Subset Gaussian Cloning and Connections With High Dimensional RRT.

Max Lovig

June 15, 2025

Ilias Zadik

Conor Sheehan

Kostas Tsirkas

n participants, m questions.

Warner's additive model outputs the matrix,

	P_1	P_2	$\ldots$	P_{m-1}	P_m
Q_1	0.12	-1.05	$\ldots$	-0.43	0.27
Q_2	-0.78	0.46	$\ldots$	0.55	-0.31
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
Q_{m-1}	-0.24	0.89	$\ldots$	0.07	-0.19
Q_m	0.45	-0.67	$\ldots$	0.78	0.11

Question vector Q and response vector P.

Adding structure to the signal!

Let $P = \mu \in \{-1, 0, 1\}^n$ with $\|\mu\|_0 = k$.

- ± 1 are two differing perspectives.
- 0 is a "moderate" or "ambivalent" perspective with respect to ± 1

Which $Q \in \mathbb{R}^n$ vector to consider?
$Q = \mathbf{1}$ is equivalent to:

"Do you agree with perspective 1?"

"1's" $\rightarrow 1$, "-1's" $\rightarrow -1$ and "0's" $\rightarrow 0$

Question vector $= \mathbf{1}$,

Response vector $= \mu$,

Warner's model becomes,

$$Y = \lambda \frac{\mathbf{1}\mu^{\top}}{n^{1/2}k^{1/2}} + G$$

Question vector $= \mathbf{1}$,

Response vector $= \mu$,

Warner's model becomes,

$$Y = \underbrace{\lambda}_{\text{SNR}} \underbrace{\frac{\mathbf{1}\mu^\top}{k^{1/2} n^{1/2}}}_{\text{Rescaled Signal}} + \underbrace{G}_{\substack{\text{Warner's} \\ \text{Additive Noise}}},$$

$$\|\mu\|_0 = k \text{ and } G_{i,j} \sim \mathcal{N}(0,1)$$

Instead of $\mathbf{1}$, consider $q \in \{-1, 0, 1\}^m$ (possible random) with $\|q\|_0 = \ell$.

Warner's matrix $Y = Y_{i,j} = q_i \mu_j$.

Re-normalizing,

$$Y = \lambda \frac{q\mu^\top}{\ell^{1/2} k^{1/2}} + G$$

Let $i \in [m]$ and $j \in [n]$.

$$q_i \neq \mu_j \neq 0 \text{ gives } Y_{i,j} = \frac{\lambda}{\sqrt{\ell k}} + G_{i,j}.$$

$$q_i = \mu_j \neq 0 \text{ gives } Y_{i,j} = \frac{-\lambda}{\sqrt{\ell k}} + G_{i,j}.$$

$$q_i = 0 \text{ or } \mu_j = 0 \text{ gives } Y_{i,j} = G_{i,j}.$$

$$Y = \lambda \frac{q\mu^\top}{\ell^{1/2} k^{1/2}} + G$$

$$(m = n) \qquad \downarrow \quad (?)$$

$$Y = \lambda \frac{\mu\mu^\top}{k^{1/2} k^{1/2}} + G$$

$$\downarrow \quad (\textit{tensor})$$

$$Y = \lambda \frac{\mu^{\otimes r}}{k^{r/2}} + G$$

In what follows, λ large allows exact recover of μ from Warner's matrix.

Why is this problematic in RRT?

We turn to the model,

$$Y = \lambda \frac{\mu^{\otimes r}}{k^{r/2}} + G,$$

$$Y \in \mathbb{R}^{n^{\otimes r}},\ \lambda_n \geq 0,\ G_{i_1,\ldots,i_r} \sim \mathcal{N}(0,1),$$
$$\mu \in \{-1,0,1\}^n \text{ and } \|\mu\|_0 = k$$

- *Many applications:* tensor PCA, compressed sensing, community detection.

- **More Importantly,** in this model exhibits a stat-comp-local gap.

First, we look at STAT.

Question?

Let ν be uniform over
$\{v \in \{-1, 0, 1\}^n : \|v\|_0 = k\}$ be the prior.

What is the optimal (in terms of exact recovery) statistical estimator of μ.

Maximum a posteriori, i.e.

$$\hat{\mu} = \arg\max_{\mu' \in \mu} \langle Y, (\mu')^{\otimes r} \rangle.$$

A Stat-Comp gaps occur when the Bayes estimator requires SNR λ but the best known "poly-time" algorithm requires SNR $\lambda' = \omega(\lambda)$.

A Stat-Comp gaps occur when the Bayes estimator requires SNR λ but the best known "poly-time" algorithm requires SNR $\lambda' = \omega(\lambda)$.

For our problem of focus,

$$\mathrm{opt}_{\mathrm{STAT}}(\lambda) = \Theta(\sqrt{k}\,\mathrm{p\text{-}log}(n))$$

and

$$\mathrm{opt}_{\mathrm{COMP}}(\lambda) = \Theta((k^{r/2} \wedge n^{r/4})\mathrm{p\text{-}log}(n)).$$

A Comp-Local gap occurs when the best known (non-"local") poly-time algorithm requires SNR λ', but there is evidence that a "local" poly-time algorithm requires $\tilde{\lambda} = \omega(\lambda')$.

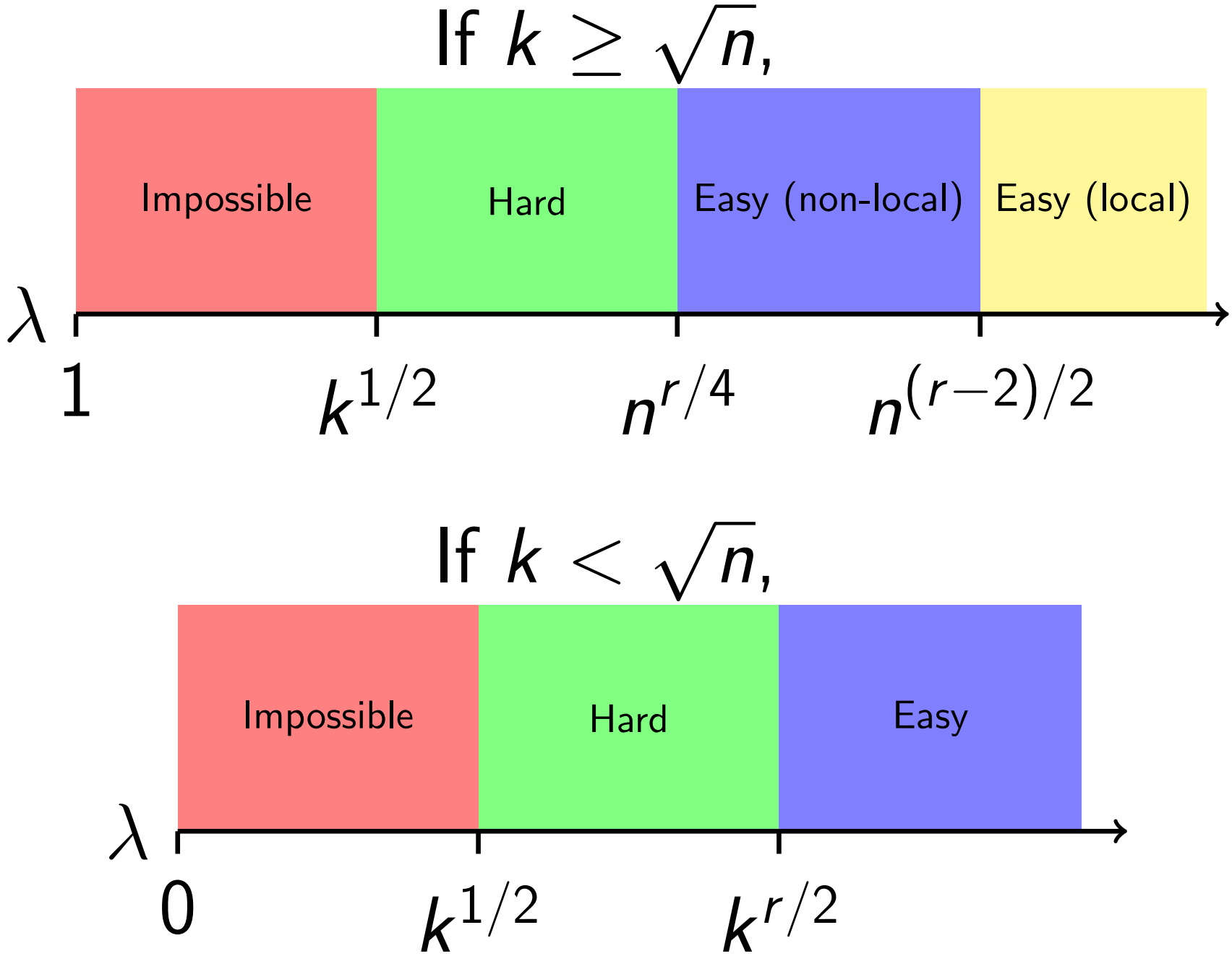
A Comp-Local gap occurs when the best known (non-"local") poly-time algorithm requires SNR λ', but there is evidence that a "local" poly-time algorithm requires $\tilde{\lambda} = \omega(\lambda')$.

Recalling the previous slide,

$$\mathrm{opt}_{\mathrm{COMP}}(\lambda) = (k^{r/2} \wedge n^{r/4})\mathrm{p\text{-}log}(n)$$

and

$$\mathrm{opt}_{\mathrm{LOCAL}} = (n^{(r-2)/2})\mathrm{p\text{-}log}(n).$$


If $k \geq \sqrt{n}$,
Impossible
Hard
Easy (non-local)
Easy (local)
λ
1
$k^{1/2}$
$n^{r/4}$
$n^{(r-2)/2}$
If $k < \sqrt{n}$,
Impossible
Hard
Easy
λ
0
$k^{1/2}$
$k^{r/2}$

Generating correlated noise,

Require: $M \in \mathbb{N}$

1: **for** $i = 1$ to n **do**

2: Sample $G_i^1, \ldots, G_i^M$ i.i.d. from $\mathcal{N}(0, M)$.

3: $\bar{G}_i \leftarrow \frac{1}{M} \sum_{j=1}^{M} G_i^j$.

4: **for** $t = 1$ to M **do**

5: $Z_{i,t} \leftarrow G_i^t - \bar{G}_i$.

6: **end for**

7: **end for**

8: **return** $((Z_i^t)_{i \in [n], t \in [M]})$

Randomized Greedy with random thresholds,

Require: $Y \in \mathbb{R}^{n^{\otimes r}}$, $S_0 \in \{-1,0,1\}^n$, $\gamma \geq 0, M \in \mathbb{N}$

$t \leftarrow 0$ and $t_i \leftarrow 0$ for each $i \in [n]$.

Let $((Z_i^t)_{i\in[n],t\in[M]})$ be the output of the above algorithm with input M

while $t \leq M$ **do**

  Let $\mathcal{N}(S_t) = \{\sigma' \in \{-1,0,1\}^n \mid d_H(S_t, \sigma') = 1\}$

  Choose a random element $\sigma' \in \mathcal{N}(S_t)$ uniformly

  Let $i \in [n]$ be such that $\sigma'_i \neq (S_t)_i$.

  $t_i \leftarrow t_i + 1, t \leftarrow t + 1$

  if $H_{\frac{r+1}{2},\gamma}(\sigma') - H_{\frac{r+1}{2},\gamma}(S_t) > 0$

    $S_{t+1} \leftarrow \sigma'$

  end if

end while

return S_{M+1}

Above $H_{\frac{r+1}{2},\gamma}(S) = \langle S, Y \rangle - \gamma \|S\|_0^{\frac{r+1}{2}}$.

For simplicity, $\lambda = \Omega(\sqrt{k}n^{(r-1)/4})$.

(1) Justify the following heuristic: If $\cos(\mu, S_1)^{r-1} \geq \frac{\sqrt{k}}{\lambda}\text{p-log}(n)$, then our algorithm recovers μ.

(2) Find an initialization where $\cos(\mu, S_1) = n^{-1/4}$.

(3) By rearranging the inequality in (1), we then have $\lambda \geq \sqrt{k}n^{\frac{r-1}{4}}\text{p-log}(n)$, as desired.

Parameterize the Hamming distance one transition by choosing a coordinate p_t of S_t to change to $q_t \neq S_t$.

Marginally,

$$H_{\frac{r+1}{2},\gamma}(S') - H_{\frac{r+1}{2},\gamma}(S_t) \approx$$

$$(q_t - (S_t)_{p_t})\mu_{p_t} \frac{\lambda}{k^{r/2}} \langle \theta, S_t \rangle^{r-1} + \langle S_t, G \rangle - (|q_t| - |S^t_{p_t}|)\gamma \|S^t\|_0^{(r-1)/2}$$

$$= \frac{\lambda}{\sqrt{k}} (q_t - (S_t)_{p_t})\mu_{p_t} \cos(S_t, \theta)^{r-1} - (|q_t| - |S^t_{p_t}|)\gamma + Z_t$$

So why not bound Z_t and do case work?

Why can't you simply bound Z_t?

- First, the random variables Z_t and $Z_{t'}$ are correlated as S_t is close to $S_{t'}$.

- Second, we have to bound Z_t over all possible paths of this walk. If we have made n transitions, this set of possible paths grow exponentially.

Alternately, we could use sub-set Gaussian cloning to establish $\lambda \geq \sqrt{k} n^{\frac{r-1}{4}} \mathrm{p\text{-}log}(n)$.

Intuition

- Many algorithms reuse noise matrix G, causing correlations between steps.
- To avoid the complexity, we could assume: $\tilde{Y}_t = \frac{\lambda}{k^{r/2}}\theta^{\otimes r} + \tilde{G}_t$ with $\tilde{G}_t \sim \mathcal{N}(0, \cdot)$, for each step t.
- Solution: design an algorithm that simulates fresh-noise behavior.

Spot the step which simulates fresh noise?

Require: $Y \in \mathbb{R}^{n^{\otimes r}}, S_0 \in \{-1,0,1\}^n, \gamma \geq 0, M \in \mathbb{N}$

1: $t \leftarrow 0$ and $t_i \leftarrow 0$ for each $i \in [n]$.
2: Let $((Z_i^t)_{i\in[n],t\in[M]})$ be the output of the above algorithm with input M
3: **while** $t \leq M$ **do**
4: Let $\mathcal{N}(S_t) = \{\sigma' \in \{-1,0,1\}^n \mid d_H(S_t, \sigma') = 1\}$
5: Choose a random element $\sigma' \in \mathcal{N}(S_t)$ uniformly
6: Let $i \in [n]$ be such that $\sigma'_i \neq (S_t)_i$.
7: $t_i \leftarrow t_i + 1, t \leftarrow t + 1$
8: **if** $H_{\frac{r+1}{2},\gamma}(\sigma') - H_{\frac{r+1}{2},\gamma}(S_t) > \|(\sigma')^{\otimes r} - (S_t)^{\otimes r}\|_F Z_i^{t_i}$ **then**
9: $S_{t+1} \leftarrow \sigma'$
10: **end if**
11: **end while**
12: **return** S_{M+1}

General setup

- Model: $Y_i = \mu_i^* + Z_i$, with $Z_i \sim \mathcal{N}(0, 1)$.
- Input: subsets $\Delta_t \subseteq [N]$, and a total number of draws M.
- Goal: output $X_{\Delta_t} \sim \mu^*_{\Delta_t} + \mathcal{N}(0, M \cdot I)$, independently across t.

Input: $Y \in \mathbb{R}^N$, subsets Δ_t, usage cap M.

1. For each $i \in [N]$, sample $G_i^1, \ldots, G_i^M \sim \mathcal{N}(0, M)$.

2. Compute $\bar{G}_i = \frac{1}{M} \sum_{j=1}^{M} G_i^j$.

3. Initialize $b_i^0 = 0$, track how many times i is used.

4. At step t:

$$\text{For each } i \in \Delta_t, \text{ set } b_i^t = b_i^{t-1} + 1$$

$$\text{Otherwise, set } b_i^t = b_i^{t-1}$$

$$X_{\Delta_t} = Y_{\Delta_t} + \left(G_{\Delta_t}^{b^t} - \bar{G}_{\Delta_t} \right)$$

5. Stop when any element is used $> M$ times (i.e. we can't inject any more fresh noise). Denote this time as T_M.

Theorem: For all $t \in [T_M]$,

$$X_{\Delta_t} \stackrel{d}{=} \mu^*_{\Delta_t} + Z'_{\Delta_t}, \quad Z'_{\Delta_t} \sim \mathcal{N}(0, M\mathrm{Id}_{|\Delta_t|})$$

- Each X_{Δ_t} is independent of the others.
- Outputs simulate fresh Gaussian noise across steps.
- Cost: variance of noise is inflated by M.

$$\lambda\sqrt{k}(q_t-(S_t)_{p_t})\mu_{p_t}\cos(S_t,\theta)^{r-1}-(|q_t|-|S^t_{p_t}|)\gamma+Z'_t\geq 0$$

Standard asymptotic theory gives
$\max_t |Z'_t| \leq 2\sqrt{M\log(n)}$ wp $1-o(1)$,

$$\lambda\sqrt{k}(q_t-(S_t)_{p_t})\mu_{p_t}\cos(S_t,\theta)^{r-1}\geq \pm 2\sqrt{M\log(n)}+\gamma(|q_t|-|S^t_{p_t}|).$$

$$\equiv (q_t-(S_t)_{p_t})\mu_{p_t}\cos(S_t,\theta)^{r-1}\geq \frac{\sqrt{k}}{\lambda}\left(\pm 2\sqrt{M\log(n)}+\gamma(|q_t|-|S^t_{p_t}|)\right).$$

We then can rigorously prove condition:

$$\cos(S_t,\theta)^{r-1}\geq \frac{\sqrt{k}}{\lambda}\text{p-log}(n).$$

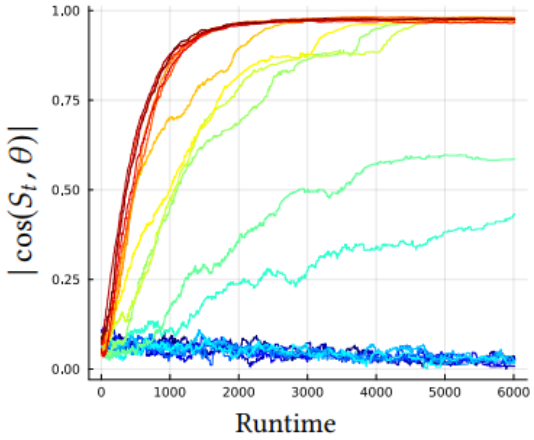


Figure 3: Randomized greedy for Sparse 3-Tensor PCA when $\theta \in \{-1, 0, 1\}^n$. Mean absolute angle vs time for $\lambda = n^\alpha$, initialized at a uniform random trinary vector. We predict that $\alpha = 1.4$ is the threshold for fast recovery. Here $n = 150$, $k = 56 \approx n^{0.8}$, and $\gamma = \log n$.

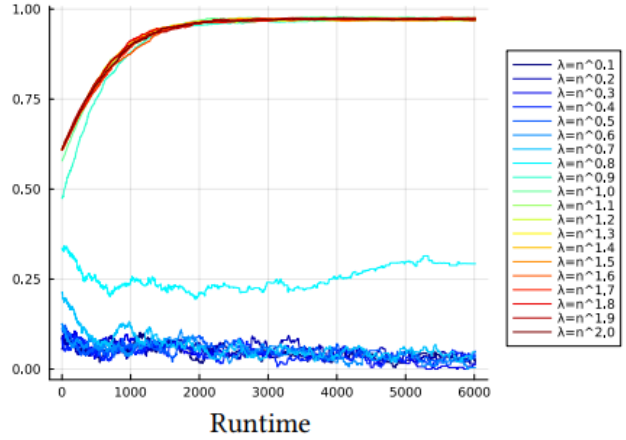


Figure 4: Randomized greedy for Sparse 3-Tensor PCA when $\theta \in \{-1, 0, 1\}^n$. Mean absolute angle vs time for $\lambda = n^\alpha$, initialized at S_{HOM}. We predict that $\alpha = 0.9$ is the threshold for fast recovery. Here $n = 150$, $k = 56 \approx n^{0.8}$, and $\gamma = \log n$.

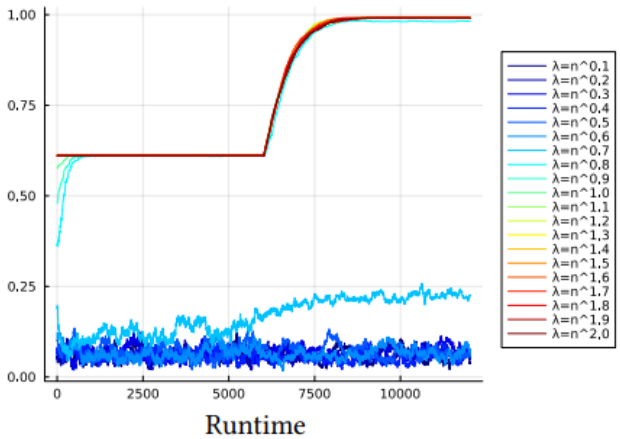


Figure 5: Two-stage algorithm for sparse 3-tensor PCA when $\theta \in \{-1, 0, 1\}^n$. Mean absolute angle vs. time for $\lambda = n^\alpha$, initialised at S_{HOM}. We predict that $\alpha = 0.75$ is the threshold for fast recovery. Here $n = 150$, $k = 56 \approx n^{0.8}$, and $\gamma = \log n$.

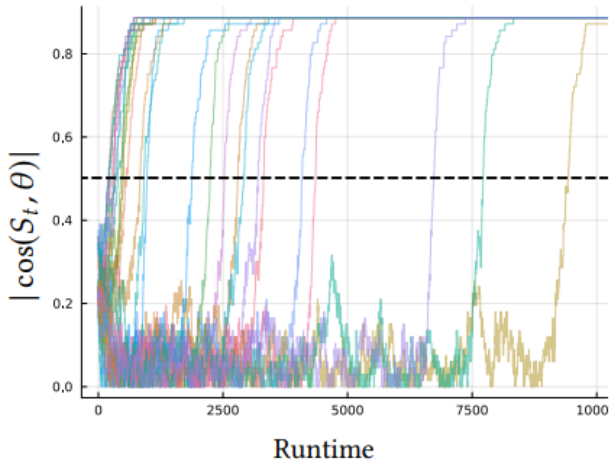


Figure 6: Stage one of the two-stage algorithm for sparse 3-tensor PCA when $\theta \in \{-1, 0, 1\}^n$. Absolute angle vs. time for $\lambda = \Theta(n^{3/4})$, initialised at S_{HOM}. Here $n = 150$, $k = 116 \approx n^{0.95}$, and we plot 29 simulations (filtered from 400 total simulations) which exhibit the most visible oscillatory phase. The dashed black line (where $|\cos(S_t, \theta)| \approx .502$) is a prediction for when this stage transitions from an oscillatory phase (below the line) to a monotonically increasing phase (above the line). Color is only used for visual clarity.

- Rigorous analysis of iterative algorithms.
- Variance inflation is mild under $M = \text{p-log}(n)$.
- Takeaway: **Subset Gaussian Cloning bridges algorithm design and clean probabilistic models**.

- Non-Gaussian
- More general algorithms
- A more convenient noising scheme