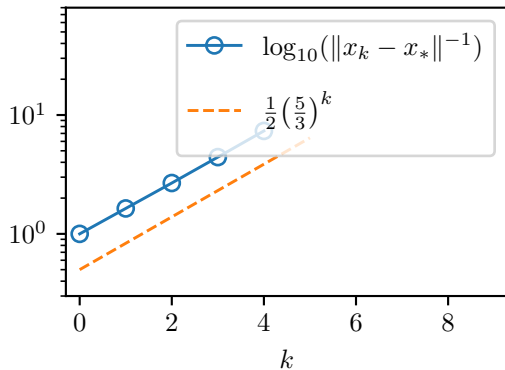
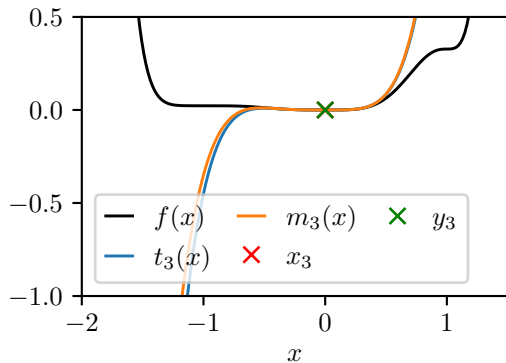
AR5, adaptive σ_k and right model minimizer



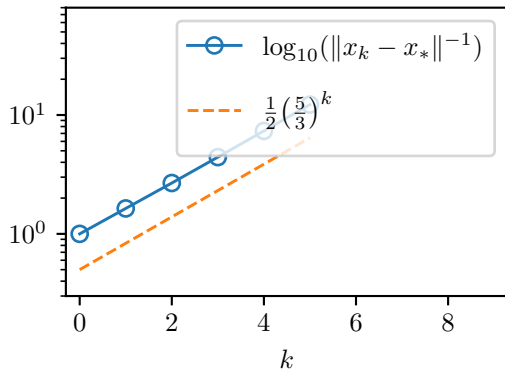
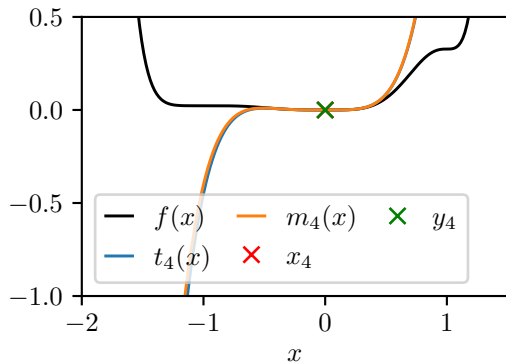
AR5, adaptive σ_k and right model minimizer



AR5, adaptive σ_k and right model minimizer



AR5, adaptive σ_k and right model minimizer



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

Lemma

If $\mathbf{x}_k, \mathbf{y}_k \in \mathcal{B}(\mathbf{x}_, r_q)$ and iteration k is successful, then the gradients satisfy*

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

Theorem

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

$$\|\nabla f(\mathbf{x}_{k_i})\| \rightarrow 0$$

at a $Q - \frac{p}{q-1}$ th-order rate

$$f(\mathbf{x}_{k_i}) \rightarrow f(\mathbf{x}_*)$$

at a $Q - \frac{p}{q-1}$ th-order rate

$$\mathbf{x}_{k_i} \rightarrow \mathbf{x}_*$$

at an $R - \frac{p}{q-1}$ th-order rate

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

Theorem

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

$$\|\nabla f(\mathbf{x}_k)\| \rightarrow 0$$

at an $R^{-\alpha+1\sqrt{\frac{p}{q-1}}}$ th-order rate

$$f(\mathbf{x}_k) \rightarrow f(\mathbf{x}_*)$$

at an $R^{-\alpha+1\sqrt{\frac{p}{q-1}}}$ th-order rate

$$\mathbf{x}_k \rightarrow \mathbf{x}_*$$

at an $R^{-\alpha+1\sqrt{\frac{p}{q-1}}}$ th-order rate

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

Lemma

Let $\mathcal{C}_k$ be the connected component of the sublevel set $\{\mathbf{y} \in \mathbb{R}^n \mid m_k(\mathbf{y}) \leq m_k(\mathbf{x}_k)\}$ that contains $\mathbf{x}_k$. There exists a radius $r_c > 0$ such that for any $\mathbf{x}_k \in B(\mathbf{x}_, r_c)$ there exists a local minimizer m_k inside $\mathcal{C}_k$ and any (approximate) stationary point $\mathbf{y}_k$ of m_k inside $\mathcal{C}_k$ gives a successful iteration.*

Theorem

Assume that $\mathbf{y}_k \in \mathcal{C}_k$ at every iteration. If $\mathbf{x}_0$ is close enough to $\mathbf{x}_$, then all iterations are successful and*

$$\|\nabla f(\mathbf{x}_k)\| \rightarrow 0$$

at a $Q - \frac{p}{q-1}$ th-order rate

$$f(\mathbf{x}_k) \rightarrow f(\mathbf{x}_*)$$

at a $Q - \frac{p}{q-1}$ th-order rate

$$\mathbf{x}_k \rightarrow \mathbf{x}_*$$

at an $R - \frac{p}{q-1}$ th-order rate.

- 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

Global analysis including subproblem $\rightarrow$ Wenqi Zhu's talk @ Wed 4:40pm (Session 9K)

Implementation of tensor methods $\rightarrow$ Yang Liu's talk @ Thu 4:40pm (Session 12N)

-  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.

Theorem

If x_0 is close enough to x_ (depending on σ_0), then either all iterates stay inside $\mathcal{B}(x_*, r_q)$ or the first iterate outside this ball satisfies $f(x_k) < f(x_*)$.*