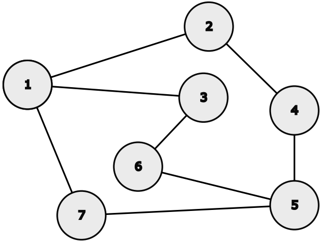
Benign nonconvexity in overparametrized group synchronization

Andrew D. McRae

Institute of Mathematics, EPFL
Joint work with Nicolas Boumal

ETHZ DACO seminar
October 17, 2023

Orthogonal group synchronization on graph



- Graph $G = (V, E)$ with vertices $V = \{1, \dots, n\}$
- Each node i has associated $r \times r$ orthogonal matrix Z_i ($Z_i Z_i^T = I_r$)
- Observed data: $R_{ij} \approx Z_i Z_j^T$ for $(i, j) \in E$
- Goal: estimate $Z_1, \dots, Z_n$
- Useful for SLAM (robotics), image alignment, and many other applications

First optimization approach

Setup:

- Graph $G = (V, E)$ with vertices $V = \{1, \dots, n\}$
- Want to estimate orthogonal matrices $Z_1, \dots, Z_n$
- Observed data: $R_{ij} \approx Z_i Z_j^T$ for $(i, j) \in E$

Least-squares optimization problem:

$$\min_{Y_i \in \mathbf{R}^{r \times r}} \sum_{(i,j) \in E} \|R_{ij} - Y_i Y_j^T\|_{\mathrm{F}}^2 \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \dots, n.$$

Equivalent problem (due to orthogonality):

$$\max_{Y_i \in \mathbf{R}^{r \times r}} \sum_{(i,j) \in E} \langle R_{ij}, Y_i Y_j^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \dots, n.$$

Nonconvex and in general has bad local optima.

Relaxed optimization problem

First optimization problem (can have **bad local minima**):

$$\max_{Y_i \in \mathbf{R}^{r\times r}} \underbrace{\sum_{(i,j)\in E} \langle R_{ij}, Y_i Y_j^T \rangle}_{=\langle C, YY^T\rangle \text{ if } Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}} \quad \text{s.t. } Y_i Y_i^T = I_r, i = 1, \dots, n.$$

C is a symmetric $rn \times rn$ matrix. YY^T is a positive semidefinite $rn \times rn$ matrix; the **semidefinite relaxation** (SDP) is then

$$\max_{X \in \mathbf{R}^{rn\times rn}} \langle C, X \rangle \text{ s.t. } X_{ii} = I_r, 1, \dots, n, X \succeq 0.$$

This is **convex** (no bad local minima) but **expensive** if n is large.

Question: Is there an approach that

- is computationally efficient **and**
- has no bad local optima?

Intermediate relaxation

SDP relaxation replaces YY^T by a general PSD matrix $X \in \mathbf{R}^{rn \times rn}$

- Original problem: for $Y \in \mathbf{R}^{rn \times r}$, YY^T has rank r
- SDP relaxation: matrix variable X can have rank as large as rn

Rank-p relaxation ($p \geq r$):

$$\max_{Y \in \mathbf{R}^{rn \times p}} \langle C, YY^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \ldots, n.$$

YY^T has rank at most p

- Bigger $p \implies$ more like (convex) SDP
- But number of optimization variables scales with p

Question: Does making p just slightly bigger than r help with nonconvexity?

- Empirically yes in robotics literature (e.g., Rosen et al. 2019; Dellaert et al. 2020)
- Can we explain/prove this?

Main result (noiseless)

Rank-p relaxation ($\boldsymbol{p}$ replaces $\boldsymbol{r}$):

$$\max_{Y \in \mathbf{R}^{rn \times \boldsymbol{p}}} \langle C, YY^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \ldots, n. \tag{1}$$

Theorem

Suppose

- *The measurement graph G is connected*
- *The measurements are **exact**: we observe $R_{ij} = Z_i Z_j^T$ for $(i, j) \in E$*
- $\boldsymbol{p \geq r + 2}$.

Then every second-order critical point Y of (1) can be written $Y = ZU$, where $UU^T = I_r$.

Consequence:

- Recovers ground truth up to global orthogonal transformation
- Local algorithms with "generic" (e.g., random) initialization reach SOCP $\Longrightarrow$ optimum

Oscillator interpretation

Kuramoto oscillator network on graph G

$$\dot{\theta}_i = \sum_{j \sim i} \sin(\theta_j - \theta_i), \quad i = 1, \ldots, n$$

- θ_i's are angles: **synchronization on unit circle** S^1
- Common research question: **for which G** does the network converge to "synchronized" state $\theta_1 = \cdots = \theta_n$ for almost every initial state?

"Oscillator" on Stiefel manifold

Our setup can be interpreted as **synchronization on the Stiefel manifold**

$$\mathrm{St}(r, p) := \{U \in \mathbf{R}^{r \times p} : UU^T = I_r\}$$

Our result: if $p \geq r + 2$, **every connected network** synchronizes on $\mathrm{St}(r, p)$

- Note d-sphere $S^d = \mathrm{St}(1, d+1)$

Proof sketch (noiseless)

Study **second-order critical points** of

$$\max_{Y \in \mathbf{R}^{rn \times p}} \langle C, YY^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \dots, n,$$

where

$$C_{ij} = \begin{cases} Z_i Z_j^T & (i, j) \in E \\ 0 & \text{otherwise.} \end{cases}$$

WLOG assume $Z_1 = \cdots = Z_n = I_r$ (same optimization landscape). Then

$$C = \underset{\substack{\uparrow \\ \text{Adjacency matrix of } G}}{A} \overset{\substack{\text{Kronecker prod.} \\ \downarrow}}{\otimes} I_r$$

Manifold optimization

Simplified: study **second-order critical points** of

$$\max_{Y \in \mathbf{R}^{rn \times p}} \langle A \otimes I_r, YY^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \dots, n.$$

Each Y_i lies on a Stiefel manifold:

$$Y_i \in \mathrm{St}(r, p) := \{U \in \mathbf{R}^{r \times p} : UU^T = I_r\}$$

Allowed perturbations (tangent vectors) at Y:

$$\dot{Y} = \begin{bmatrix} \dot{Y}_1 \\ \vdots \\ \dot{Y}_n \end{bmatrix} \text{ with } \dot{Y}_i Y_i^T + Y_i \dot{Y}_i^T = 0.$$

Criticality conditions

Optimization problem:

$$\max_{Y \in \mathbf{R}^{rn \times p}} \langle A \otimes I_r, YY^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \ldots, n. \tag{2}$$

Let L be the graph Laplacian matrix of G, i.e.,

$$L_{ij} = \begin{cases} \sum_k A_{ik} & i = j \\ -A_{ij} & \text{otherwise.} \end{cases}$$

Set

$$S(Y) = \overline{L} - \underset{\substack{\uparrow \\ \text{Sym. blk. diag.}}}{\text{SBD}}(\overline{L}YY^T), \text{ where } \overline{L} = L \otimes I_r.$$

Critical points of (2):

- **First-order:** $S(Y)Y = 0$ (Riemannian gradient is zero)
- **Second-order:** for any tangent vector $\dot{Y}$, $\langle S(Y), \dot{Y}\dot{Y}^T \rangle \geq 0$ (Riemannian Hessian is PSD)

Using (second-order) criticality

SOCP Y satisfies, for any tangent vector $\dot{Y} \in \mathbf{R}^{rn \times p}$,

$$\langle \bar{L} - \mathrm{SBD}(\bar{L}YY^T), \dot{Y}\dot{Y}^T \rangle \geq 0.$$

Analysis key: choose the right $\dot{Y}$

Intuition: move in the direction of ground truth

$$Z = \begin{bmatrix} I_r \\ \vdots \\ I_r \end{bmatrix} = \mathbf{1}_n \otimes I_r.$$

What if:

$$\dot{Y}\dot{Y}^T = ZZ^T = (\mathbf{1}_n \mathbf{1}_n^T) \otimes I_r$$

Then

$$\langle \bar{L}, ZZ^T \rangle \geq \langle \mathrm{SBD}(\bar{L}YY^T), ZZ^T \rangle = \mathrm{tr}(\bar{L}YY^T) = \langle \bar{L}, YY^T \rangle$$

Structure of $\overline{L}$

L graph Laplacian of G:

- L is PSD
- $L\mathbf{1}_n = 0$; smallest eigenvalue $\lambda_1(L) = 0$ with the constant eigenvector
- G connected $\implies$ every other eigenvalue **strictly positive**

What does this mean for $\overline{L} = L \otimes I_r$?

- $\overline{L}$ is PSD
- $\overline{L}Z = (L \otimes I_r)(\mathbf{1}_n \otimes I_r) = 0$; r zero eigenvalues corresponding to columns of Z
- All other eigenvalues strictly positive

So

$$\langle \overline{L}, YY^T \rangle \leq \langle \overline{L}, ZZ^T \rangle = 0 \implies \overline{L}Y = 0 \implies Y = ZU$$

for some $r \times p$ matrix U. $Y_1Y_1^T = I_r \implies UU^T = I_r$. **Done!**

How to make a valid argument?

We showed

$$\dot{Y}\dot{Y}^T = ZZ^T \implies \langle \bar{L}, YY^T \rangle \leq \langle \bar{L}, ZZ^T \rangle \implies \text{result}$$

This is potentially **illegal**: $\dot{Y}$ needs to be a tangent vector, i.e.,

$$\dot{Y}_i Y_i^T + Y_i \dot{Y}_i^T, i = 1, \dots, n$$

$\dot{Y}_i$ is valid if and only if it has the form

$$\dot{Y}_i = \Gamma_i - Y_i \Gamma_i^T Y_i$$

Intuition: $\dot{Y}\dot{Y}^T \approx ZZ^T = (\mathbf{1}_n \mathbf{1}_n^T) \otimes I_r \iff \dot{Y}_i \dot{Y}_j^T \approx I_r$ requires

$$\dot{Y}_1 \approx \cdots \approx \dot{Y}_n \implies \Gamma_1 = \cdots = \Gamma_n$$

Randomization

We want

$$\dot{Y}_i \dot{Y}_j^T \approx I_r, i, j = 1, \dots, n.$$

We will try

$$\dot{Y}_i = \Gamma - Y_i \Gamma^T Y_i, i = 1, \dots, n.$$

Then

$$(\dot{Y}\dot{Y}^T)_{ij} = \dot{Y}_i \dot{Y}_j^T = \Gamma\Gamma^T - (Y_i\Gamma^T)^2 - (\Gamma Y_j^T)^2 + Y_i\Gamma^T Y_i Y_j^T \Gamma Y_j^T$$

🙂 😬 😱

Simplified by choosing **random** Γ:

$$\Gamma_{k\ell} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1) \Longrightarrow \mathbf{E}\, \dot{Y}_i \dot{Y}_j^T = (p-2)I_r + \operatorname{tr}(Y_i Y_j^T) Y_i Y_j^T.$$

😁 🤔

Plugging in

We have random $\dot{Y}$ with $\mathbf{E}\,\dot{Y}_i\dot{Y}_j^T = (p-2)I_r + \mathrm{tr}(Y_iY_j^T)Y_iY_j^T$
Second-order criticality works under expectation:

$$\langle \bar{L} - \mathrm{SBD}(\bar{L}YY^T), \dot{Y}\dot{Y}^T \rangle \geq 0 \Longrightarrow \langle \bar{L} - \mathrm{SBD}(\bar{L}YY^T), \mathbf{E}\,\dot{Y}\dot{Y}^T \rangle \geq 0.$$

First, the easy part:

$$\mathbf{E}\,\dot{Y}_i\dot{Y}_i^T = (p+r-2)I_r \Longrightarrow \langle \mathrm{SBD}(\bar{L}YY^T), \mathbf{E}\,\dot{Y}\dot{Y}^T \rangle = (p+r-2)\langle \bar{L}, YY^T \rangle.$$

The harder part

$$
\begin{aligned}
\langle \overline{L}, \mathbf{E}\, \dot{Y}\dot{Y}^T \rangle &= \langle L \otimes I_r, \mathbf{E}\, \dot{Y}\dot{Y}^T \rangle \\
&= \sum_{i,j} L_{ij} \operatorname{tr}(\mathbf{E}\, \dot{Y}_i \dot{Y}_j^T) \\
&= \sum_{i,j} L_{ij} \operatorname{tr}((p-2)I_r + \operatorname{tr}(Y_i Y_j^T) Y_i Y_j^T) \\
&= (p-2)r \underbrace{\sum_{i,j} L_{ij}}_{=0} + \sum_{i,j} L_{ij} \operatorname{tr}^2(Y_i Y_j^T) \\
&= \sum_{i,j} L_{ij} \operatorname{tr}^2(Y_i Y_j^T).
\end{aligned}
$$

A helpful trick

We need to calculate

$$\langle \overline{L}, \mathbf{E}\, \dot{Y}\dot{Y}^T \rangle = \sum_{i,j} L_{ij}\, \mathrm{tr}^2(Y_i Y_j^T)$$

Because $Y_i Y_i^T = I_r$,

$$\begin{aligned} \mathrm{tr}(Y_i Y_j^T) &= \frac{1}{2}\, \mathrm{tr}(Y_i Y_j^T + Y_j Y_i^T) \\ &= \mathrm{tr}\left(I_r - \frac{1}{2}(Y_i - Y_j)(Y_i - Y_j^T) \right) \\ &= \left(r - \frac{1}{2}\|Y_i - Y_j\|_F^2 \right). \end{aligned}$$

Then, after more tedious computations,

$$\mathrm{tr}^2(Y_i Y_j^T) = -r^2 + 2r\, \mathrm{tr}(Y_i Y_j^T) + \frac{1}{4}\|Y_i - Y_j\|_F^4.$$

Finishing

We have shown

$$\langle \bar{L}, \mathbf{E}\,\dot{Y}\dot{Y}^T \rangle = \sum_{i,j} L_{ij}\left(-r^2 + 2r\,\mathrm{tr}(Y_i Y_j^T) + \frac{1}{4}\|Y_i - Y_j\|_F^4 \right)$$
$$= 2r\langle \bar{L}, YY^T \rangle - \frac{1}{4}\sum_{i,j} A_{ij}\|Y_i - Y_j\|_F^4.$$

Then

$$(p+r-2)\langle \bar{L}, YY^T \rangle = \langle \mathrm{SBD}\,\bar{L}YY^T, \mathbf{E}\,\dot{Y}\dot{Y}^T \rangle \leq \langle \bar{L}, \mathbf{E}\,\dot{Y}\dot{Y}^T \rangle$$
$$\Longrightarrow \underbrace{(p-r-2)\langle \bar{L}, YY^T \rangle}_{\geq 0 \text{ if } \boldsymbol{p \geq r+2}} + \underbrace{\frac{1}{4}\sum_{i,j} A_{ij}\|Y_i - Y_j\|_F^4}_{>0 \text{ if } Y_i\text{'s not equal}} \leq 0.$$

So we must have $Y_1 = \cdots = Y_n$, which completes the proof!

How to handle noise?

Now

$$\underbrace{(p-r-2)\langle \overline{L}, YY^T\rangle}_{\text{quadratic}} + \underbrace{\frac{1}{4}\sum_{i,j} A_{ij}\|Y_i - Y_j\|_F^4}_{\text{quartic}} \leq \text{noise terms}$$

For simplicity, assume $p > r + 2$ and use only quadratic term.

$$\langle \overline{L}, YY^T\rangle \geq \lambda_2(L)\underbrace{\left(1 - \frac{\|Z^T Y\|_F^2}{n^2 r}\right)}_{\text{error (normalized)}} nr$$

- Robustness to noise depends on graph G through $\lambda_2(L)$
- Give error bound but not **landscape** result...

Landscape analysis with noise

First-order critical point condition:

$$S(Y)Y = 0, \text{ where } S(Y) = \bar{L} + \Delta - \mathrm{SBD}((\bar{L} + \Delta)YY^T)$$

Note

$$\Delta \text{ small} \underset{\uparrow \atop \text{error bound}}{\Longrightarrow} Y \approx ZU \underset{\uparrow \atop \bar{L}Z=0}{\Longrightarrow} \bar{L}Y \approx 0$$

Then

$$\left.\begin{array}{r} S(Y) \approx \bar{L} \\ S(Y)Y = 0 \end{array}\right\} \Longrightarrow \mathrm{rank}(Y) = r < p$$

- Y second-order critical **and** rank-deficient $\Longrightarrow YY^T$ solves **SDP relaxation**
- $\Longrightarrow Y$ is globally optimal for original problem!

Noisy landscape result

Theorem

Suppose

- *Connected measurement graph* G *on* $1, \dots, n$ *with edges* E.
- *We observe* $R_{ij} = Z_i Z_j^T + \Delta_{ij} \in \mathbf{R}^{r \times r}, (i, j) \in E$.
- $p > r + 2$, *and we solve the rank-relaxed problem*

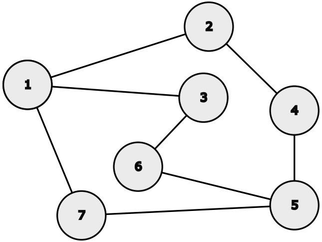
$$\max_{Y \in \mathbf{R}^{rn \times p}} \langle C, YY^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, i = 1, \dots, n,$$

where $C_{ij} = R_{ij}$ *for* $(i, j) \in E$.

Then, if $\|\boldsymbol{\Delta}\|_{\ell_2} \leq \boldsymbol{C_{p,r}} \frac{\boldsymbol{\lambda_2(L)}}{\sqrt{\boldsymbol{n}}}$, *any second-order critical point* Y

- *is a global optimum and*
- *has rank* r *(i.e., we lost nothing from relaxation).*

Recap of results



- Relative $r \times r$ orthogonal group measurements $R_{ij} \approx Z_i Z_j^T$ for (i, j) edges of a connected graph
- Rank-relaxed estimator of Z:

$$\max_{Y \in \mathbf{R}^{rn \times p}} \langle C, YY^T \rangle \text{ s.t. } Y_i Y_i^T = I_r, \quad i = 1, \ldots, n. \tag{3}$$

- If $p > r + 2$ and noise is small enough vs. graph connectivity:
  - (3) has **benign landscape**
  - Rank relaxation is **tight** (optima have rank r)

Literature comparison

Landscape of rank-relaxed group synchronization

- Prior theoretical work (e.g., Ling 2022) focused on complete-graph case
- We extend to general graphs (losing some noise tolerance)

"Oscillator" networks over Stiefel manifold $\mathrm{St}(r, p)$

- Previous best result (e.g., Markdahl, Thunberg, and Goncalves 2020) requires $\boldsymbol{2p \geq 3(r+1)}$ (ours is $\boldsymbol{p \geq r+2}$)
- Our result best possible by topological argument (Markdahl 2021)
- Prior work conjectured our result from topology and empirical evidence

Our proof technique is similar to both but uses a better (randomized) perturbation

Future directions

- Improve size dependence in noisy results (currently requires SNR $= \frac{\lambda_2}{\|\Delta\|_{\ell_2}} \gtrsim \sqrt{n}$)
- Complex case (possibly-suboptimal results in our preprint)
- Oscillator synchronization on other manifolds
- Other problems with optimization over low-rank matrices
  - stochastic block model
  - low-rank matrix sensing
  - sensor network localization/MDS

Preprint: Andrew D. McRae and Nicolas Boumal (2023). "Benign landscapes of low-dimensional relaxations for orthogonal synchronization on general graphs". In: arXiv: 2307.02941 [math.OC]

References I

Dellaert, Frank et al. (Aug. 2020). "Shonan Rotation Averaging: Global Optimality by Surfing $SO(p)^n$". In: *Proc. European Conf. Comput. Vis. (ECCV)*. Virtual conference, pp. 292–308.

Ling, Shuyang (2022). "Solving Orthogonal Group Synchronization via Convex and Low-Rank Optimization: Tightness and Landscape Analysis". In: *Math. Program.* Forthcoming.

Markdahl, Johan (2021). "Synchronization on Riemannian Manifolds: Multiply Connected Implies Multistable". In: *IEEE Trans. Autom. Control* 66.9, pp. 4311–4318.

Markdahl, Johan, Johan Thunberg, and Jorge Goncalves (2020). "High-dimensional Kuramoto models on Stiefel manifolds synchronize complex networks almost globally". In: *Automatica* 113.

McRae, Andrew D. and Nicolas Boumal (2023). "Benign landscapes of low-dimensional relaxations for orthogonal synchronization on general graphs". In: arXiv: `2307.02941 [math.OC]`.

References II

Rosen, David M. et al. (2019). "SE-Sync: A certifiably correct algorithm for synchronization over the special Euclidean group". In: *Int. J. Robot. Res.* 38.2-3, pp. 95–125.