Mathematical Institute

Adaptive Levenberg-Marquardt Third-Order Newton's Method

Yubo Cai [1] [2]

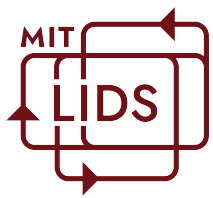
Zardini Lab

[1]Laboratory for Information and Decision Systems, Massachusetts Institute of Technology

[2]Mathematical Institute, University of Oxford

31st LIDS Student Conference, January 29, 2026

Civil and Environmental Engineering

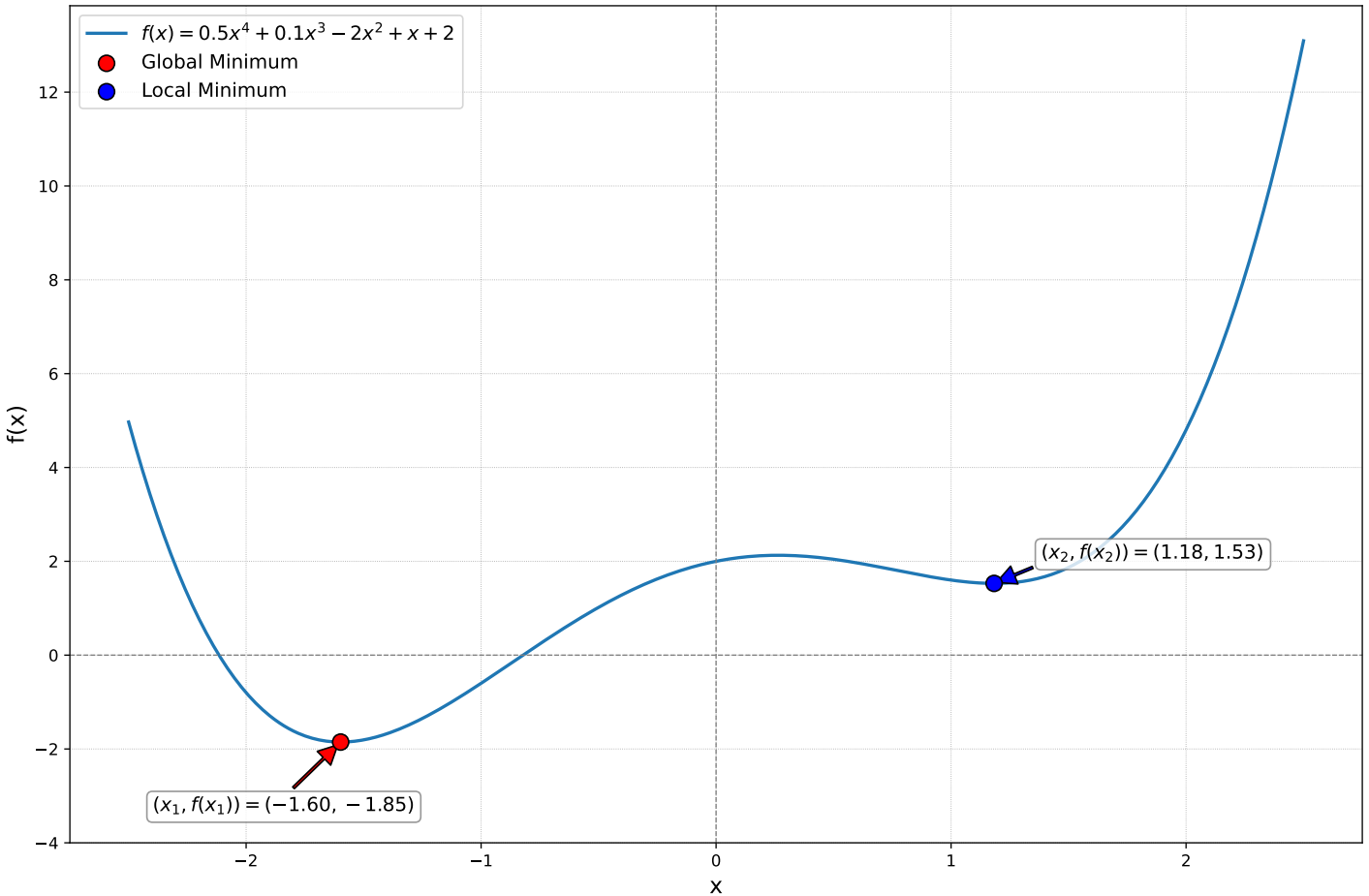


Background: Unconstrained Nonconvex Optimization

$$\text{minimize } f(x) \quad \text{subject to } x \in \mathbb{R}^n. \tag{UP}$$

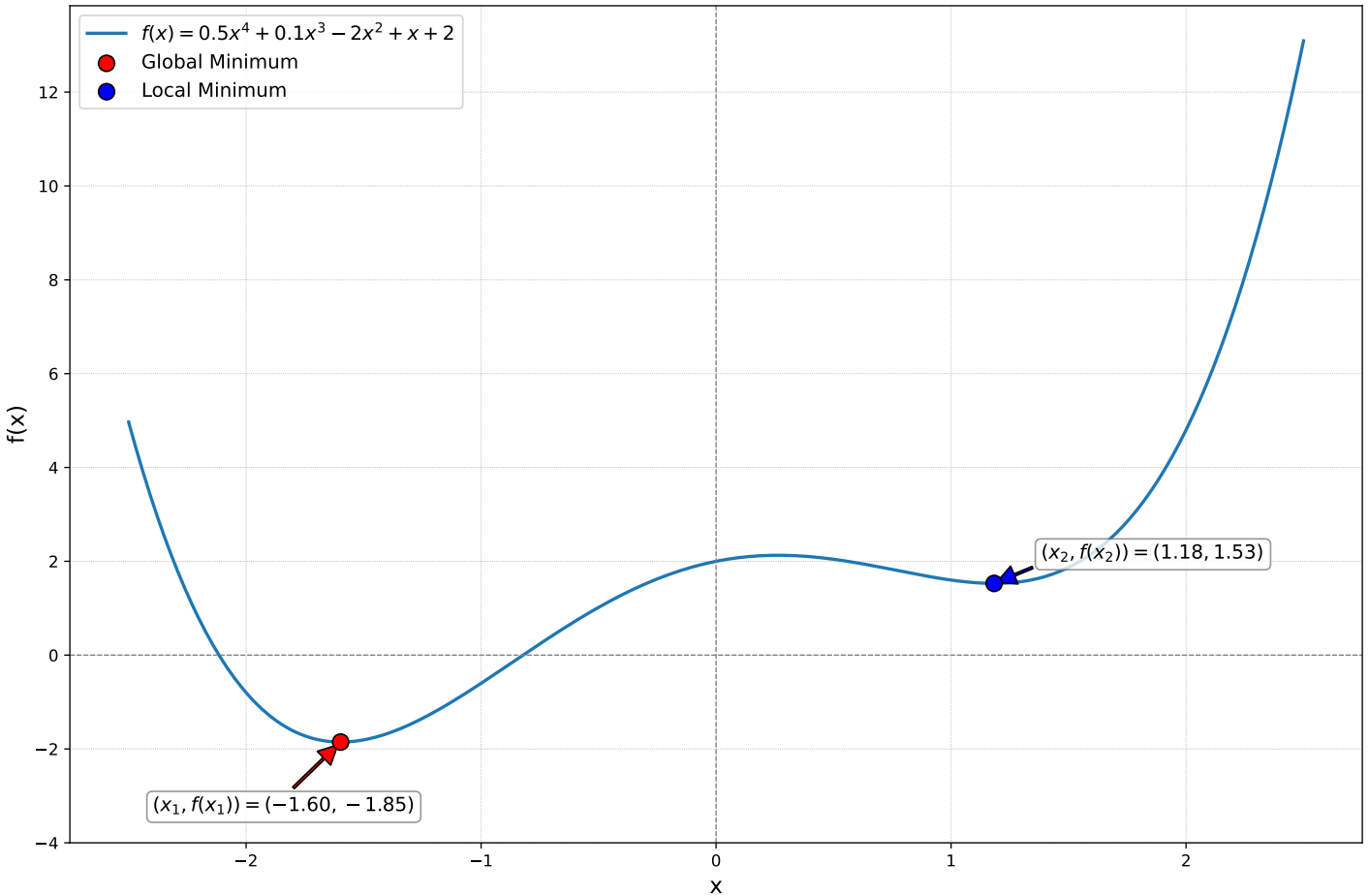
Background: Unconstrained Nonconvex Optimization

$$\text{minimize } f(x) \quad \text{subject to } x \in \mathbb{R}^n. \tag{UP}$$



Background: Unconstrained Nonconvex Optimization

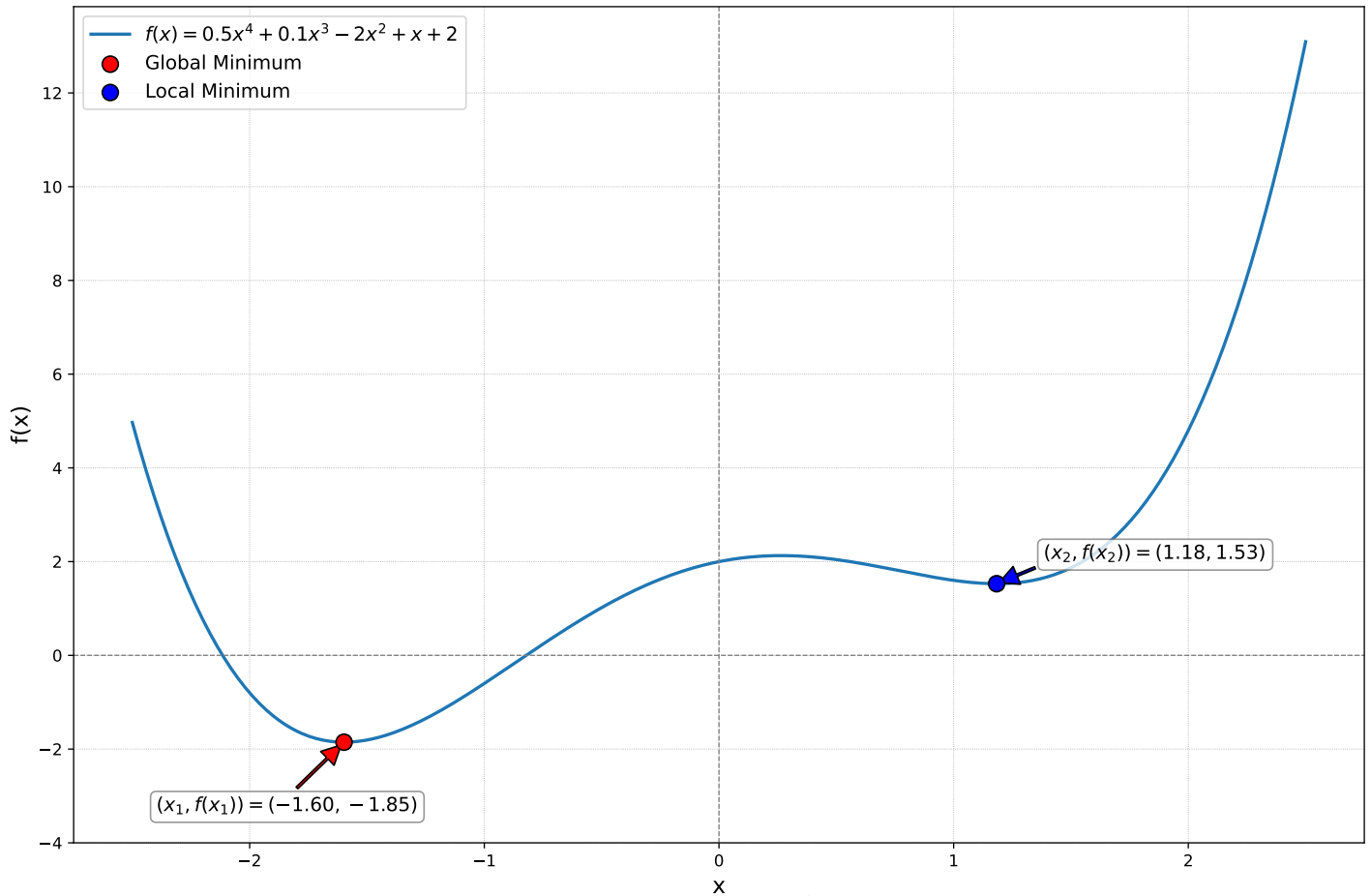
$$\text{minimize } f(x) \quad \text{subject to } x \in \mathbb{R}^n. \tag{UP}$$



x^* **global minimizer** of f (over $\mathbb{R}^n$) $\iff f(x) \geq f(x^*), \forall x \in \mathbb{R}^n$. Here is x_1.

Background: Unconstrained Nonconvex Optimization

$$\text{minimize } f(x) \quad \text{subject to } x \in \mathbb{R}^n. \tag{UP}$$



x^* **local minimizer** of f (over $\mathbb{R}^n$) $\iff$ there exists $\mathcal{N}(x^*, \delta)$ such that $f(x) \geq f(x^*)$, for all $x \in \mathcal{N}(x^*, \delta)$, where $\mathcal{N}(x^*, \delta) := \{x \in \mathbb{R}^n : \|x - x^*\| \leq \delta\}$ and $\|\cdot\|$ is the Euclidean norm. Here is x_2.

Motivation: Applications of Unconstrained Nonconvex Optimization

- **Deep Learning:** Train neural networks by minimizing nonconvex loss functions
  - Saddle points, local minima, high dimensionality
- **Phase Retrieval:** Recover signals from magnitude-only measurements
  - Phase information lost, multiple equivalent solutions
- **Portfolio Optimization:** Optimize asset allocation with transaction costs
  - Non-smooth costs, risk constraints, cardinality limits

Common Thread: All require efficient algorithms for nonconvex optimization with global convergence guarantees.

Background: Iterative Methods

A Generic Method (GM)

1: **Input:** $\epsilon > 0$ (tolerance), $x_0 \in \mathbb{R}^n$ (initial point)
2: **while** $\|\nabla f(x_k)\| > \epsilon$ **do**
3: Build local model $m_{f,x_k}(x)$ around x_k (Taylor expansion)
4: Compute trial point: $\bar{x} \leftarrow \arg\min_x \ m_{f,x_k}(x)$
5: **If** $f(\bar{x}) < f(x_k)$: accept $x_{k+1} \leftarrow \bar{x}$ (descent condition)
6: **Else:** adjust model parameters and retry
7: $k \leftarrow k + 1$
8: **end while**
9: **Output:** x_k with $\|\nabla f(x_k)\| \leq \epsilon$

- **Global convergence:** For any $x_0 \in \mathbb{R}^n$: $\nabla f(x_k) \to 0$ as $k \to \infty$. (**REMARK.** Not necessarily converge to the **global** minimizer!)
- **Local convergence:** If x_0 sufficiently close to local minimizer x^*: $x_k \to x^*$.
- **Complexity:** Count the number of iterations to achieve $\|\nabla f(x_k)\| \leq \epsilon$.
- **p-Rate of convergence:** $x_k \to x^*$ with rate $p \geq 1$ if $\exists \rho > 0$, $k_0 \geq 0$ s.t. $\|x_{k+1} - x^*\| \leq \rho \|x_k - x^*\|^p$, $\forall k \geq k_0$.

How to Choose the Local Model m_{f,x_k}?

Key Question: What is the best choice for the local model in iterative methods?

Two Design Choices:

- **Order:** How many derivatives to use? ($p = 1, 2, 3, \ldots$)
- **Regularization:** Add $\sigma_k \|x - x_k\|^{p+1}$ term? (Yes/No)

	Unregularized	**Regularized (ARp)**
$p = 2$	Newton's Method Local quadratic conv.	ARC [CGT11] Global conv. $\mathcal{O}(\epsilon^{-3/2})$
$p = 3$	3rd-Order Newton [SZ22] Local cubic conv.	AR3 [CGT20] Global conv. $\mathcal{O}(\epsilon^{-4/3})$
$p \geq 4$	NP-hard subproblem Not tractable	ARp [CGT20] $\mathcal{O}(\epsilon^{-(p+1)/p})$

Going Beyond 2nd Order: ARp

Adaptive Regularization with Models of Order p (ARp):

Going Beyond 2nd Order: ARp

Adaptive Regularization with Models of Order p (ARp):

$$\widetilde{m}_{f,x_k}(x,\sigma_k) = \underbrace{f(x_k) + \sum_{j=1}^{p} \frac{1}{j!} \nabla_x^j f(x_k)\,[x - x_k]^j}_{\Phi^p_{f,x_k}(x)\text{: } p\text{-th order Taylor expansion at } x_k} + \underbrace{\frac{\sigma_k}{p+1}\,\|x - x_k\|^{p+1}}_{\text{Regularization } (\sigma_k > 0)} \qquad (1)$$

Going Beyond 2nd Order: ARp

Adaptive Regularization with Models of Order p (ARp):

$$\widetilde{m}_{f,x_k}(x,\sigma_k) = \underbrace{f(x_k) + \sum_{j=1}^{p} \frac{1}{j!} \nabla_x^j f(x_k)\,[x-x_k]^j}_{\Phi^p_{f,x_k}(x)\text{: } p\text{-th order Taylor expansion at } x_k} + \underbrace{\frac{\sigma_k}{p+1}\,\|x-x_k\|^{p+1}}_{\textbf{Regularization } (\sigma_k>0)} \tag{1}$$

Update: Solve the subproblem

$$x_{k+1} = \arg\min_{x\in\mathbb{R}^n} \widetilde{m}_{f,x_k}(x,\sigma_k) \tag{2}$$

Going Beyond 2nd Order: ARp

Adaptive Regularization with Models of Order p (ARp):

$$\widetilde{m}_{f,x_k}(x,\sigma_k) = \underbrace{f(x_k) + \sum_{j=1}^{p} \frac{1}{j!} \nabla_x^j f(x_k)\,[x - x_k]^j}_{\Phi^p_{f,x_k}(x)\text{: } p\text{-th order Taylor expansion at } x_k} + \underbrace{\frac{\sigma_k}{p+1}\,\|x - x_k\|^{p+1}}_{\text{Regularization } (\sigma_k > 0)} \tag{1}$$

Update: Solve the subproblem

$$x_{k+1} = \arg\min_{x \in \mathbb{R}^n} \widetilde{m}_{f,x_k}(x,\sigma_k) \tag{2}$$

Various methods: Sum-of-Squares (SOS) [ACZ24], Gradient Descent, ...

Pros

- Global convergence guarantees
- Higher-order convergence rates

Cons

- No uniform method for (2)
- Most methods are NP-hard

Unregularized Third-Order Newton: Motivation

A surprisingly tractable case for third order!

Theorem ([SZ22, AZ22])

A local minimum of a cubic polynomial can be found by solving semidefinite programs (SDP) of size linear in the number of variables.

SDP format:

$$\begin{array}{ll} \min_{X \in \mathbb{S}^{n \times n}} & \operatorname{Tr}(CX) \\ \text{s.t.} & \operatorname{Tr}\left(A_i X\right) = b_i, i = 1, \ldots, m \\ & X \succeq 0 \end{array}$$

Geometric meaning: intersections of a convex cone and some hyperplanes.

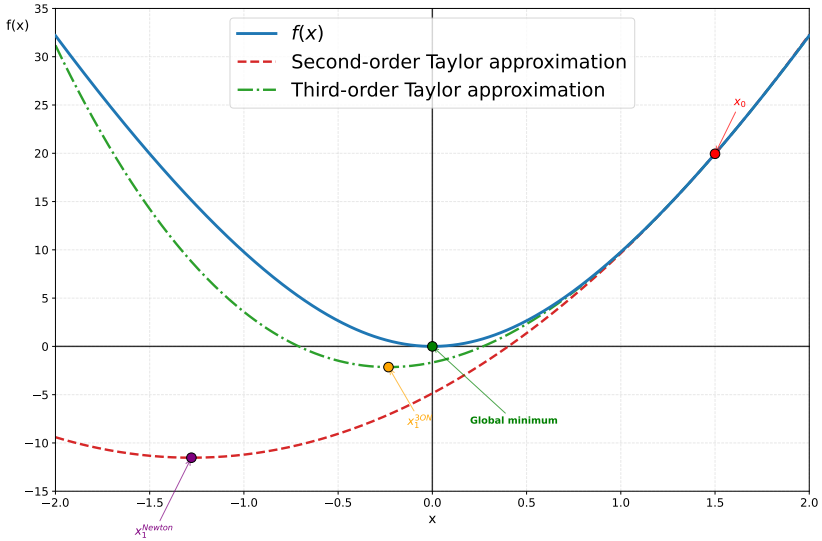
REMARK. SDP can be solved to arbitrary accuracy in polynomial time [VB96].

Theorem (Locally cubic convergence [SZ22])

If f is strongly convex and has a Lipschitz third derivative, the 3^{rd}-order unregularized Newton method has local cubic convergence.

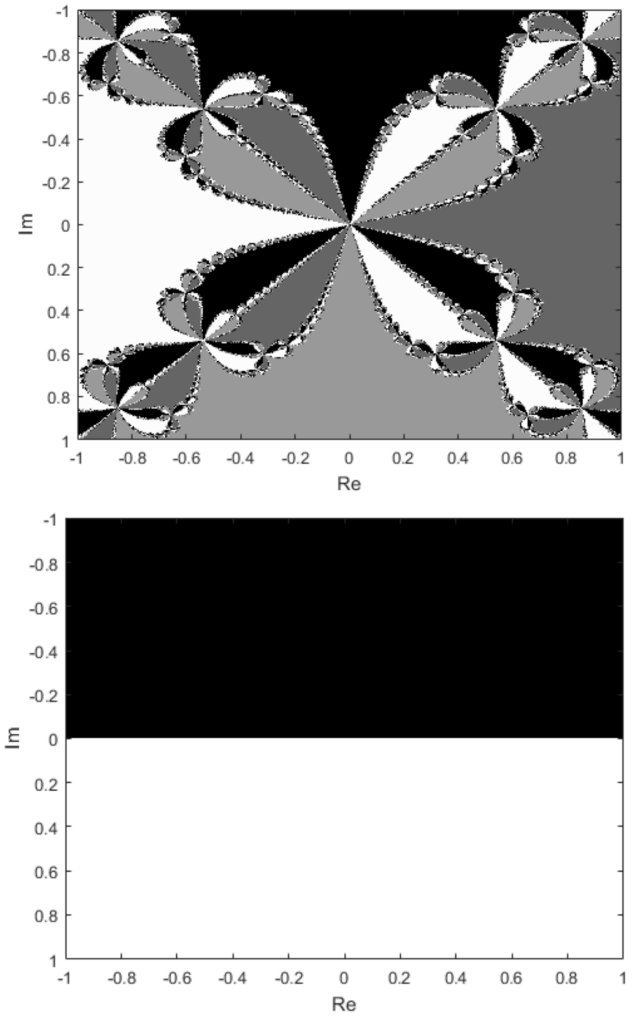
Unregularized Third-Order Newton: Advantages

Faster Convergence:



Better approximation $\Longrightarrow$ fewer iterations

Larger Basin of Attraction:



Left: Newton Right: 3rd-Order Newton

More robust to initialization

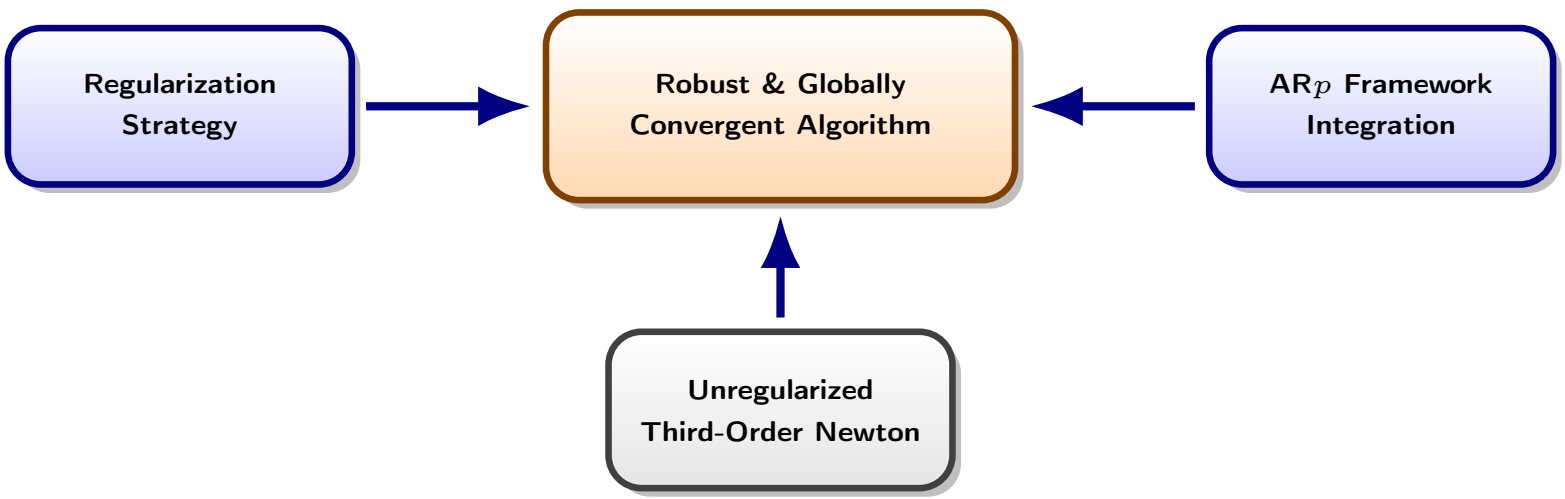
Unregularized Third-Order Newton: Issues

But still issues:

- Locally cubic convergence? **TOO STRONG**.
- $\Phi^3_{f,x_k}(x)$ has no local minimum $\Longrightarrow$ SDP infeasible: regularization?
- No global convergence: ARp?

Solution:

Proposed Solution Framework



Unregularized Third-Order Newton: Regularization

Q: still want to keep cubic polynomial, how?

A: Levenberg Marquardt Regularization: $\|\cdot\|^2$.

Existence Condition [SZ22]

$$\sigma_k \geq \alpha_k^{LM} \quad \Longrightarrow \quad \text{local minimum of } \Phi^3_{f,x_k} + \sigma_k \|x - x_k\|^2 \text{ exists.}$$

$$\alpha_k^{LM} := \underbrace{\sqrt{\frac{3}{2}\left(\|g_k\|\|h_k\| + g_k{}^\top h_k\right)}}_{\text{Gradient-Tensor Interaction}} - \underbrace{\min\{0, \lambda_k\}}_{\text{Hessian Spectral Shift}}$$

λ_k: Hessian Info

$$\lambda_k = \lambda_{\min}\left(\nabla^2 f(x_k)\right)$$

g_k: Gradient Info

$$g_k := \begin{bmatrix} |\nabla_1 f(x_k)| \\ \vdots \\ |\nabla_n f(x_k)| \end{bmatrix}$$

h_k: Tensor Info

$$h_k := \begin{bmatrix} \|\nabla_1^3 f(x_k)\| \\ \vdots \\ \|\nabla_n^3 f(x_k)\| \end{bmatrix}$$

ALMTON: Assumptions

ALMTON: Adaptive Levenberg-Marquardt Third-Order Newton's Method

Standard Assumptions

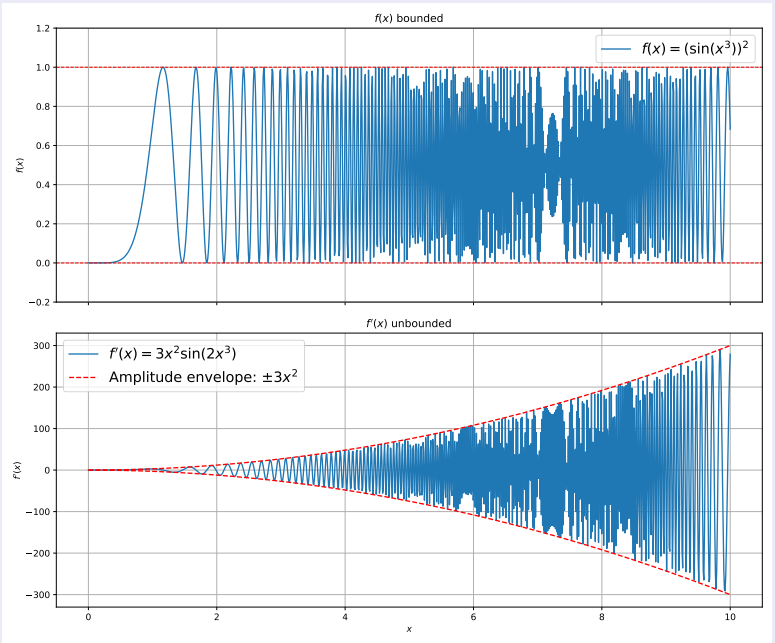
A1. *Smoothness & Lower Bound*

- $f \in \mathcal{C}^p(\mathbb{R}^n)$
- $f \geq f_{\text{low}}$ (bounded below)
- $\nabla^p f$ is L-Lipschitz

A2. *Uniform Tensor Bound for all* x_k

- $\left\|\nabla^j f(x_k)\right\|_{[j]} \leq \Lambda_j$
- $\forall k \geq 0,\ j = 1, \ldots, p$

Why Assumption 2?

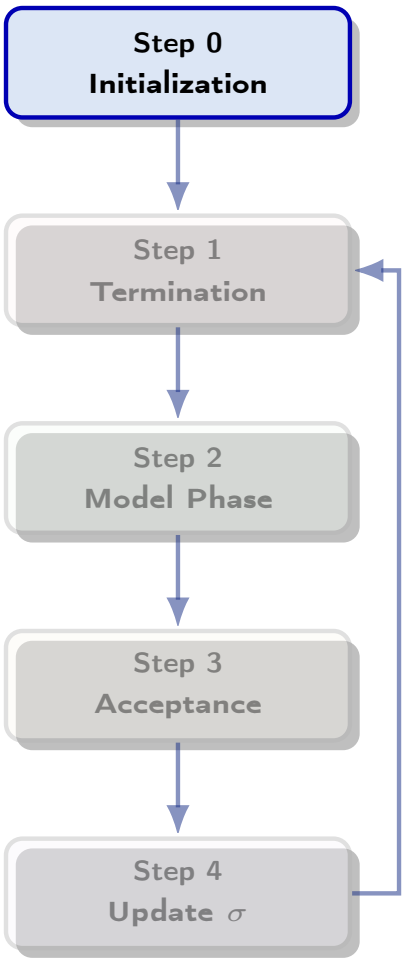


Takeaway

Monotonicity $\not\Rightarrow$ Bounded Derivatives.

We need **A2** to prevent oscillation (as shown above).

ALMTON: Step-by-Step



Step 0: Initialization

Given:

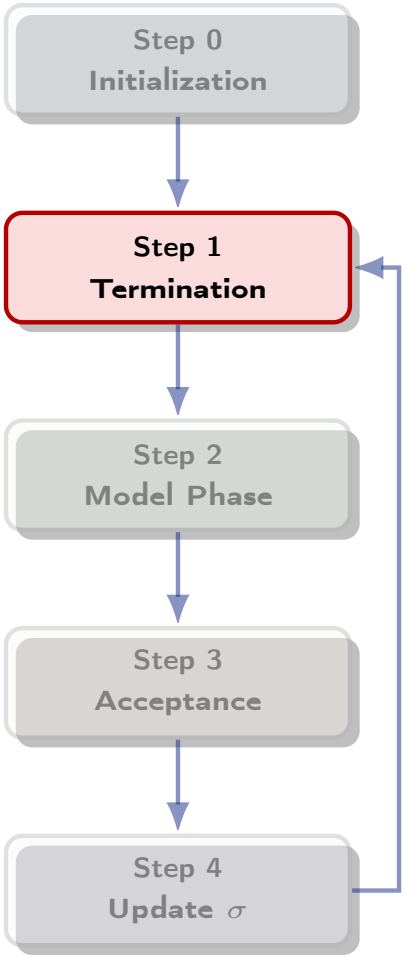
- $x_0 \in \mathbb{R}^n$: initial point
- $\epsilon > 0$: gradient tolerance
- $\sigma_0 = 0$: initial regularization
- $c, l > 0$: curvature constants (small)
- $\eta \in (0, 1)$: acceptance threshold
- $\gamma > 1$: regularization growth factor

Action: Compute $f(x_0)$ and set $k \leftarrow 0$.

Key Insight

Start with **no regularization** ($\sigma_0 = 0$) to preserve the unregularized third-order Newton step whenever possible.

ALMTON: Step-by-Step



Step 1: Termination Check

Evaluate: $\nabla f(x_k)$

If $\|\nabla f(x_k)\| \leq \epsilon$:

- ✓ **Terminate** with $x_\epsilon = x_k$

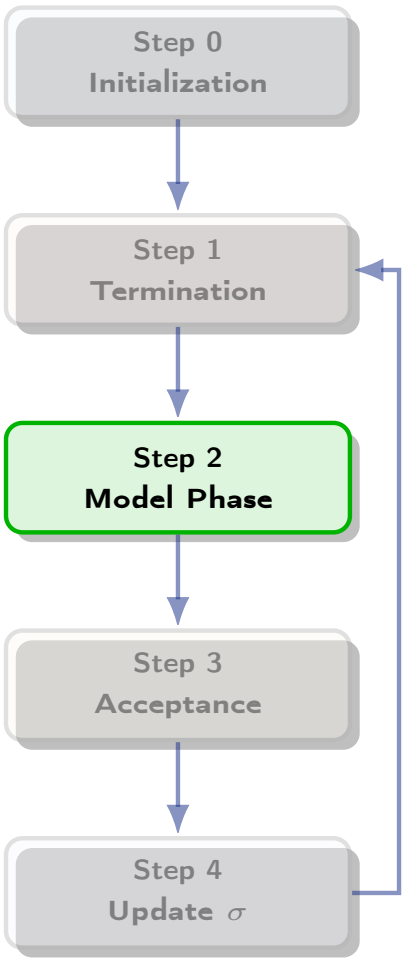
Else:

- Compute derivatives: $\nabla f(x_k)$, $\nabla^2 f(x_k)$, $\nabla^3 f(x_k)$
- Compute LM threshold: $\alpha_{LM}(x_k)$
- Proceed to **Step 2**

Stopping Criterion

First-order optimality: $\|\nabla f(x_k)\| \leq \epsilon$ guarantees an ϵ-approximate stationary point.

ALMTON: Step-by-Step



Step 2: Model Phase (SDP Subproblem)

Model: $m_{f,x_k}(x;\sigma_k) = \Phi^3_{f,x_k}(x) + \sigma_k \|x - x_k\|^2$

Try $\tilde{\sigma} \leftarrow \sigma_k$ (prefer unregularized):

- Solve SDP for local minimizer $\bar{x}$
- Compute $\bar{\lambda}_k = \lambda_{\min}(\nabla^2 f(\bar{x}) + 2\tilde{\sigma} I)$

If $\bar{\lambda}_k < c$ or no minimizer:

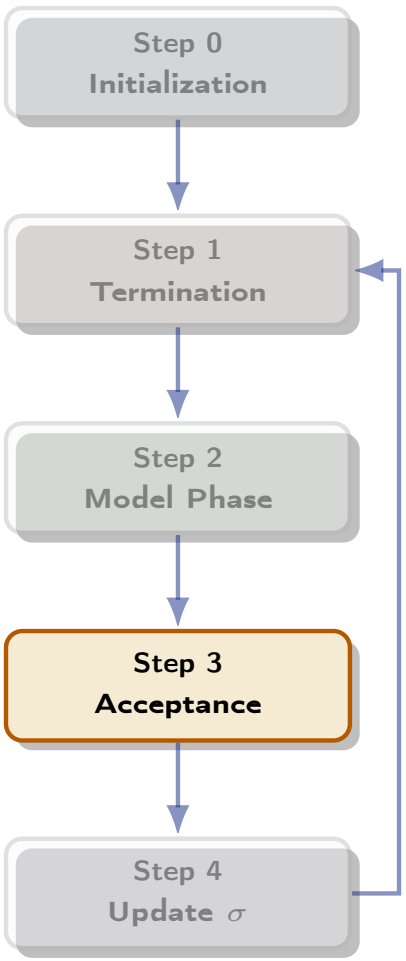
- Increase $\tilde{\sigma} \leftarrow \max\{\alpha_{LM}, \gamma\tilde{\sigma}, \tilde{\sigma} + (c - \bar{\lambda}_k)_+\}$
- Repeat until valid step found

Output: Trial step $s_k = \bar{x} - x_k$, phase value $\tilde{\sigma}_k$

Key Feature

The model remains **cubic** $\Rightarrow$ unified SDP solver for all $\sigma_k \geq 0$.

ALMTON: Step-by-Step



Step 3: Acceptance of Trial Point

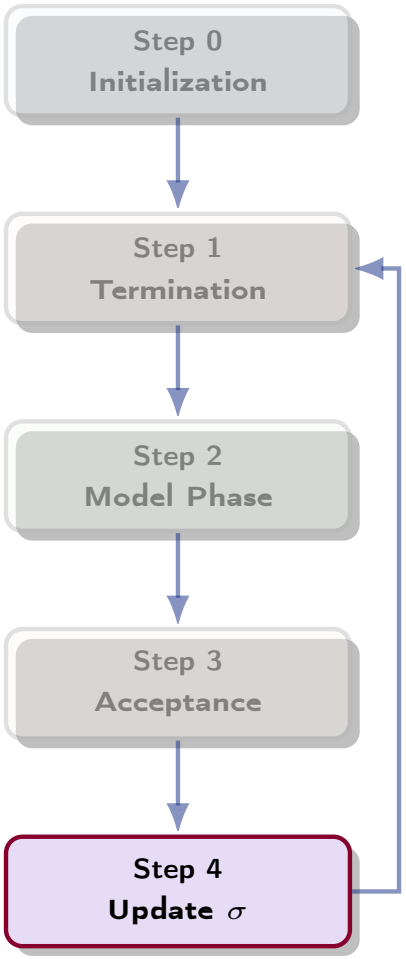
Evaluate: $f(\bar{x})$

Compute acceptance ratio:

$$\rho_k = \begin{cases} \dfrac{f(x_k) - f(\bar{x})}{l \, \|s_k\|^2} & \text{if } \tilde{\sigma}_k = 0 \\ \dfrac{f(x_k) - f(\bar{x})}{f(x_k) - m_{f,x_k}(\bar{x}; \tilde{\sigma}_k)} & \text{if } \tilde{\sigma}_k > 0 \end{cases}$$

Decision:

- $\rho_k \geq \eta$: **Success!** Set $x_{k+1} = \bar{x}$
- $\rho_k < \eta$: **Failure.** Set $x_{k+1} = x_k$



Step 4: Regularization Update

Update regularization parameter:

$$\sigma_{k+1} = \begin{cases} 0 & \text{if } \rho_k \geq \eta \text{ (success)} \\ \max\{\alpha_{LM}, \gamma\} & \text{if } \rho_k < \eta \text{ and } \tilde{\sigma}_k = 0 \\ \gamma \cdot \tilde{\sigma}_k & \text{if } \rho_k < \eta \text{ and } \tilde{\sigma}_k > 0 \end{cases}$$

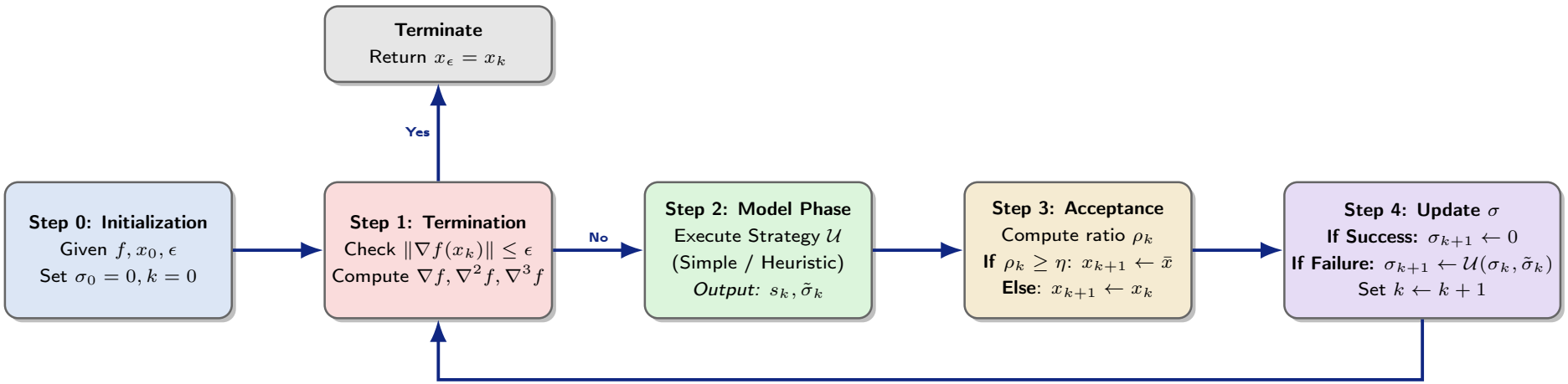
Increment: $k \leftarrow k + 1$

Loop: Go to **Step 1**

Reset Strategy

After **success**: reset $\sigma_{k+1} = 0$ to re-test unregularized step.
After **failure**: increase σ to ensure progress.

ALMTON: Full Algorithm Flow



Success Path ($\rho_k \geq \eta$)

$x_{k+1} = \bar{x}$, $\sigma_{k+1} = 0$ (reset)

Failure Path ($\rho_k < \eta$)

$x_{k+1} = x_k$, $\sigma_{k+1} \uparrow$ (increase)

ALMTON: Global Convergence – Three Pillars

Pillar 1: Descent

Monotonicity

$$\rho_k \geq \eta \implies$$

$$\boxed{f(x_{k+1}) < f(x_k)}$$

Pillar 2: Bounded σ

Controlled Regularization

$$\boxed{\sigma_k \leq \sigma_{\max}}$$

Large σ $\implies$ small step $\implies$ success $\implies$ reset

Pillar 3: Step Size

Non-Vanishing Step

$$\boxed{\|s_k\| \geq \delta(\epsilon) > 0}$$

Curvature floor $\bar{\lambda}_k \geq c$ $\implies$ steps stay significant

Combined Effect: (1) Descent + (2) Bounded σ + (3) Non-vanishing step $\implies$ **Cannot stall or oscillate indefinitely**

ALMTON: Complexity Analysis – $\mathcal{O}(\epsilon^{-2})$

Total Budget (Finite Capacity)

Total possible decrease is bounded: $f(x_0) - f_{\textbf{low}} \leq C_{\textbf{budget}} < \infty$

$\Downarrow$

Minimum Cost Per Successful Step

Each successful iteration "costs" at least:

$$f(x_k) - f(x_{k+1}) \geq \eta \cdot l \cdot \|s_k\|^2 \geq \underbrace{\kappa_s^{-1}}_{\textbf{const}} \cdot \epsilon^2$$

$\Downarrow$

Result: Iteration Complexity

$$|\mathcal{S}_k| = \frac{\text{Budget}}{\text{Cost per step}} \leq \frac{f(x_0) - f_{\textbf{low}}}{\kappa_s^{-1}\epsilon^2} = \boxed{\mathcal{O}(\epsilon^{-2})}$$

Unsuccessful Iterations (Batch Logic)

$|\mathcal{U}_k| \leq C \cdot |\mathcal{S}_k|$ (geometric growth of σ_k bounds failures between successes) $\implies$ **Total iterations** $= |\mathcal{S}_k| + |\mathcal{U}_k| = \mathcal{O}(\epsilon^{-2})$

Numerical Experiments: Setup

Baseline Hierarchy

First-Order:

- Gradient Descent (GD, $\alpha \in \{0.01, 0.05\}$)

Second-Order:

- Damped Newton's Method
- Newton-CG (Truncated Newton)
- AR2-Interp [CHL+24]

Third-Order:

- Unregularized 3rd-Order Newton [SZ22]
- AR3-Interp [CHL+24]

Implementation

Environment: `Python 3.12`
SDP Solver: `MOSEK` (fallback: `SCS`, `CVXOPT`)
Tolerance: $\|\nabla f(x_k)\| \leq 10^{-8}$

Exp 1: Robustness

Goal: Basin of attraction analysis
Method: Dense grid (30×30) on 4 non-convex functions
Metric: Dolan-Moré profiles

Exp 2: Scalability

Goal: High-dimensional stress test
Method: Rosenbrock-n ($n = 5, 20$)
Metric: Success rate, time, iterations

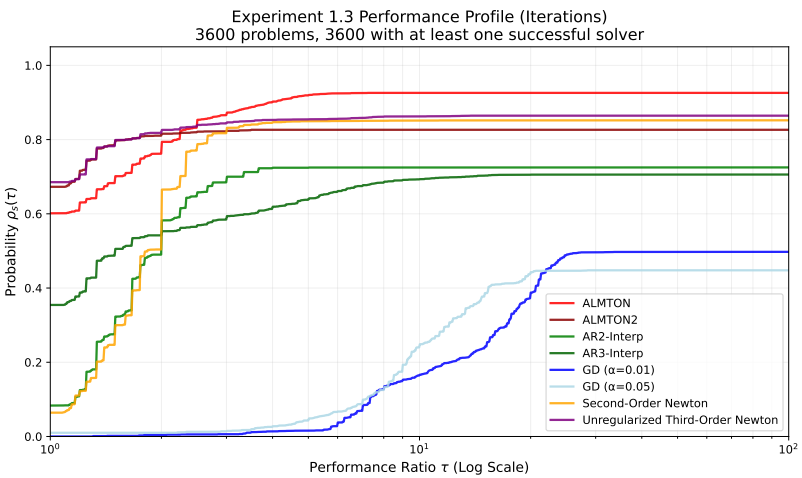
Exp 3: Geometry

Goal: Navigate pathological valleys
Method: Slalom & Hairpin Turn functions
Metric: Trajectory analysis, success rate

Exp 1: Robustness – Dolan-Moré Performance Profiles

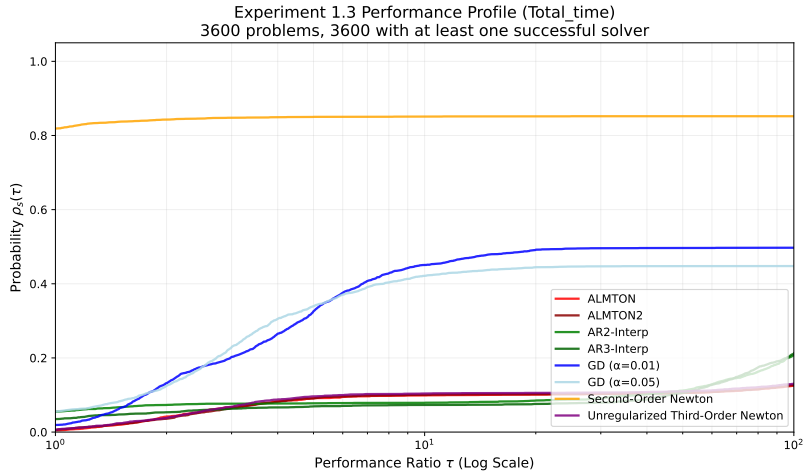
Test set: **4 nonconvex functions** $\times$ 30 $\times$ 30 **grid** = **3600 problem instances**

Iteration Complexity



- $\tau = 1$: win rate (best solver)
- ALMTON: 70%, AR3-Interp: 35%
- $\tau \to 100$: robustness (solved $\geq$ 90%)
- ALMTON: 90%, AR3-Interp: 70%

Wall-Clock Time



- $\tau = 1$: SDP overhead $\Rightarrow$ lower win rate
- $\tau \to \infty$: converges to AR3 level
- Damped Newton: fastest but unreliable
- **Trade-off:** per-step cost vs. robustness

Takeaway: ALMTON achieves highest robustness (90% vs 70%) with best iteration efficiency (70% win rate).

Exp 2: Scalability – High-Dimensional Rosenbrock

Rosenbrock Function:

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

Solver	Success	Iters	Time
	Dimension $N = 5$		
L-BFGS	100%	38	0.001s
Newton-CG	100%	30	0.002s
ALMTON	54%	795	31.259s
	Dimension $N = 20$		
L-BFGS	100%	110	0.006s
Newton-CG	100%	50	0.007s
ALMTON	9%	>1000	>170s

Scalability Bottleneck

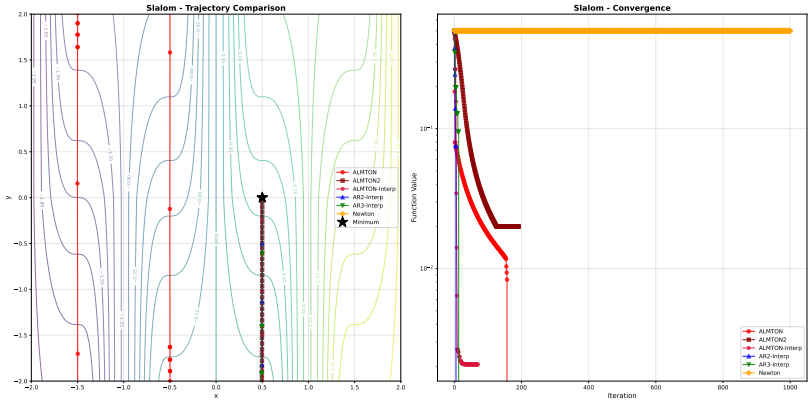
Root Cause: SDP subproblem complexity

- Lifting: $\mathbb{R}^n \to \mathbb{S}^{n+1}$
- SOTA-IPM complexity [JKL+20]: $\mathcal{O}(n^{3.5})$
- Standard commercial solvers: higher than $\mathcal{O}(n^{4.5})$ (e.g., MOSEK)
- Ill-conditioning $\implies$ large σ_k

Operational Boundary:
ALMTON is effective for Low-Dimensional problems.
High-Dimensional problems $\implies$ use AR3-Interp instead.

Exp 3: Geometry – Navigating Pathological Valleys

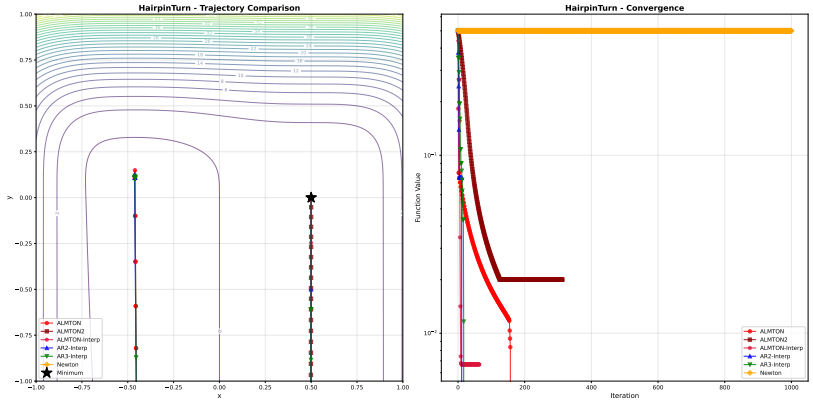
Slalom Function [CHL+24]



Curved valley with changing curvature

- Newton: zig-zag between walls, stagnates
- ALMTON: long curvilinear steps
- $\nabla^3 f$ reveals valley "twist" $\implies$ shortcuts

Hairpin Turn [CHL+24]



Sharp bend with barrier functions

- Newton-CG: 25.5% (crashes into barrier)
- ALMTON: 99.5% (smooth trajectory)
- Cubic model anticipates direction change

Key Insight: $\nabla^3 f(x_k)$ captures the "twist" in valleys $\implies$ cubic model *bends with the landscape*, where quadratic approximation *structurally fails*.

Numerical Summary & Future Directions

ALMTON Strengths

- ✓ **Robustness:** Largest basin of attraction
- ✓ **Geometric Intelligence:** Navigates complex curvature
- ✓ **Iteration Efficiency:** Fewer steps than SOTA AR3-Interp
- ✓ **Unified Subproblem:** Single SDP template

Current Limitations

- × **Scalability:** Effective only for Low-Dimensional problems
- × **SDP Cost:** Per-iteration overhead
- × **Solver Sensitivity:** Depends on SDP solver accuracy

Future Research

1. **Approximate Solvers:**
Krylov subspace / Lanczos methods

2. **Tensor Compression:**
Tensor-train decomposition, sketching

3. **Hybrid Strategies:**
Combine ALMTON with AR3-Interp for High-Dimensional problems

4. **Applications:**
Simulation-based optimization (expensive f evaluations)

Bottom Line: ALMTON excels in **low-dimensional, geometrically complex** problems where function evaluations dominate the cost.

References I


Amir Ali Ahmadi, Abraar Chaudhry, and Jeffrey Zhang, *Higher-order newton methods with polynomial work per iteration*, Advances in Mathematics **452** (2024), 109808.


Amir Ali Ahmadi and Jeffrey Zhang, *Complexity aspects of local minima and related notions*, Advances in Mathematics **397** (2022), 108119.


Coralia Cartis, Nicholas IM Gould, and Philippe L Toint, *Adaptive cubic regularisation methods for unconstrained optimization. part i: motivation, convergence and numerical results*, Mathematical Programming **127** (2011), no. 2, 245–295.


Coralia Cartis, Nick IM Gould, and Ph L Toint, *A concise second-order complexity analysis for unconstrained optimization using high-order regularized models*, Optimization Methods and Software **35** (2020), no. 2, 243–256.


Coralia Cartis, Raphael Hauser, Yang Liu, Karl Welzel, and Wenqi Zhu, *Efficient implementation of third-order tensor methods with adaptive regularization for unconstrained optimization*, arXiv preprint arXiv:2501.00404 (2024).

References II


Haotian Jiang, Tarun Kathuria, Yin Tat Lee, Swati Padmanabhan, and Zhao Song, *A faster interior point method for semidefinite programming*, 2020.


Olha Silina and Jeffrey Zhang, *An unregularized third order newton method*, arXiv preprint arXiv:2209.10051 (2022).


Lieven Vandenberghe and Stephen Boyd, *Semidefinite programming*, SIAM review **38** (1996), no. 1, 49–95.