Sandia National Laboratories

Exceptional service in the national interest

HYBRID METHODS FOR TENSOR DECOMPOSITIONS THAT LEVERAGE STOCHASTIC AND DETERMINISTIC OPTIMIZATION

Jeremy M. Myers[1,2], Daniel M. Dunlavy[1]

June 2, 2023

MS250: Geometric Perspectives in Optimization

SIAM Conference on Optimization (OP23)

Seattle, WA

[1]Sandia National Laboratories, Albuquerque, NM and Livermore, CA
[2]College of William and Mary, Williamsburg, VA

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly owned subsidiary of Honeywell International Inc. for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

SAND2023-04027C

U.S. DEPARTMENT OF ENERGY

NNSA

MOTIVATING STOCHASTIC + DETERMINISTIC TENSOR ALGORITHMS

Recent trend in theoretical computer science & numerical linear algebra (and elsewhere!):

- Use randomization to solve very large, hard problems
  - data mining, information science, compression, scientific computing
- Often faster with equivalent levels of error
- Examples: low-rank matrix decompositions, streaming, regression, linear systems [1]

Typical approach: use stochasticity for a fast approximation and determinism for refinement to yield effective algorithms with theoretical guarantees.

How can we extend the existing approaches to **low-rank tensor decompositions**?

[1] Martinsson and Tropp, Randomized Numerical Linear Algebra: Foundations & Algorithms, *Acta Numerica*, 2020.

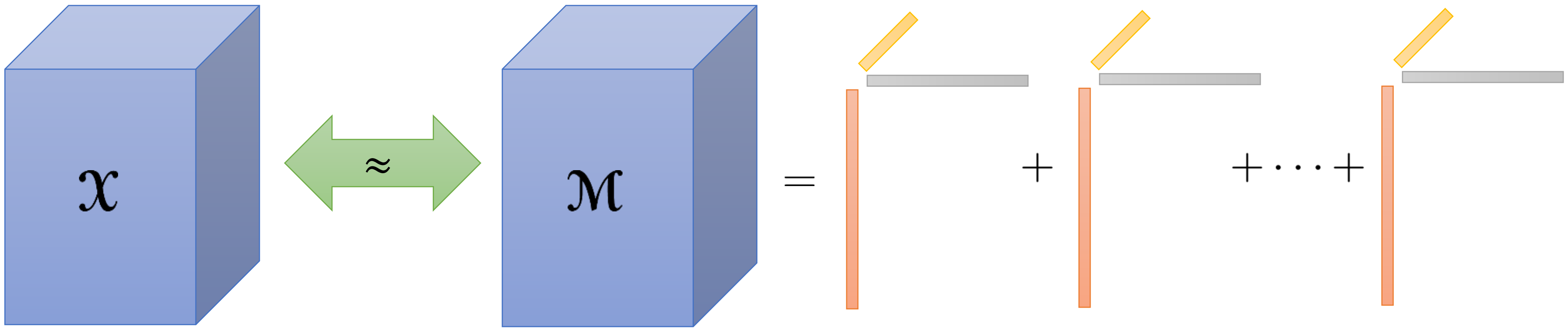
LOW-RANK TENSOR APPROXIMATION

Canonical polyadic decomposition (CPD)

Data

CP model

Sum of r outer product tensors



$$\mathcal{X} \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d} \qquad \mathcal{M} = [\![\mathbf{A}_1, \mathbf{A}_2, \ldots, \mathbf{A}_d]\!] \in \mathbb{R}^{n_1 \times n_2 \times \cdots \times n_d}$$

- $\mathcal{X}$ is the data tensor in d dimensions or modes.
- $\mathcal{M}$ is the model tensor.
- $\mathbf{A}_k$ is an $n_k \times r$ factor matrix.
- $\mathbf{i} = (i_1, i_2, \ldots, i_d)$ is a multi-index

Low-rank CPD

- Assume $\operatorname{rank}(\mathcal{X}) = r$.
- Typically choose $r \ll \min\{n_1, n_2, \ldots, n_d\}$.

Poisson CPD

$$\mathcal{X}_{\mathbf{i}} \sim \text{Poisson}(\mathcal{M}_{\mathbf{i}})$$

LOW-RANK TENSOR APPROXIMATION

Poisson tensor maximum likelihood estimation

Statistical method to compute low-rank Poisson CPD

$$\min_{\mathcal{M}} f_{\mathcal{X}}(\mathcal{M}) = \min \sum_{\mathbf{i}} m_{\mathbf{i}} - x_{\mathbf{i}} \log(m_{\mathbf{i}})$$

$$\text{where } a_{i_1}^{(1)} a_{i_2}^{(2)} \dots a_{i_d}^{(d)} = m_{\mathbf{i}} \text{ are the optimization variables}$$

- This a **nonlinear, nonconvex** optimization problem.
- The **maximum likelihood estimator (MLE)** corresponds to the **global optimizer** $\mathcal{M}^*$ for this problem.
- The typical approach is to *flatten* or *unfold* the tensors into matrices and use **local** methods.
  - Stochastic: Generalized Canonical Polyadic (GCP) tensor decomposition [2, 3]
  - Deterministic: Canonical Polyadic Alternating Poisson Regression (CPAPR) [4]

[2] Hong, Kolda, and Duersch, Generalized Canonical Polyadic Tensor Decomposition, *SIAM Review*, 2020

[3] Kolda and Hong, Stochastic Gradients for Large-Scale Tensor Decomposition, *SIAM Journal on Mathematics of Data Science*, 2020

[4] Chi and Kolda, On Tensors, Sparsity, and Nonnegative Factorizations, *SIAM Journal on Matrix Analysis and Applications*, 2012

OUR NEW STOCHASTIC + DETERMINISTIC TENSOR ALGORITHMS

How can current local methods be leveraged together to improve likelihood of finding the MLE/global optimizer?

Proposed methods

Hybrid GCP-CPAPR

- Inspired by Simulated Annealing.
- Improves probability of convergence to global optimizer and reduces cost compared to standalone methods.

Restarted CPAPR with LookAhead

- Uses novel heuristic to avoid suboptimal solutions w.r.t. global optimizer.
- Saves computation by restarting when the iterates are detected to be headed to a suboptimal solution.

HYBRID GCP-CPAPR

Hybrid GCP-CPAPR (HybridGC) intuition

1. Use stochastic optimization to compute a fast approximate solution.
2. Use deterministic optimization to refine approximate solution.

```
Algorithm HybridGC(tensor $\boldsymbol{\mathcal{X}}$, rank $r$, initial guess $\boldsymbol{\mathcal{M}}_0$)
    $\boldsymbol{\mathcal{M}}_1 \leftarrow \mathrm{GCP}(\boldsymbol{\mathcal{X}}, r, \boldsymbol{\mathcal{M}}_0)$
    $\boldsymbol{\mathcal{M}}_2 \leftarrow \mathrm{CPAPR}(\boldsymbol{\mathcal{X}}, r, \boldsymbol{\mathcal{M}}_1)$
    return model tensor $\widehat{\boldsymbol{\mathcal{M}}} = \boldsymbol{\mathcal{M}}_2$ as estimate to $\boldsymbol{\mathcal{M}}^*$
```

NUMERICAL EXPERIMENTS WITH HYBRID GCP-CPAPR

Methodology

1. Generate N random starting points.
2. Compute decompositions with CPAPR & GCP separately.
3. HybridGC step: refine GCP decompositions with CPAPR.
4. Analyze average behavior of our experiments.

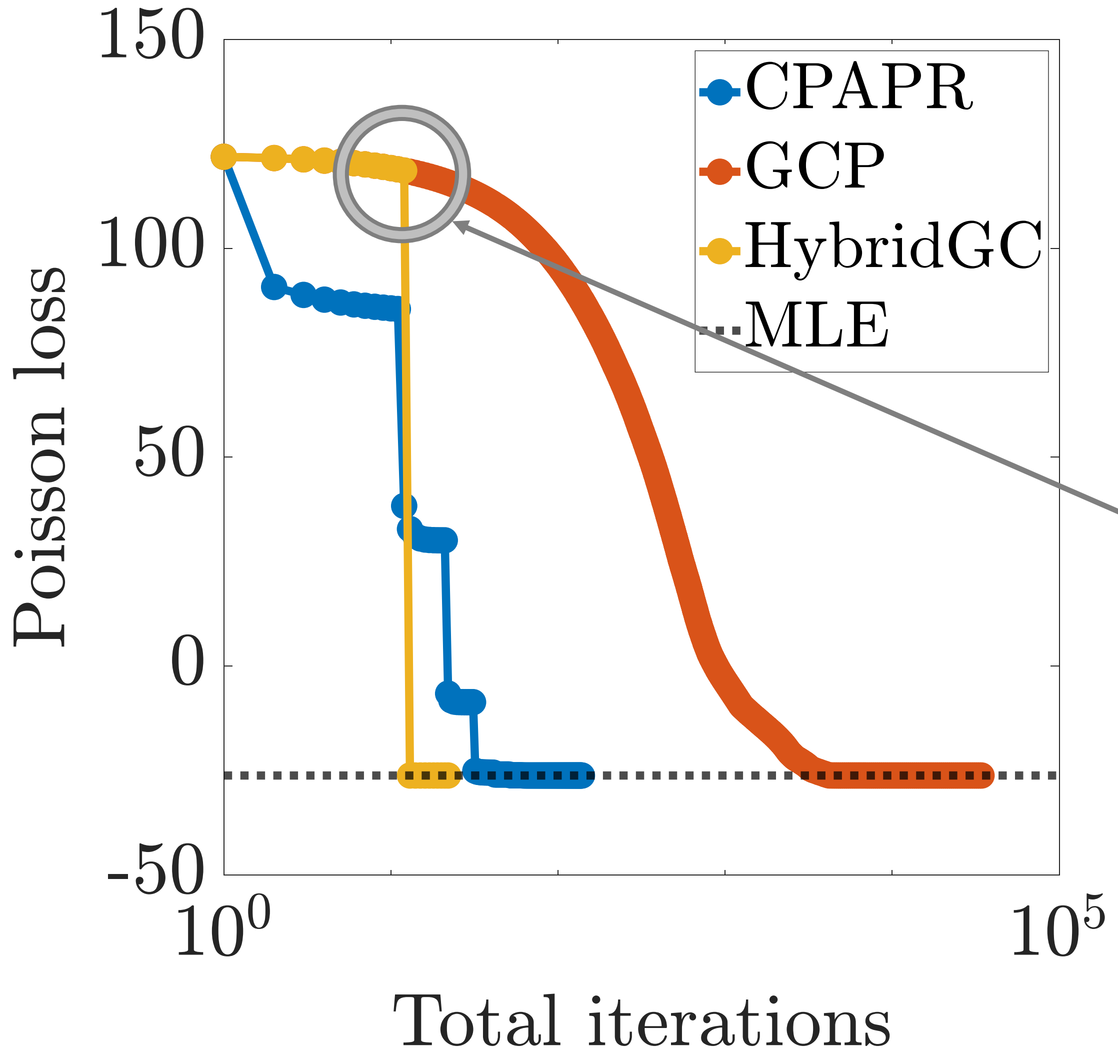
Datasets

1. **Small**: 4 x 6 x 8, 17 nonzeros, $r = 3$, $N > 110$k
2. **Large**: 1k x 1k x 1k, 98k nonzeros, $r = 20$, $N = 100$

Error measures

- Based on loss function values.
- Probability estimate of finding MLE/global optimizer.
- Spectral properties of unfolded tensor.

HYBRID GCP-CPAPR RESULTS: OPTIMIZATION VARIABLES VIEW



Small dataset

- HybridGC initially has the same search path as GCP and makes slow progress.
- Upon switch to deterministic solver, HybridGC converges in fewer iterations than either GCP or CPAPR alone.

PROBABILITY OF FINDING MAXIMUM LIKELIHOOD ESTIMATOR (MLE)

Small dataset

ϵ	CPAPR	GCP	HybridGC
10^{-1}	0.964	0.963	**0.968**
10^{-2}	0.964	0.963	**0.968**
10^{-3}	0.964	0.880	**0.968**
10^{-4}	0.964	0.003	**0.968**
10^{-5}	0.964	0.001	**0.968**

Large dataset

ϵ	CPAPR	GCP	HybridGC
10^{-1}	1.000	0.993	1.000
10^{-2}	0.414	0.023	**1.000**
10^{-3}	0.049	0.000	**0.880**
10^{-4}	0.001	0.000	**0.190**
10^{-5}	0.000	0.000	**0.020**

Relative distance from MLE

$$\epsilon = \frac{|f\mathcal{X}(\widehat{\mathcal{M}}) - f\mathcal{X}(\mathcal{M}^*)|}{|f\mathcal{X}(\mathcal{M}^*)|}$$

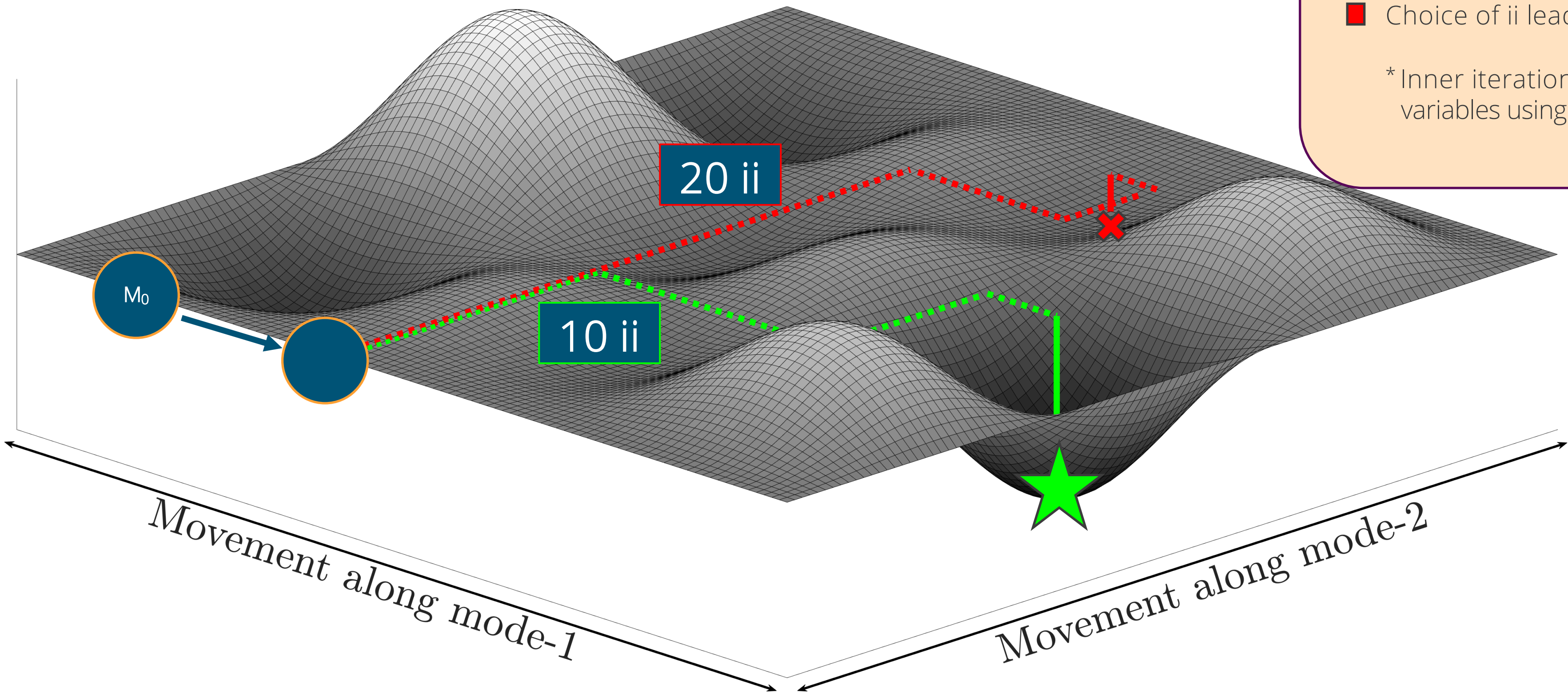
For small choices of ϵ, HybridGC is the most likely to estimate MLE/global optimizer.

OUR FUNDAMENTAL (YET SIMPLE) QUESTION

Why and when do these methods fail?

We'll try to answer this for CPAPR.

CONVERGENCE DEPENDS ON NUMBER OF STEPS IN SEARCH DIRECTION



KEY

ii Maximum # inner iterations*

☆ Maximum likelihood estimator (MLE)

✖ Local minimizer

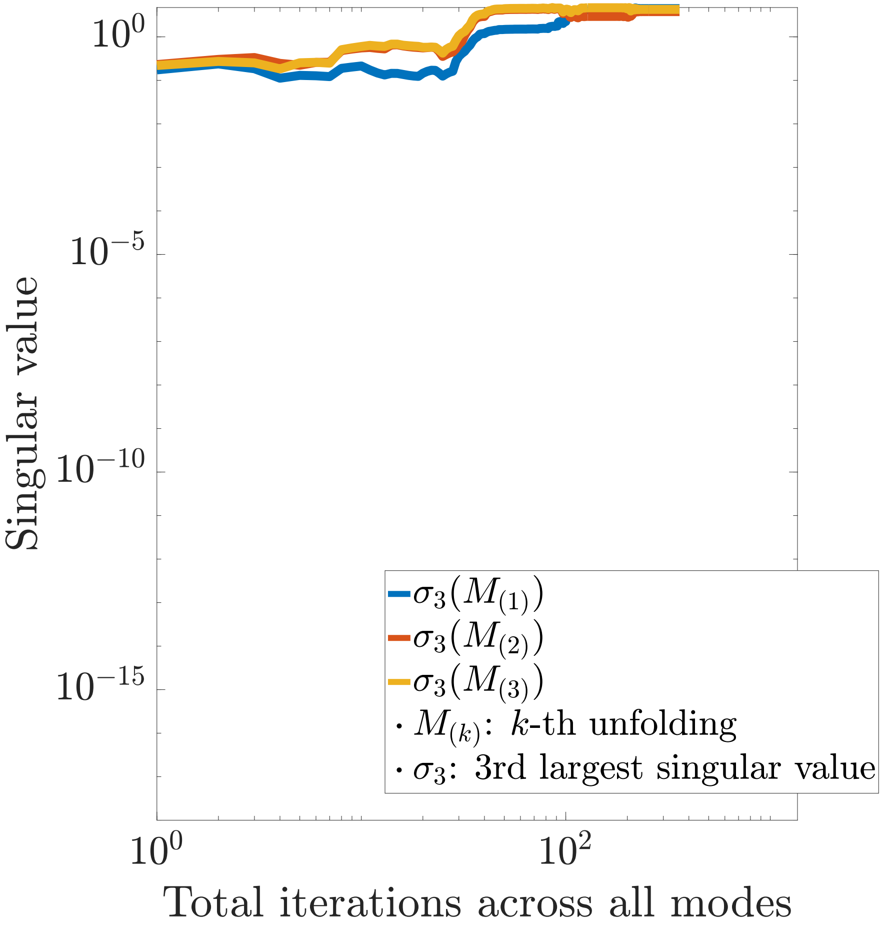
🟩 Choice of ii leads to MLE

🟥 Choice of ii leads to local minimizers

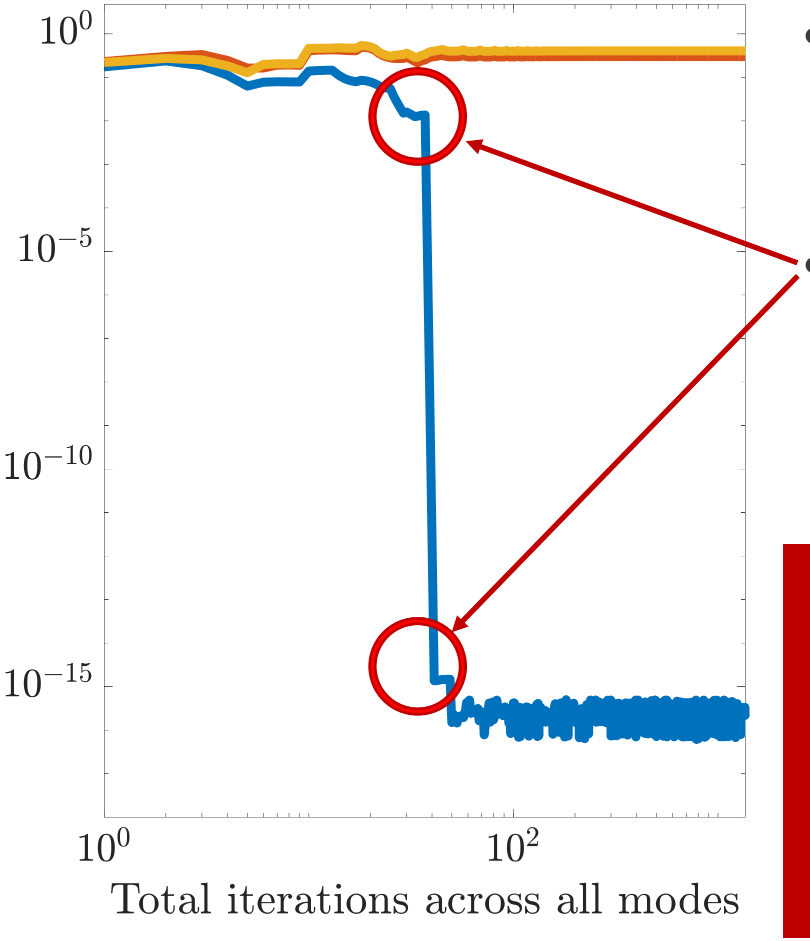
[*] Inner iteration = update of subset of variables using multiplicative updates

CHALLENGING BEHAVIOR ON SMALL DATASET

"Goldilocks" # of inner iterations; leads to MLE



Too many inner iterations; leads to other local minimizer



- CPAPR can be very sensitive to the number of steps in search direction.
- **Spectral property**: the ratio of successive singular values may indicate there is a problem.

What if we looked ahead to see if current path is heading toward suboptimal solution?

RESTARTED CPAPR WITH LOOKAHEAD (OVERVIEW)

Choose the following parameters:

- k_{max}: Maximum number of outer iterations
- l_{max} : Maximum number of inner iterations
- j: Number of steps to look ahead ($j \leq l_{max}$)
- γ: Maximum threshold of spectral properties for acceptable search path (e.g., $\gamma = \mathbf{10^6}$)

While (not converged), compute a rank-R decomposition with CPAPR:

1. Every step, update current model.
2. Every j steps, compute spectral properties of current model.
3. If (spectral properties) $< \gamma$, checkpoint and continue.
4. Otherwise, choose a new initial guess and **restart**.

RESTARTED CPAPR WITH LOOKAHEAD (DETAILED)

Choose the following parameters:

- k_{max}: Maximum number of outer iterations
- l_{max} : Maximum number of inner iterations
- j: Number of steps to look ahead ($j \leq l_{max}$)
- γ: Maximum threshold of spectral properties for acceptable search path (e.g., $\gamma = 10^6$)

1. Choose an initial guess
2. While not converged, compute a rank-R decomposition with CPAPR:
   a. At the i-th iteration in mode-k, compute the R-th largest singular value $\sigma_{(k)}[R]^{(i)}$.
   b. Proceed for j iterations.
   c. At the $(i + j)$-th iteration in mode-k, compute the R-th largest singular value $\sigma_{(k)}[R]^{(i+j)}$.
   d. If $\sigma_{(k)}[R]^{(i)}/\sigma_{(k)}[R]^{(i+j)} < \gamma$, set $\sigma_{(k)}[R]^{(i)} \leftarrow \sigma_{(k)}[R]^{(i+j)}$ and continue.
   e. Otherwise, **restart**: go to 1.

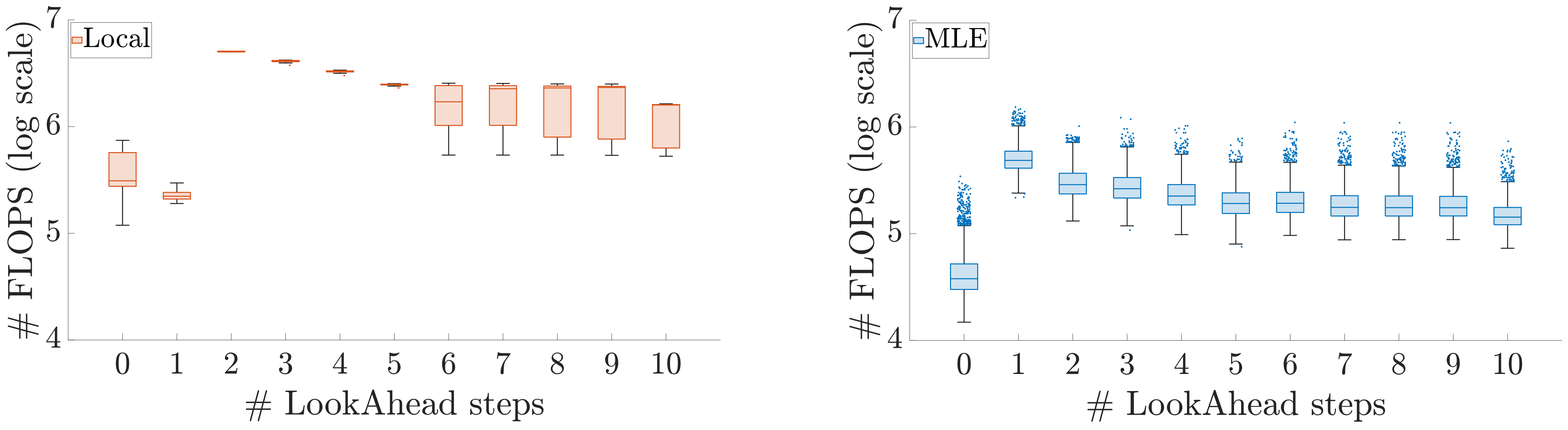
CONVERGENCE AND PERFORMANCE RESULTS

Probability of convergence to MLE vs. local minimizer with LookAhead, $\epsilon = 10^{-8}$

Previous best had relative distance $\epsilon = 10^{-5}$

# Steps	0	1	2	3	4	5
Local	3.6738×10^{-2}	2.4486×10^{-4}	1.8138×10^{-5}	1.4510×10^{-4}	2.0859×10^{-4}	1.9952×10^{-4}
MLE	9.6326×10^{-1}	9.9976×10^{-1}	9.9998×10^{-1}	9.9985×10^{-1}	9.9979×10^{-1}	9.9980×10^{-1}
# Steps	6	7	8	9	10	
Local	1.3150×10^{-3}	7.3459×10^{-4}	6.1669×10^{-4}	5.5321×10^{-4}	4.8066×10^{-4}	
MLE	9.9868×10^{-1}	9.9927×10^{-1}	9.9938×10^{-1}	9.9945×10^{-1}	9.9952×10^{-1}	

Computational cost of convergence to MLE vs. local minimizer with LookAhead



CONCLUSIONS

- LookAhead has the highest likelihood of finding MLE in our experiments.
- The method can be prohibitively expensive when it does fail, but this is rare.

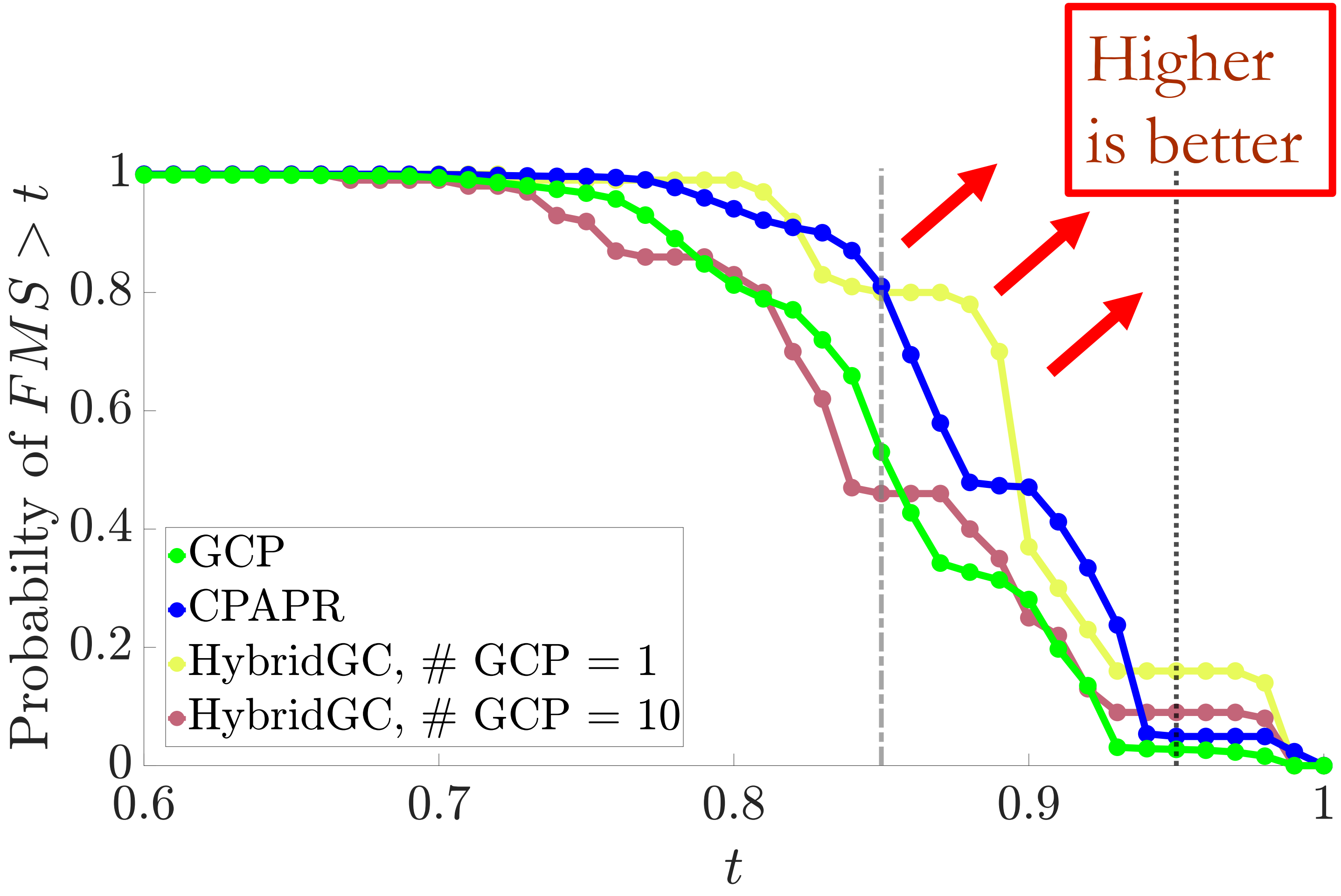
FUTURE WORK

- It's unclear how sensitive LookAhead is to the complex interplay parameters.
- Experiments on **Small** dataset are very limited – do they generalize?
- Are low-accuracy singular values useful?

Contact: {jermyer, dmdunla}@sandia.gov
Paper (to be updated soon): https://arxiv.org/abs/2207.14341

BACKUP SLIDES

HYBRID GCP-CPAPR RESULTS: ALGEBRAIC STRUCTURES VIEW



HybridGC is better when using fewer iterations of GCP (yellow line = 1) than more iterations of GCP (red line = 10) when using more iterations.