

# **An Alternative to the Trust-Region: Homogeneous Second-Order Descent Method**

**KURT-FEST!, SIOPT, JUNE 2, 2023**

**Yinyu Ye**

**Stanford University and CUHKSZ (Sabbatical Leave)**

# Early Complexity Analyses for Nonconvex Optimization

- $\min f(x), x \in X$  in  $\mathbb{R}^n$ ,  
where  $f$  is nonconvex and twice-differentiable,  
 $g_k = \nabla f(x_k), H_k = \nabla^2 f(x_k)$
- Goal: find  $x_k$  such that:
  - $\| \nabla f(x_k) \| \leq \epsilon$  (primary, first-order condition)
  - $\lambda_{\min}(H_k) \geq -\sqrt{\epsilon}$  (in active subspace, second-order condition)
- For the ball-constrained nonconvex QP:  $\min c^T x + 0.5x^T Qx$  s.t.  $\|x\|_2 \leq 1$   
 $O(\log \log(\epsilon^{-1}))$ ; see Vavasis&Zippel (1990), Y (1989,93).
- For nonconvex QP with polyhedral constraints:  $O(\epsilon^{-1})$ ; see Y (1998), Vavasis (2001)

# Second-order Methods for General Optimization

## SOM (Hessian-Type Methods) with $M$ -Lipschitz cont. Hessian

- Trust-region (More 70, Sorenson 80). Fixed-radius TR  $O(\epsilon^{-\frac{3}{2}})$ , see the lecture notes by Y since 2005
- Cubic regularization,  $O(\epsilon^{-3/2})$ , see Nesterov and Polyak (2006), Cartis, Gould, and Toint (2011)
- An adaptive trust-region framework,  $O(\epsilon^{-3/2})$ , Curtis, Robinson, and Samadi (2017)

## SOM for convex functions

- Cubic regularization,  $O(\epsilon^{-1/2})$ , see Nesterov and Polyak (2006),
- Accelerated SOMs,  $O(\epsilon^{-1/3})$ ,  $O(\epsilon^{-1/3.5})$ , see Monteiro and Svaitor (2013), Nesterov (2008), Doikov et al. (2022)
- Linearly convergent SOMs, self-concordance, see Nesterov and Nemirovskii (1994); scaled Lipschitz, see Kortanek and Zhu (1993), Anderson and Ye (1998); generalized concordance, see Sun (2019).

Disadvantage: each iteration requires  $O(n^3)$  operations: **How to reduce it?**

# An Integrated Descent Direction Using the SDP Homogeneous Model I (Zhang et al. SHUFE, 2022)

- Recall the fixed-radius trust-region method minimizes the Taylor quadratic model

$$\begin{aligned} \min_{d \in \mathbb{R}^n} m_k(d) &:= g_k^T d + \frac{1}{2} d^T H_k d \\ \text{s.t. } \|d\| &\leq \Delta_k. \end{aligned} \quad \longrightarrow \quad \begin{aligned} \min_{[d,t] \in \mathbb{R}^{n+1}} m_k(d) &:= t \cdot g_k^T d + \frac{1}{2} d^T H_k d + \frac{1}{2} \delta \cdot (1-t^2) \\ \text{s.t. } \|d\|^2 + t^2 &= \Delta_k^2 + 1 \end{aligned}$$

where  $\Delta_k = \epsilon^{1/2} / M$  is the trust-ball radius.

- $-g_k$  is the first-order steepest descent direction but ignores Hessian;
- the most-left eigenvector of  $H_k$ -would be a descent direction for the second order term
- Could we construct a direction integrating both?

**Answer:** Use the most-left eigenvector of the SDP homogenized quadratic function!

(see Rojas 2001, a specialized Lanczos method for the Trust-region Subproblem with a given radius; and Adachi 2017 for solving more Generalized Trust-region Subproblems)

# An Integrated Descent Direction Using the SDP Homogeneous Model II (Zhang et al. SHUFE, 2022)

$$\psi_k(\xi_0, t; \delta) := \frac{1}{2} \begin{bmatrix} \xi_0 \\ t \end{bmatrix}^T \begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} \xi_0 \\ t \end{bmatrix} = \frac{t^2}{2} \begin{bmatrix} \xi_0/t \\ 1 \end{bmatrix}^T \begin{bmatrix} H_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} \xi_0/t \\ 1 \end{bmatrix}$$

- **Find the direction  $\xi = \xi_0/t$  (if  $t = 0$  then set  $t=1$ ) by the leftmost eigenvector:**

$$\min_{\|[\xi_0; t]\| \leq 1} \psi_k(\xi_0, t; \delta)$$

with a suitable  $\delta_k$  and use  $\xi$  as the direction to go – a single loop

algorithm to solve the original problem.

- **Accessible at the cost of  $O(n^2 \epsilon^{-1/4})$  via the randomized Lanczos method and needs only Hessian-Vector-Product (HVP).**



## How to Set $\delta$ : Theoretical Guarantees of HSODM

- Consider using the second-order homogenized direction, and let the length of each step  $\|\eta\xi\|$  be fixed:  $\|\eta\xi\| \leq \Delta_k = \frac{2\sqrt{\epsilon}}{M}$ , where  $f(x)$  has  $L$ -Lipschitz gradient and  $M$ -Lipschitz Hessian.
- **Theorem 1** (Global convergence rate) : Let  $f(x)$  satisfy the Lipschitz Assumption and fix  $\delta = \sqrt{\epsilon}$ , and let  $x_{k+1} = x_k + \eta_k \xi$  where  $\eta_k = \Delta_k / \|\xi\|$ , then algorithm has  $O(\epsilon^{-3/2})$  iteration complexity to second-order stationarity, where each iteration compute the most-left eigenvector of the homogenized matrix to  $\epsilon$  accuracy.
- **Theorem 2** (Local convergence rate): If the iterate  $x_k$  of HSODM converges to a strict local optimum  $x^*$ , HSODM possesses a local superlinear (quadratic) speed of convergence:  $\|x_{k+1} - x^*\| = O(\|x_k - x^*\|^2)$ .

# HSODM with Line-Search Methods

- **Fixed** step length  $\eta_k$  may be too conservative.
- **Observation I:** homogenized direction  $\xi$  can be used with **any** Line-search (e.g., Hager-Zhang)
- **Theorem 3 (Global convergence with Line-search, informal)** : If we apply the backtrack to compute  $\eta_k$  with parameter  $\beta \in (0,1)$  then the algorithm converges in  $O\left(\epsilon^{-\frac{3}{2}} |\log_{\beta}(\epsilon)|\right)$  iterations.

# Application I: HSODM for Policy Optimization in Reinforcement Learning

- Consider policy optimization of linearized objective in reinforcement learning

$$\max_{\theta \in \mathbb{R}^d} L(\theta) := L(\pi_\theta),$$

$$\theta_{k+1} = \theta_k + \alpha_k \cdot M_k \nabla \eta(\theta_k),$$

- The Natural Policy Gradient (NPG) method (Kakade, 2001) uses the Fisher information matrix where  $M_k$  is the inverse of

$$F_k(\theta) = \mathbb{E}_{\rho_{\theta_k}, \pi_{\theta_k}} [\nabla \log \pi_{\theta_k}(s, a) \nabla \log \pi_{\theta_k}(s, a)^T]$$

- Based on KL divergence, TRPO (Schulman et al. 2015) uses KL divergence in the constraint:

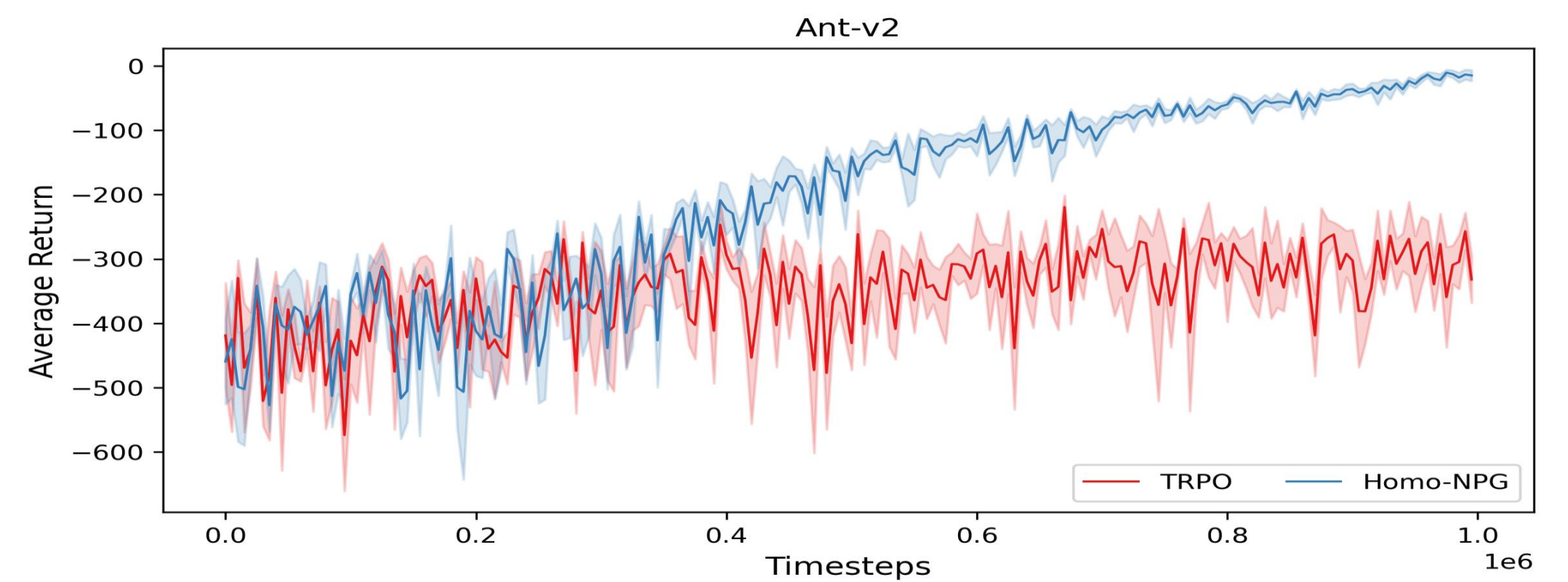
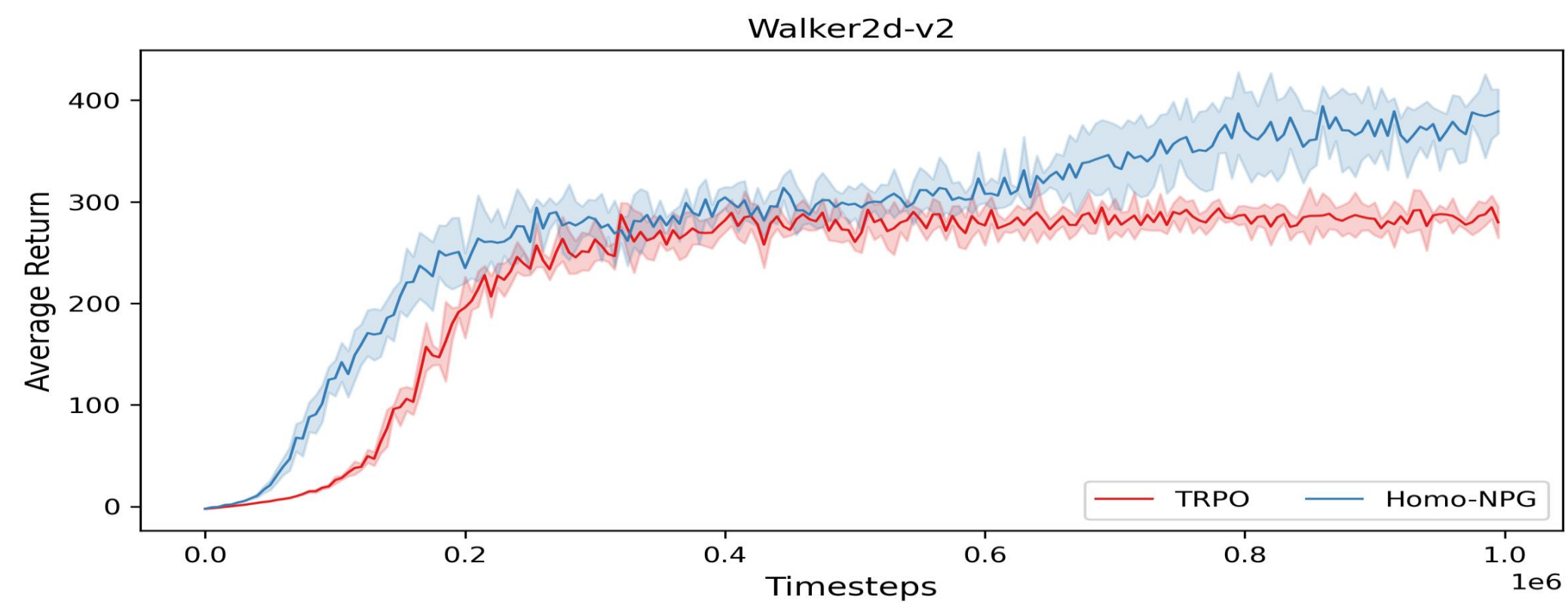
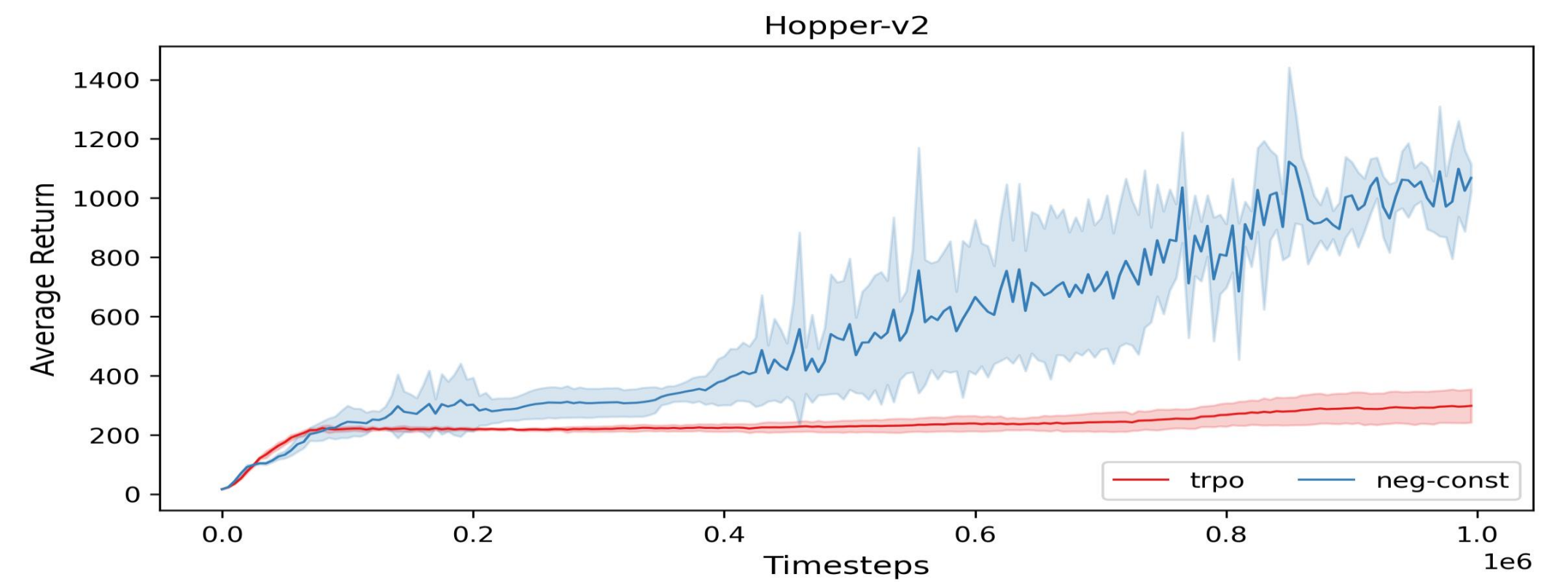
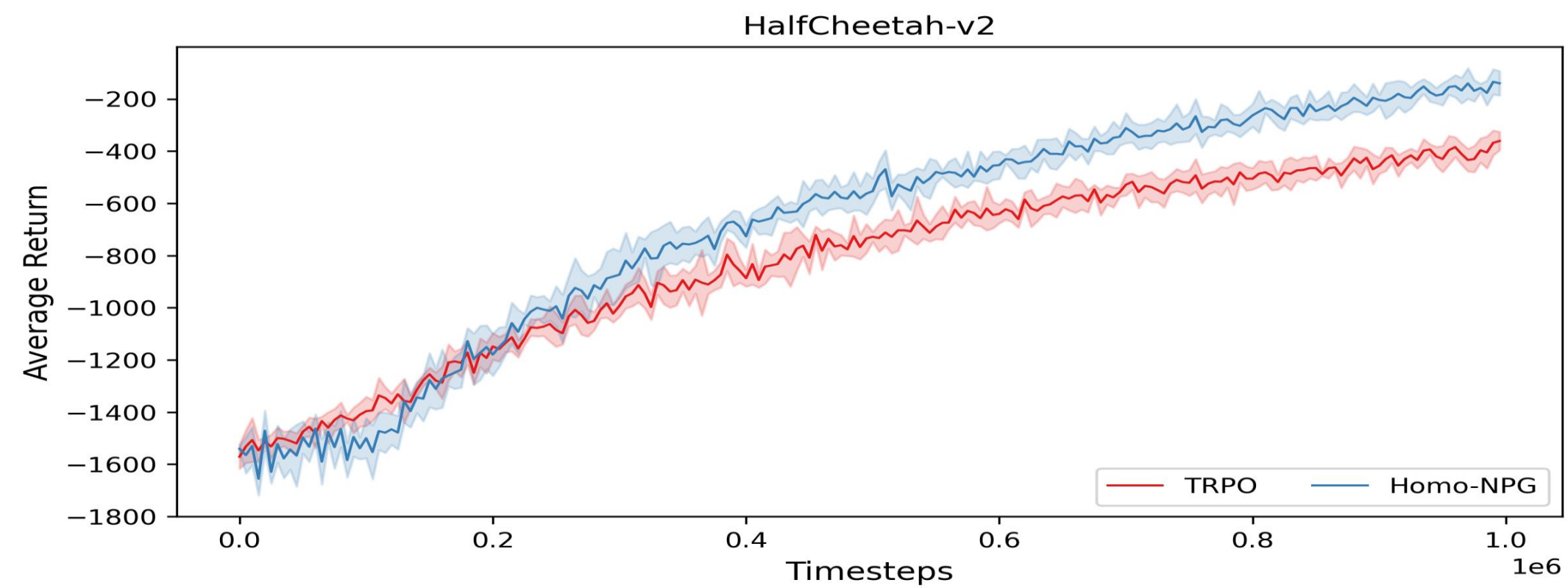
$$\begin{aligned} \max_{\theta} \nabla L_{\theta_k}(\theta_k)^T (\theta - \theta_k) \\ \text{s.t. } \mathbb{E}_{s \sim \rho_{\theta_k}} [D_{KL}(\pi_{\theta_k}(\cdot | s); \pi_{\theta}(\cdot | s))] \leq \delta. \end{aligned} \quad \longrightarrow \quad \min_{\| [v; t] \| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} F_k & g_k \\ g_k^T & -\delta \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}$$

**Homogeneous Natural Policy Gradient:  
Apply the Homogenized Direction with Line Search!**



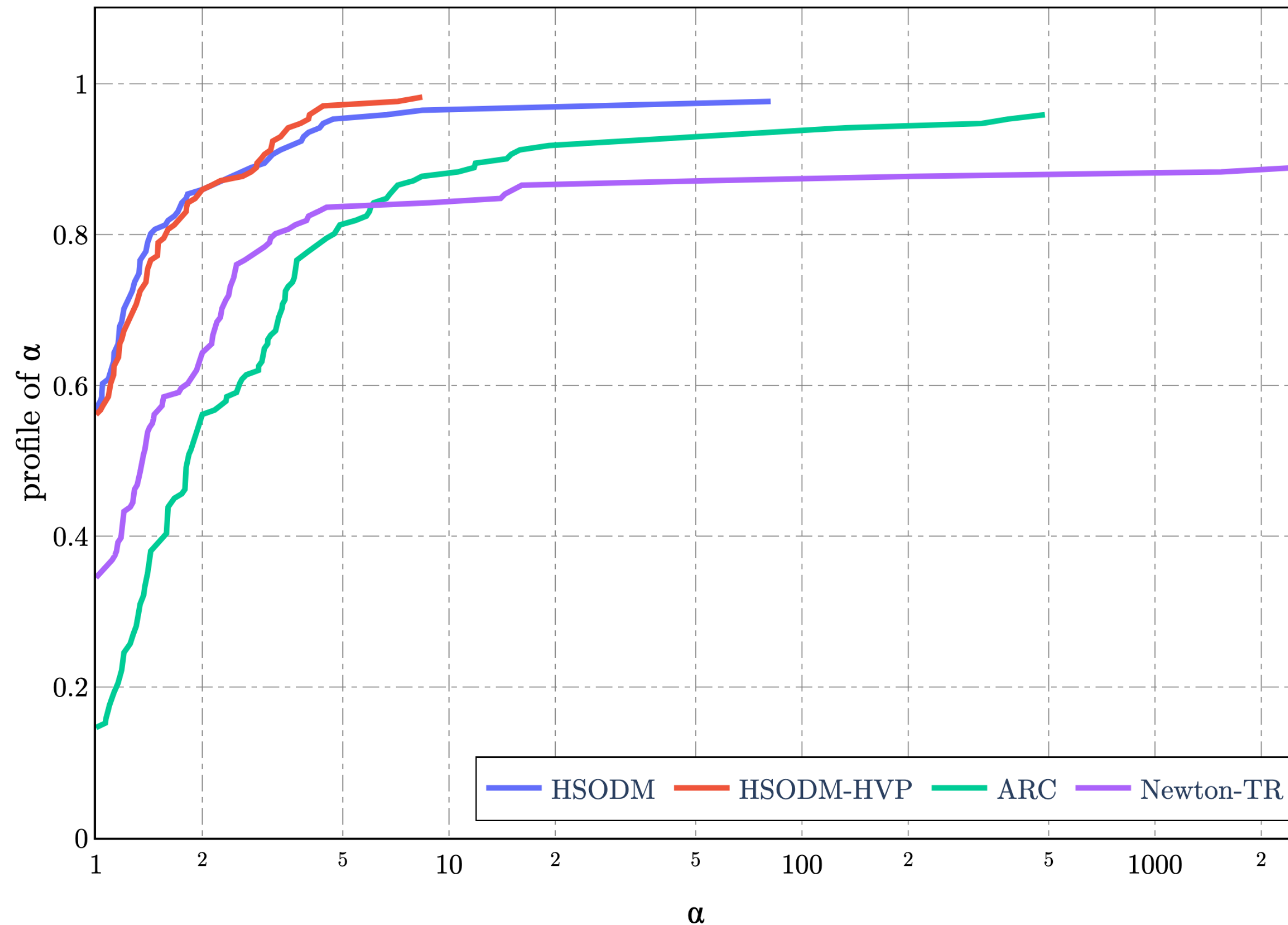
# HSODM for Policy Optimization in RL II

- A comparison of Homogeneous NPG and Trust-region Policy Optimization (Schultz, 2015)



- **Homogeneous model provides significant improvements over TRPO (public open-source solver)**

# Application II: HSODM for CUTEst Benchmark



- Compare HSODM (with Hessian), HSODM-HVP (with HVP), Newton TR and ARC
- Compare performance metrics in SGM

method	$\mathcal{K}$	$\bar{t}_G$	$\bar{k}_G$	$\bar{k}_G^f$	$\bar{k}_G^g$	$\bar{k}_G^H$
Newton-TR	155.00	15.41	216.59	211.99	219.58	203.82
HSODM	170.00	4.13	80.22	159.76	180.04	80.22
HSODM-HVP	171.00	5.25	110.61	193.07	1080.57	0.00
ARC	167.00	5.32	185.03	185.03	888.35	0.00

- $\mathcal{K}$  – success #,  $t_G$  - geometric mean running time (SGM),  $k_G$  - geometric mean iteration # (SGM)

## Performance Profile of iteration #

$\alpha$  – iteration # compared to the best

$profile(\alpha)$  – percentage of solved instances within  $\alpha$

- Newton-TR and ARC are public solvers



# Application III: HSODM for Sensor Network Localization

- Consider Sensor Network Location (SNL)

$$N_x = \{(i, j) : \|x_i - x_j\| = d_{ij} \leq r_d\}, N_a = \{(i, k) : \|x_i - a_k\| = d_{ik} \leq r_d\}$$

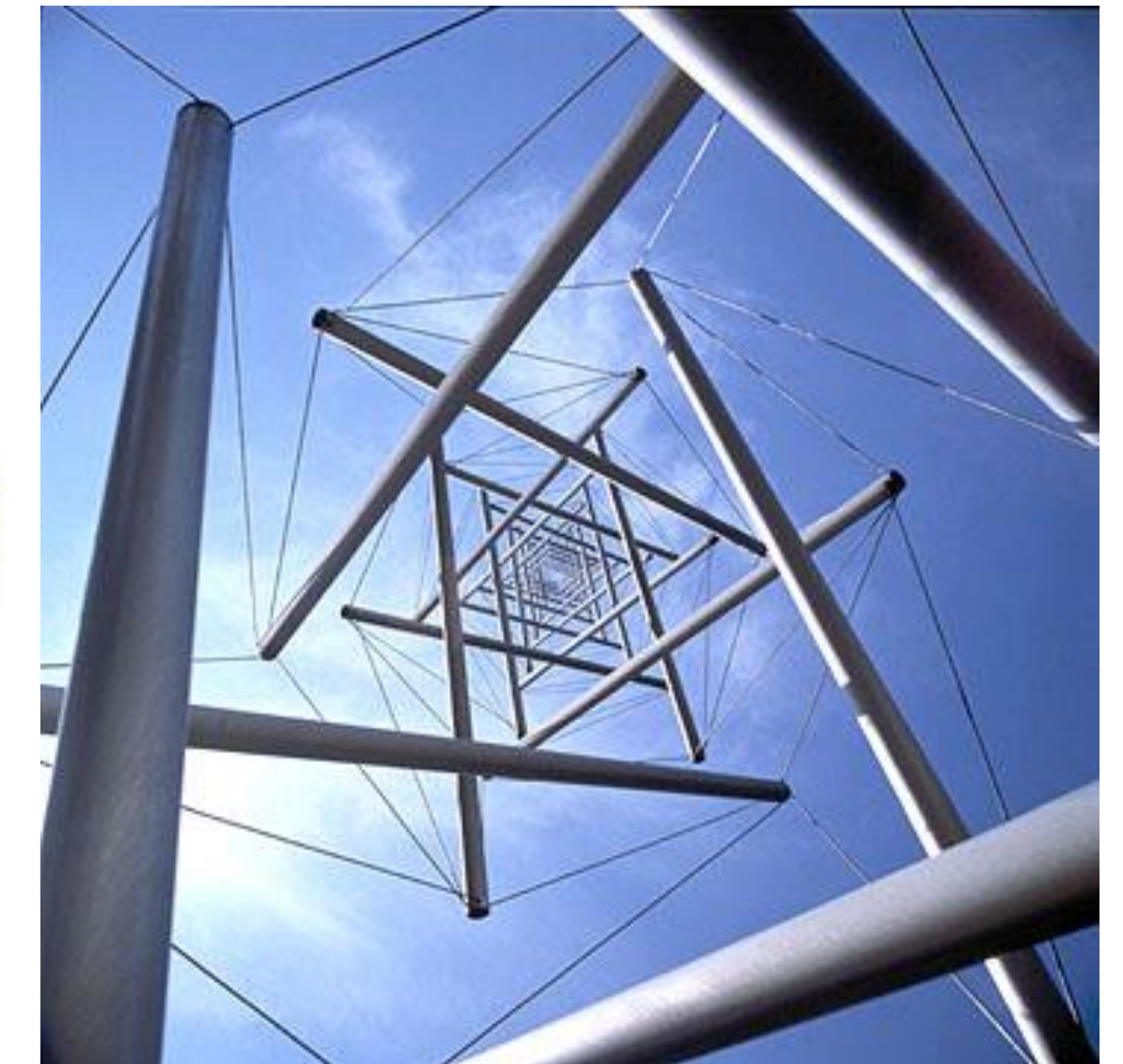
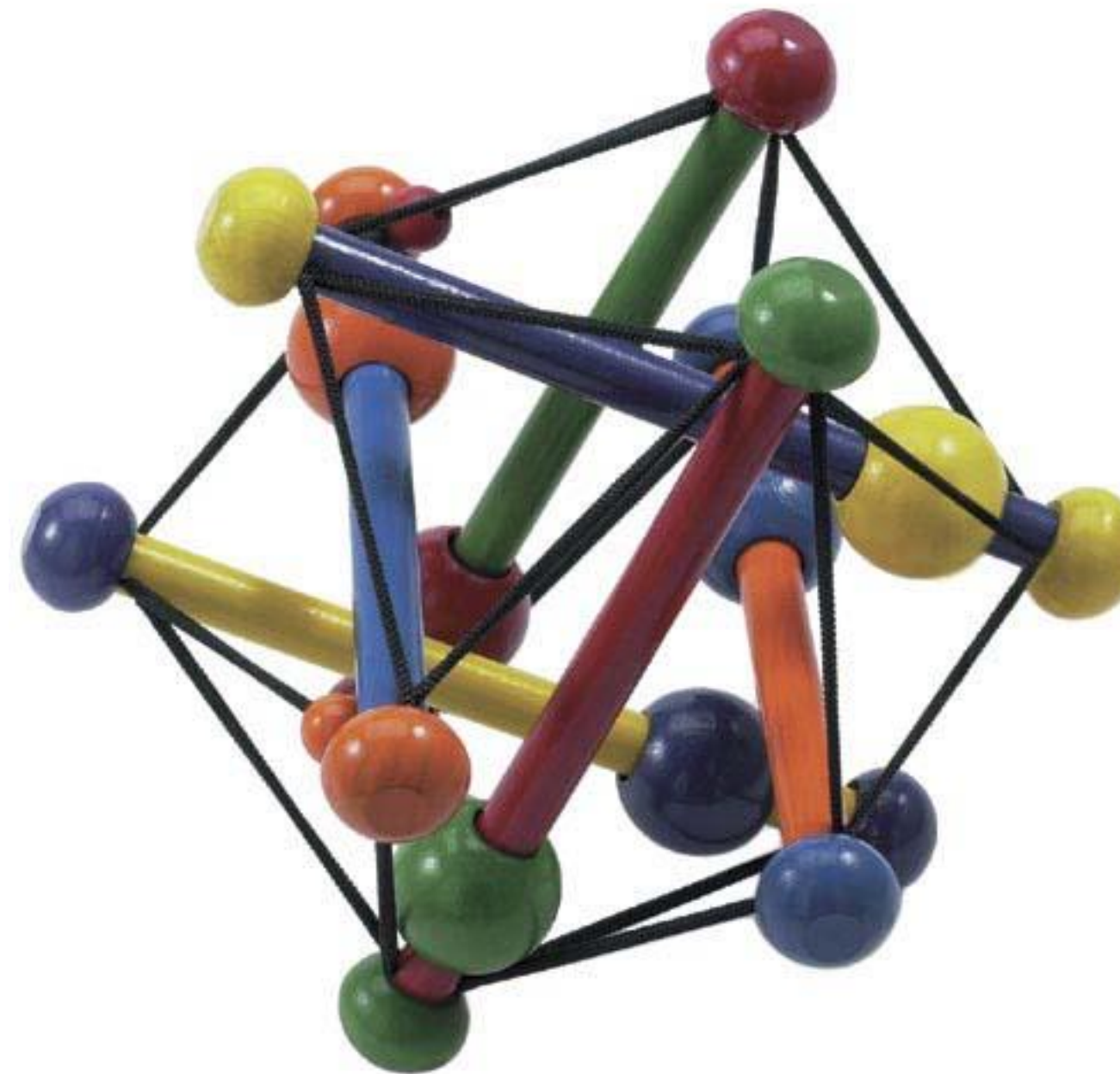
where  $r_d$  is a fixed parameter known as the radio range. The SNL problem considers the following QCQP feasibility problem,

$$\|x_i - x_j\|^2 = d_{ij}^2, \forall (i, j) \in N_x$$

$$\|x_i - a_k\|^2 = \bar{d}_{ik}^2, \forall (i, k) \in N_a$$

- We can solve SNL by the nonconvex nonlinear least square (NLS) problem

$$\min_X \sum_{(i,j) \in N_x} (\|x_i - x_j\|^2 - d_{ij}^2)^2 + \sum_{(k,j) \in N_a} (\|a_k - x_j\|^2 - \bar{d}_{kj}^2)^2.$$

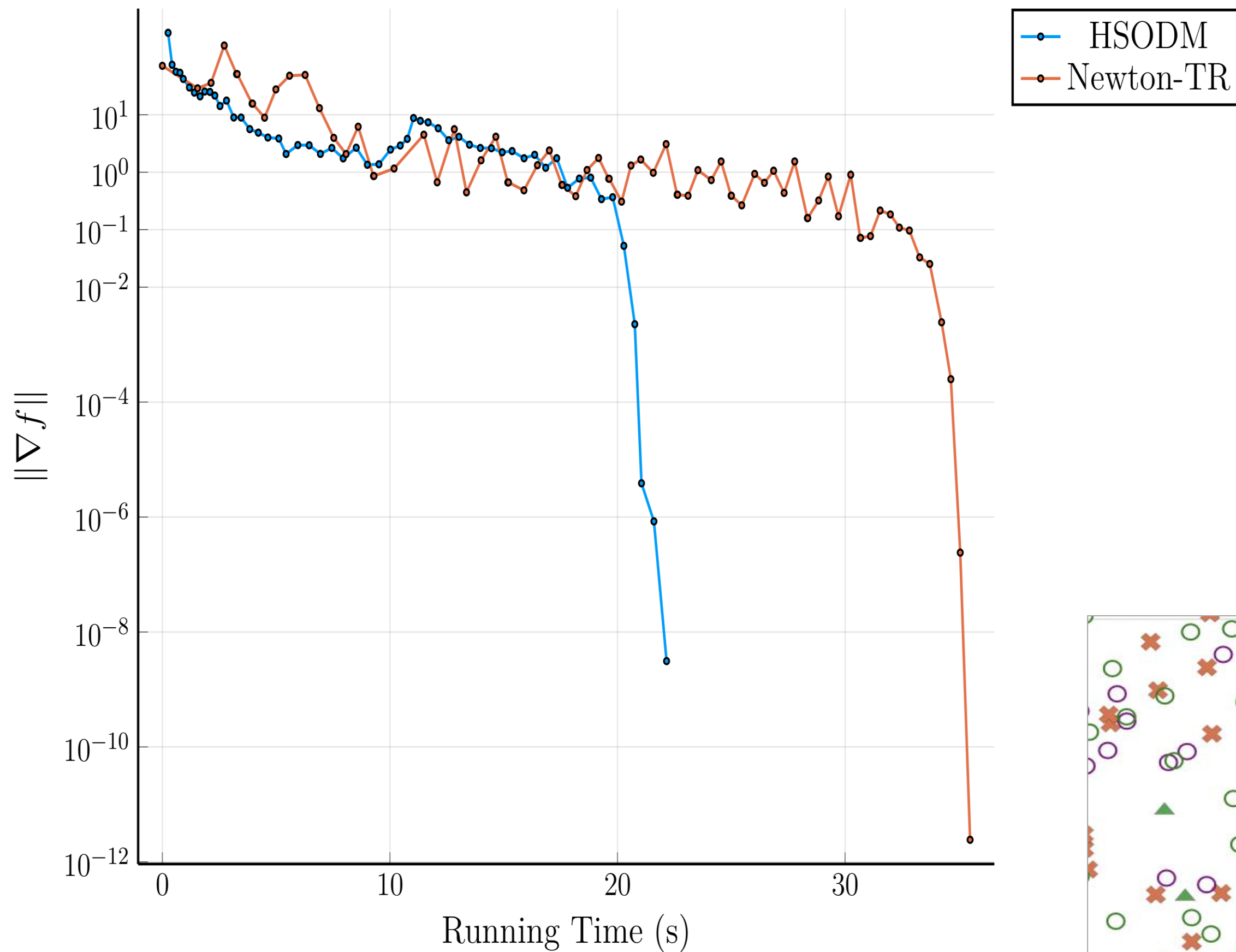


**Kurt's Collection**

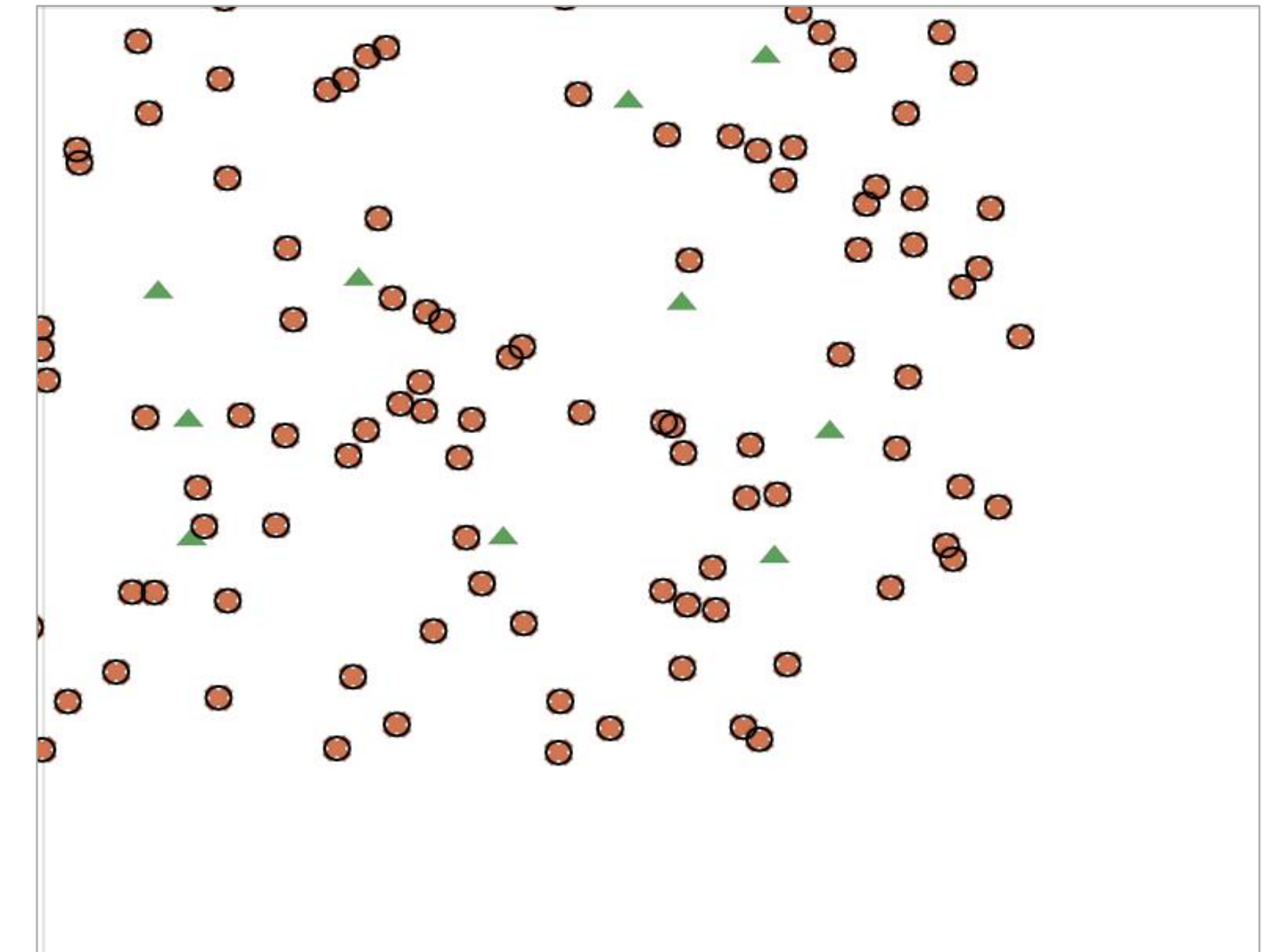
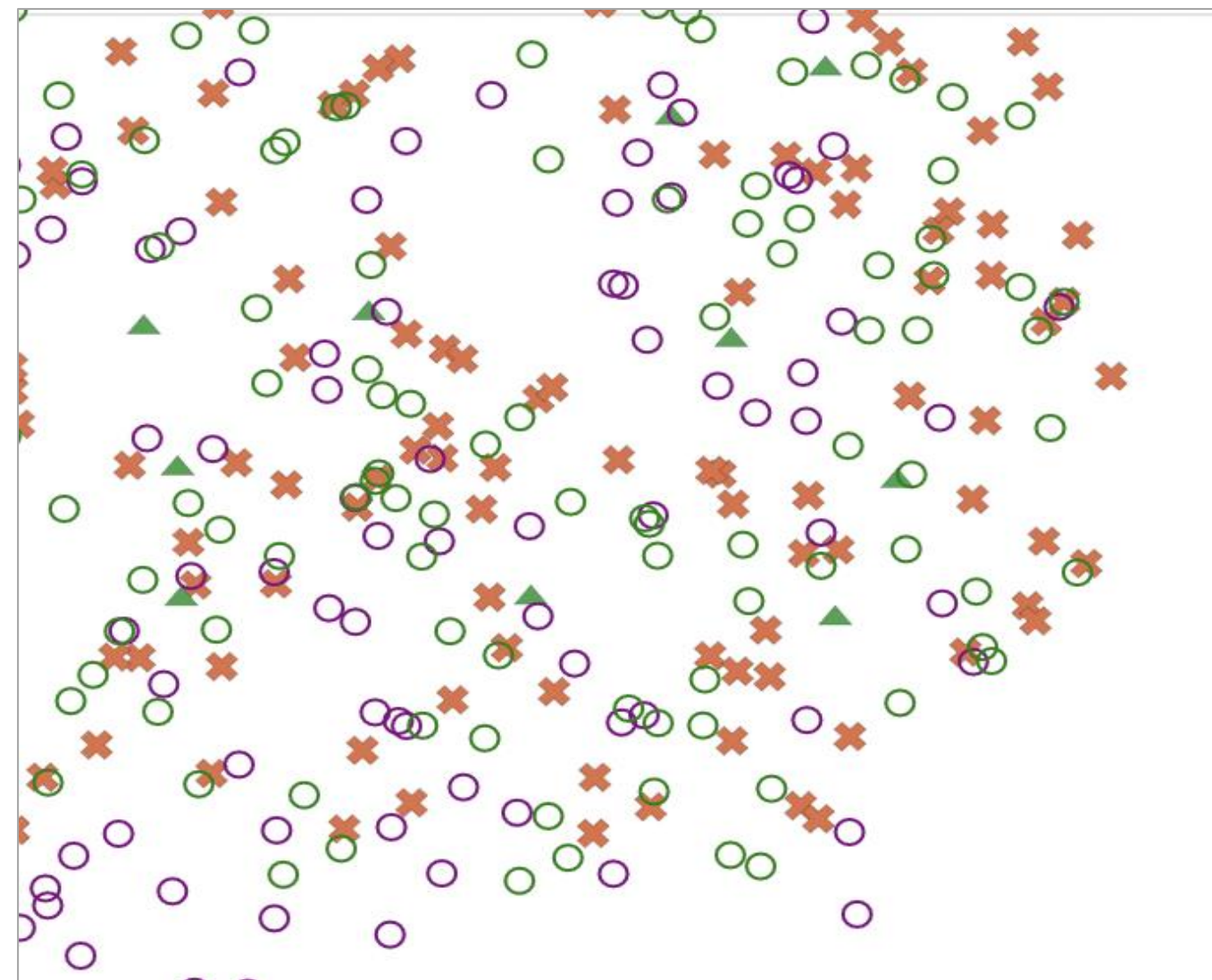


# Application III: HSODM for Sensor Network Localization

SNL,  $n := 200$ ,  $m := 20$



- Compare HSODM (with HVP), and Newton-TR Method.
- HSODM is faster due to the eigenvalue procedure
- The solution quality is much better than the FOMs

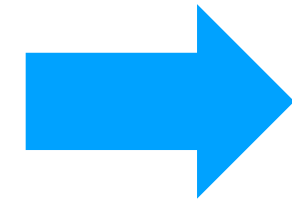


# Adaptive HSODM for 2<sup>nd</sup> order Lipschitz functions I

- Establish an equivalence of HSODM to Adaptive Trust-Region Method:

Method:

Adjust  $\delta_k \nearrow$



Implicit controls:  $|d_k(\delta_k)| \nearrow$

- Establish an equivalence of HSODM to Cubic Regularized Newton Method

$$d_k = \operatorname{argmin} \quad g_k^T d + \frac{1}{2} d^T H_k d + \frac{\sqrt{h_k(\delta_k)}}{3} \|d\|^3 \quad \longrightarrow \quad h_k(\delta_k) = \frac{\theta_k^2}{\|d_k\|^2}$$

where  $\theta_k$  is the dual variable; therefore one can tune  $\delta_k$  adaptively using a **bisection** to find proper  $h_k$

**Takeaway: "O(n<sup>3</sup>) Newton" can be replaced by  $O(n^2 \epsilon^{-1/4})$**

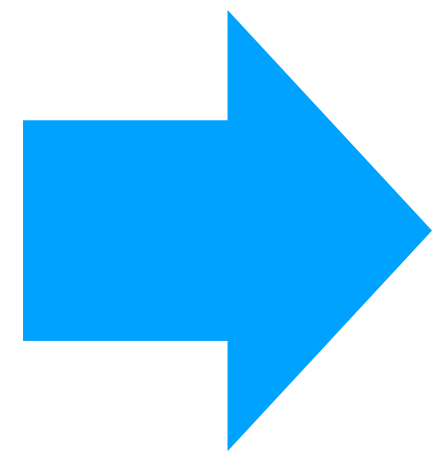


# Generalized Homogeneous Model (GHM) and HSODM

- Can we equivalent HSODM to more second-order frameworks?
- Introduce *Generalized Homogeneous Model (GHM)*

$$\begin{bmatrix} H_k & g_k \\ g_k^T & \delta \end{bmatrix} \Rightarrow \begin{bmatrix} H_k & \phi_k \\ \phi_k^T & \delta_k \end{bmatrix},$$

- Adaptive  $\delta_k$  and smart choice of  $\phi_k$  ( $g_k$  suffices in most case)



Method	Adaptive Controls		Complexity	References
	$\phi_k$	$\delta_k$		
Gradient Regularization		✓	$O(\epsilon^{-0.5})$	Mishchenko 2022, Doikov 2022
ARC	†	✓	$O(\epsilon^{-1.5}), O(\epsilon^{-0.5})$	Nesterov and Polyak 2006, Curtis et al. 2011
Trust-region Method	†	✓	$O(\epsilon^{-1.5})$	Curtis et al. 2017
Homotopy method (new)	✓	✓	$O(\log(\epsilon^{-1}))$	Luenberger and Ye 2021 Lecture notes by Ye, 2015

# Concordant Second-Order Lipschitz condition I

- Consider  $\min_x f(x)$ , where  $f(x)$  satisfies

$$\|\nabla f(x+d) - \nabla f(x) - \nabla^2 f(x)d\| \leq \beta \cdot d^T \nabla^2 f(x)d$$

whenever  $\|d\| \leq O(1)$ .

- This condition is called *the concordant second-order Lipschitz condition (CSOLC)*, first introduced in Luenberger & Ye (2015, 2022).
- CSOLC is motivated from the Scaled Lipschitz Condition, which was widely used in the IPMs and MCPs. see Zhu(1992), Kortane&Zhu(1993), Andersen&Ye(1999).

# Concordant Second-Order Lipschitz condition II

## Properties of CSOLC:

- Closed under positive scalar multiplications and summations;
- Closed under affine transformation: if  $f(x)$  satisfies CSOLC, then  $f(Ax$

## Examples of CSOLC:

- Convex quadratic functions, exponential functions;
- $\gamma$ -Regularized logistic regression:  $f(x) = \frac{1}{m} \sum_{i=1}^m \log \left( 1 + e^{-b_i \cdot a_i^T x} \right) + \frac{\gamma}{2} |x|^2$

# The Homotopy Model

- The homotopy model:

$$x_{\mu_T} = \arg \min f(x) + \frac{\mu_T}{2} \|x\|^2$$

Where  $\mu_T \rightarrow 0$ . We say  $\{X_{\mu_T}\}$  forms a central path.

- At each iterate solve the homotopy model *inexactly* (*approximate “centering” condition, ACC*):

$$\|\nabla f(x_{T,k}) + \mu_T \cdot x_{T,k}\| \leq \frac{\mu_T}{1 + 3(\beta + 1)}.$$

- Use GHM with proper  $\delta_k$  and  $\phi_k$  !

# Homotopy HSODM I

- For each homotopy model, we apply GHM to solve it:

$$\min_{\|[v;t]\| \leq 1} \begin{bmatrix} v \\ t \end{bmatrix}^T \begin{bmatrix} H_{T,k} & g_{T,k} + \mu_T \cdot x_{T,k} \\ (g_{T,k} + \mu_T \cdot x_{T,k})^T & -\mu_T \end{bmatrix} \begin{bmatrix} v \\ t \end{bmatrix}$$

- **Lemma 2(a): (fixed distance from the “central” path)**

$$\|x_{T,k} - x_{\mu_T}\| \leq \frac{1}{1 + 3(\beta + 1)}$$

- **Lemma 2(b): (finite convergence for each epoch) For any  $\mu_T$ , ACC can be satisfied within  $K \leq 2$  steps, specifically**

$$K = \left\lceil \log_2 \left( \frac{\log(1 + 3(\beta + 1)) - \log(\beta + 1)}{\log 3 - \log 2} \right) \right\rceil$$



# Homotopy HSODM II

## A Non-Interior **Homotopy** HSODM:

- Linearly decrease  $\mu_T \rightarrow$  simultaneously adaptive  $\delta_k$  and  $\phi_k$

$$\mu_{T+1} = \frac{1 + \|x_{T,k}\|}{1 + 3(1 + \beta)(1 + \|x_{T,k}\|)} \cdot \mu_T \quad \rightarrow \quad x_{T+1,0} := x_{T,k}$$

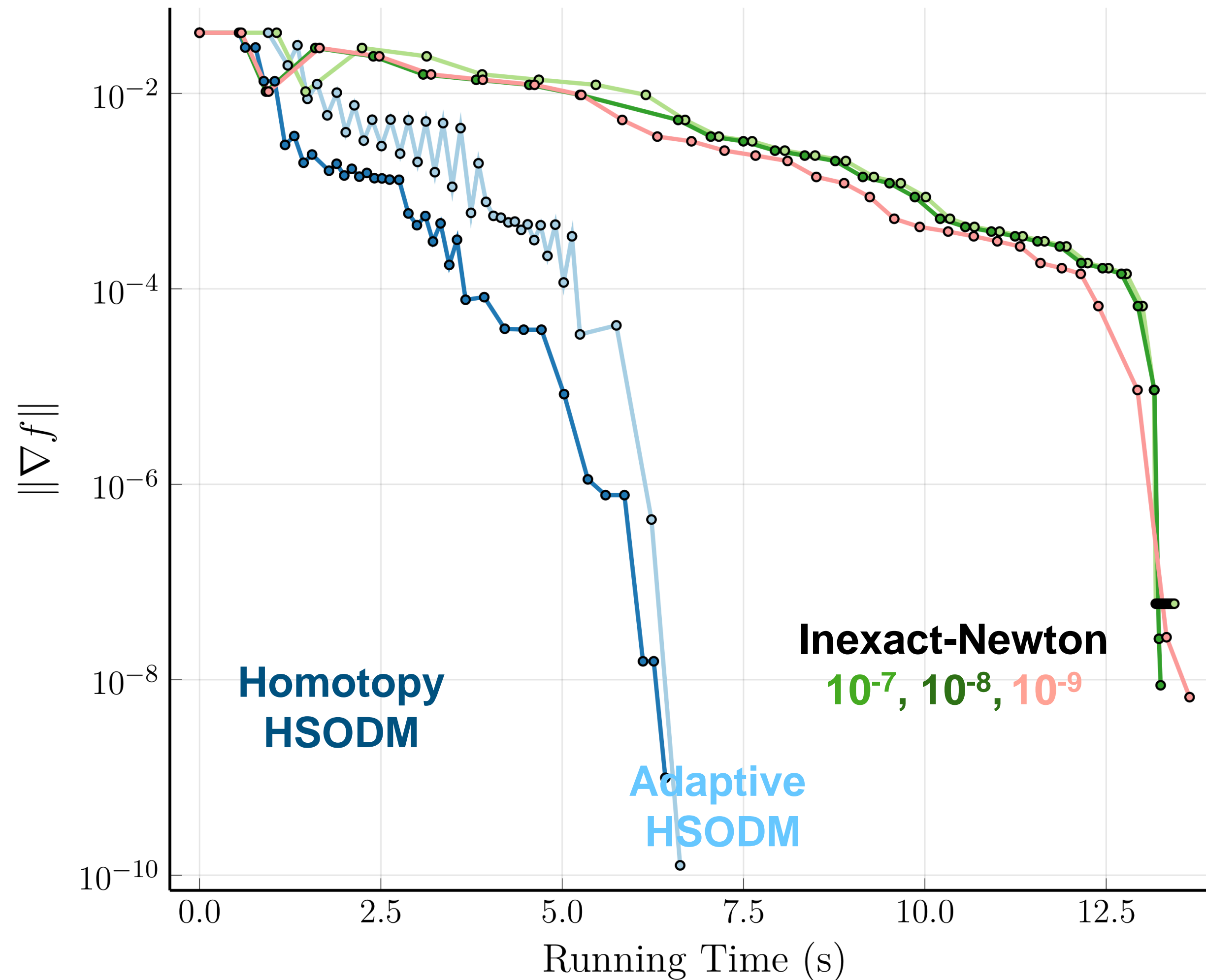
- Use GHMs as each subproblem at  $\mu_T$  with finite convergence
- **Theorem: (global rate of convergence)** After at most

$$\bar{T} = \left\lceil \log_{\tau} \left( \frac{(1 + 3(\beta + 1))\epsilon}{2(\beta + 1)(1 + \|\nabla f(0)\|^2)((3\beta + 4)\|x^*\| + 2)} \right) \right\rceil$$

**iterates**, we could find an iterate that satisfies  $|\nabla f(x_{\bar{T}+1,0})| \leq \epsilon$

# Application IV: A Comparison in $L_2$ - Logistic regression, $\gamma = 1e-5$

Logistic Regression name := rcv1,  $n := 47236$ ,  $N := 20242$



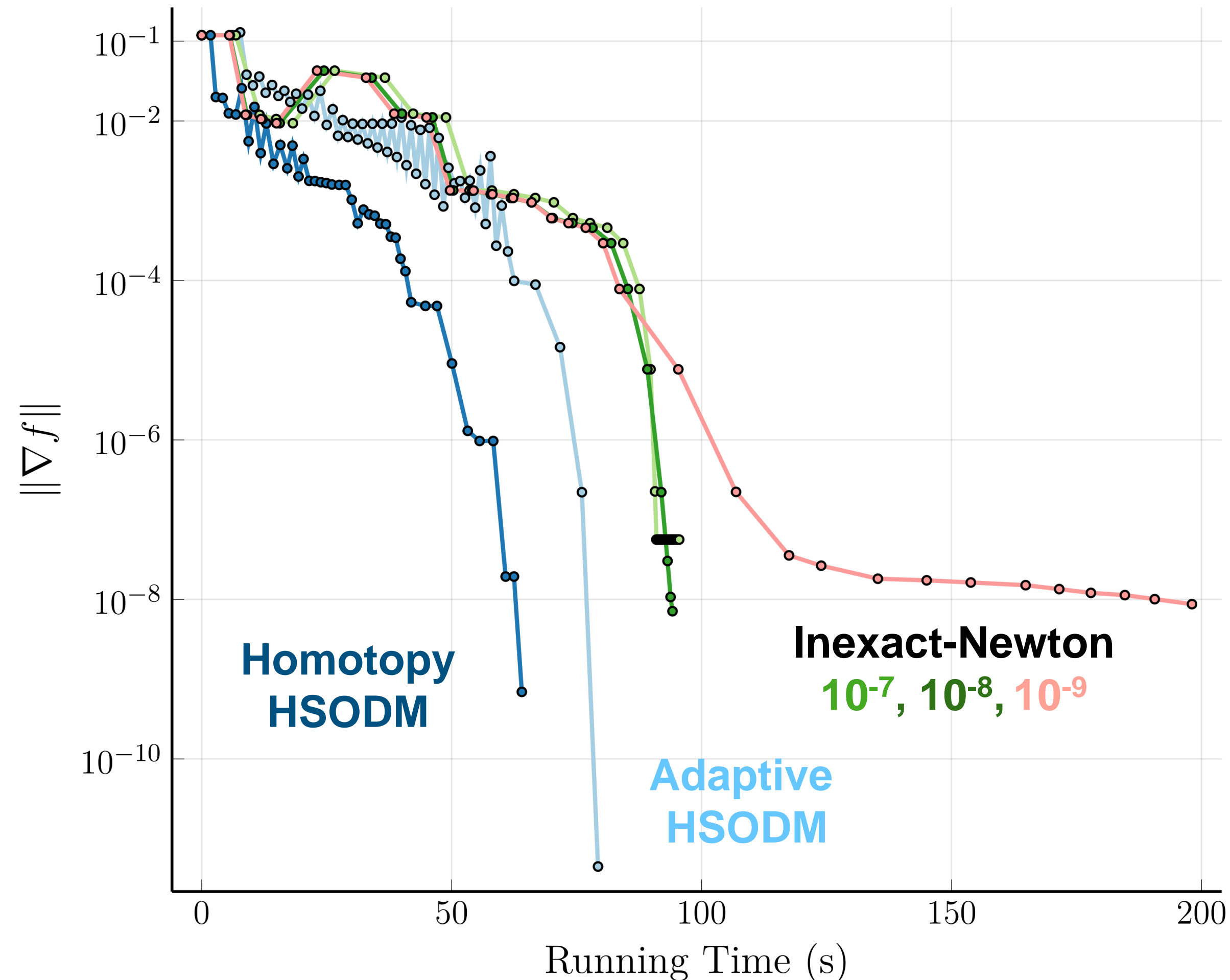
- $L_2$  -Logistic regression:

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log \left( 1 + e^{-b_i \cdot a_i^T x} \right) + \frac{\gamma}{2} |x|^2$$

- Compare **Homotopy-HSODM, Adaptive HSODM**
- and **inexact Newton** with different accuracy (public open-source code)

# A Comparison in $L_2$ - Logistic regression, $\gamma = 1e-5$

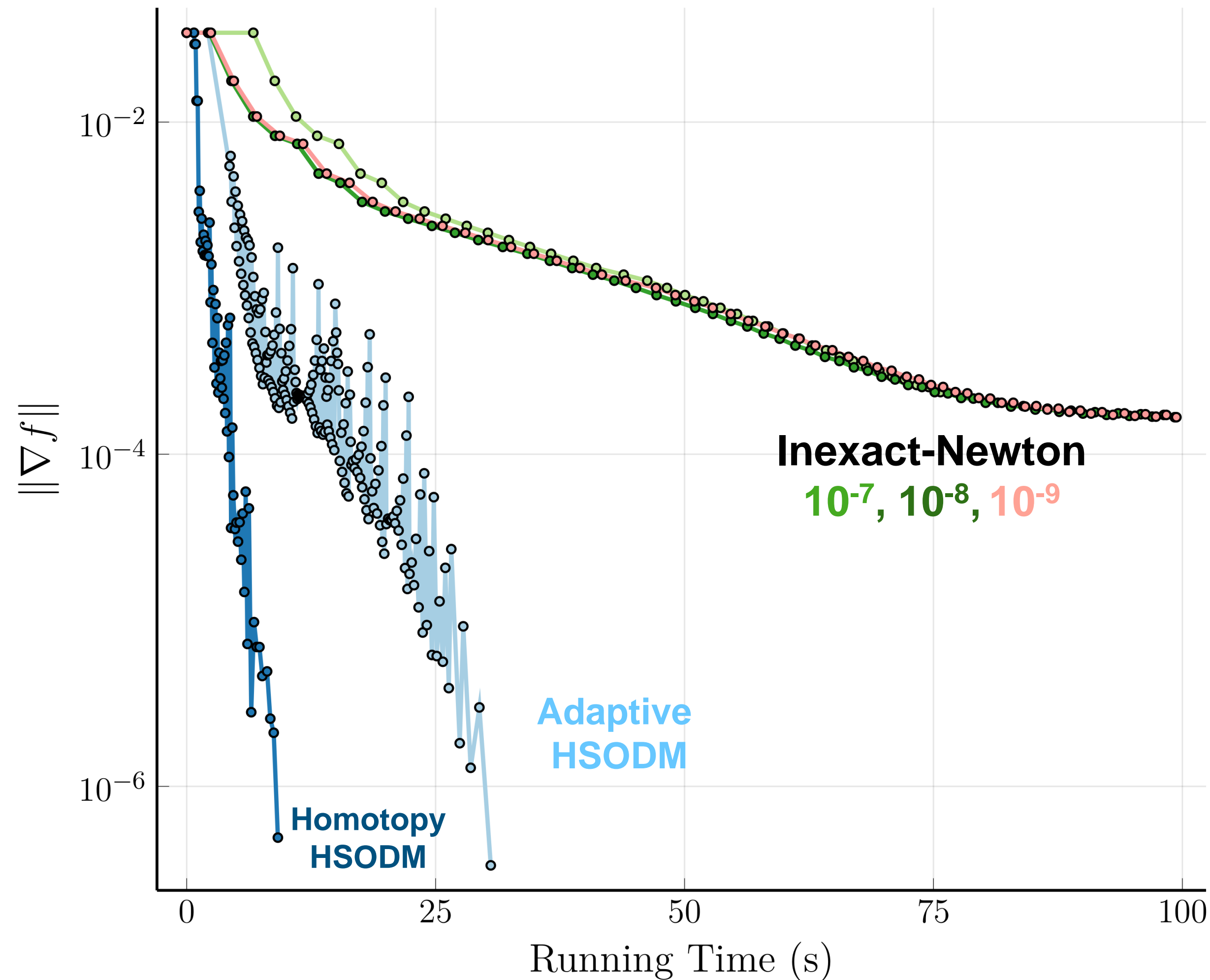
Logistic Regression name := news20,  $n := 1355191$ ,  $N := 19996$



- A larger dataset **news20**
- Large dimension but relatively few data
- HSODM can benefit when dimension  $n$  gets large
- Similar results were observed in Rojas 2001, Adachi 2017 for solving Trust-region Subproblems.

# Resilience of Homotopy-HSODM for small $\gamma$ , $\gamma = 1e-7$

Logistic Regression name := rcv1,  $n := 47236$ ,  $N := 20242$



- With same dataset **rcv1**

$$f(x) = \frac{1}{m} \sum_{i=1}^m \log \left( 1 + e^{-b_i \cdot a_i^T x} \right) + \frac{\gamma}{2} |x|^2$$

- Sensitivity study from  $\gamma = 1e-5 \rightarrow 1e-7$
- Homotopy-HSODM is resilient to small  $\gamma$  (almost degenerate case)

# Takeaways

**Homogeneous second-order direction as an extreme eigenvalue computation is a “cheaper” alternative to the Trust-Region or Newton step computation**

**Generalized Homogeneous direction is flexible using different  $\delta_k$  and  $\phi_k$  and substitutes for other SOM step**

**Happy Retirement, Kurt!**