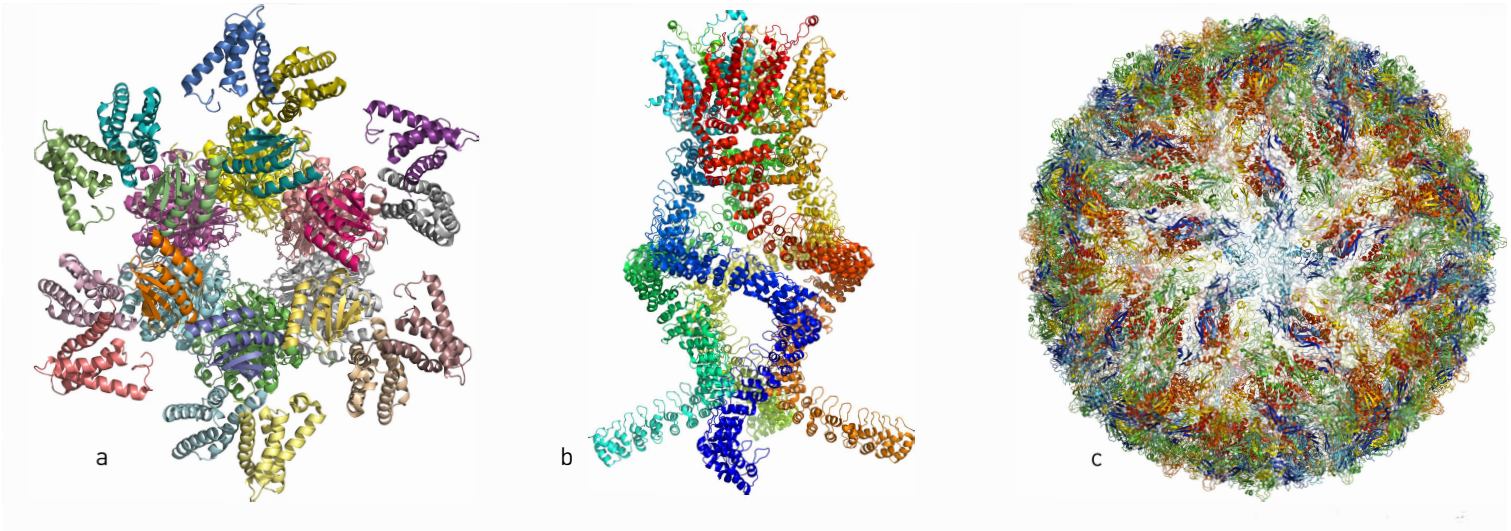
The sample complexity of multi-reference alignment (and a few words about cryo-EM)

Tamir Bendory (Tel Aviv University, EE)

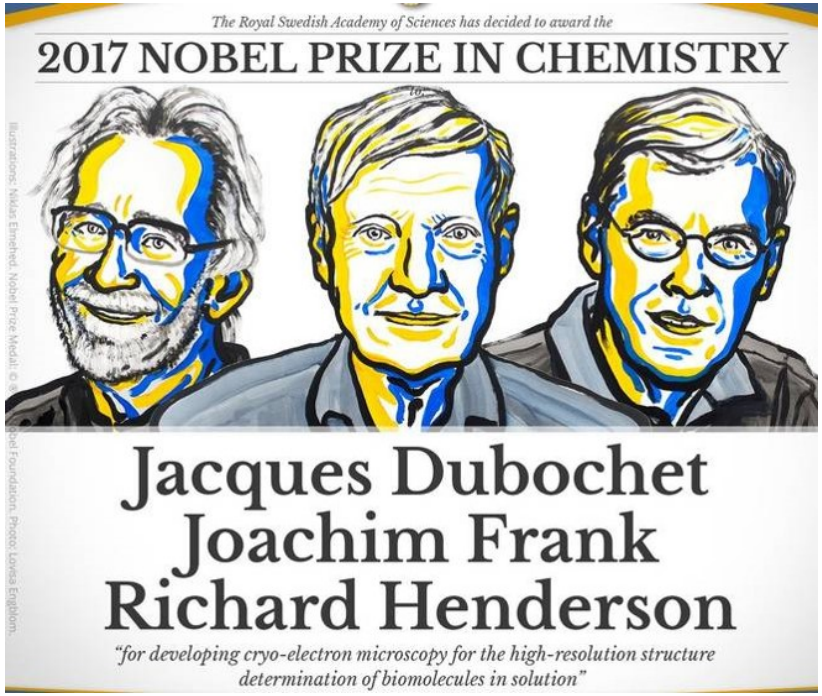
Structural biology

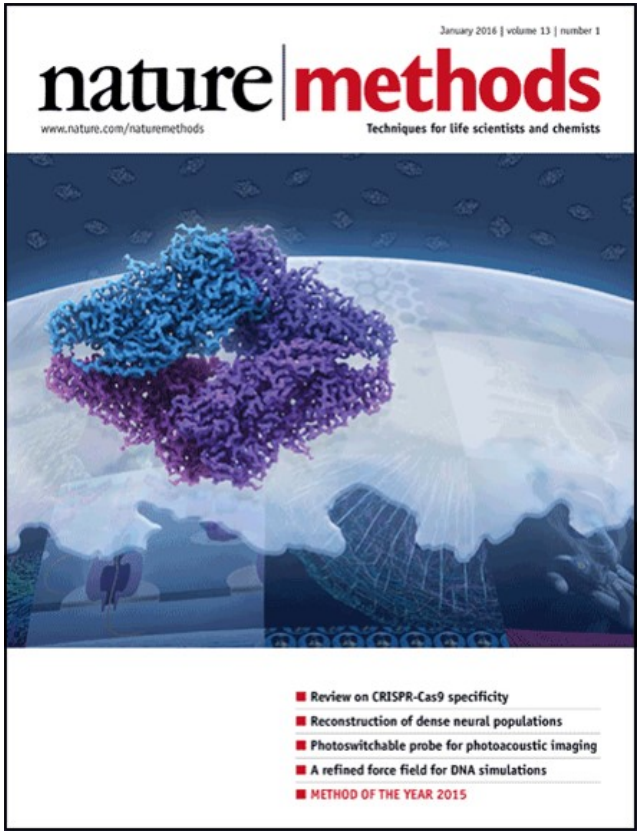
Structural biology is the study of the molecular structure and dynamics of biological macromolecules, particularly proteins.



(left) A protein complex that governs the circadian rhythm. (middle) A sensor of the type that reads pressure changes in the ear and allows us to hear. (right) The Zika virus.

Exciting times for cryo-electron microscopy (cryo-EM)



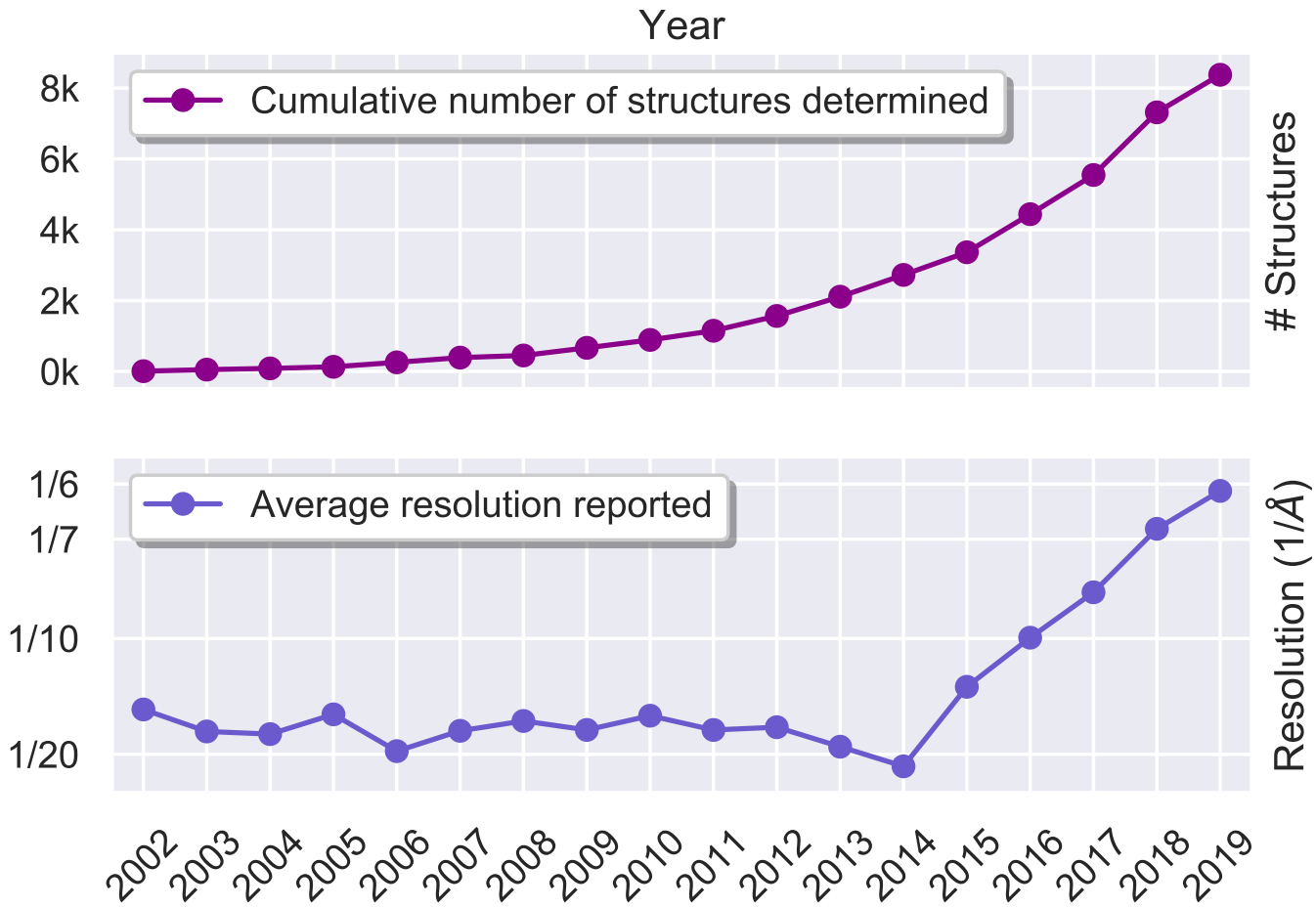


Why cryo-EM?

- Does not require crystallization and thus can capture molecules in their native states

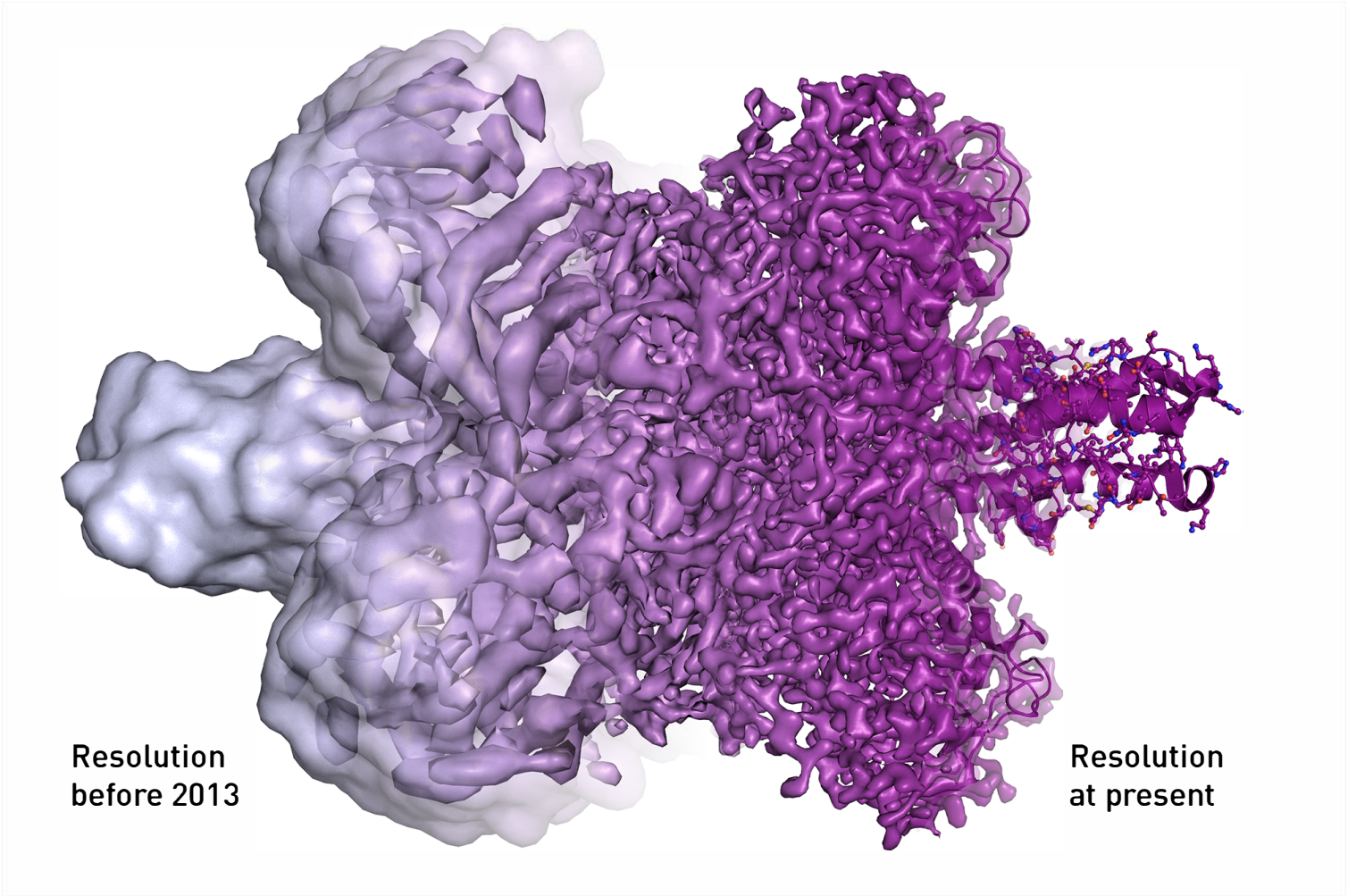
- Has the potential to analyze conformationally heterogeneous mixtures and, consequently, can be used to determine the structures of complexes in different functional states

The recent growth in the number of high-resolution structures produced by cryo-EM



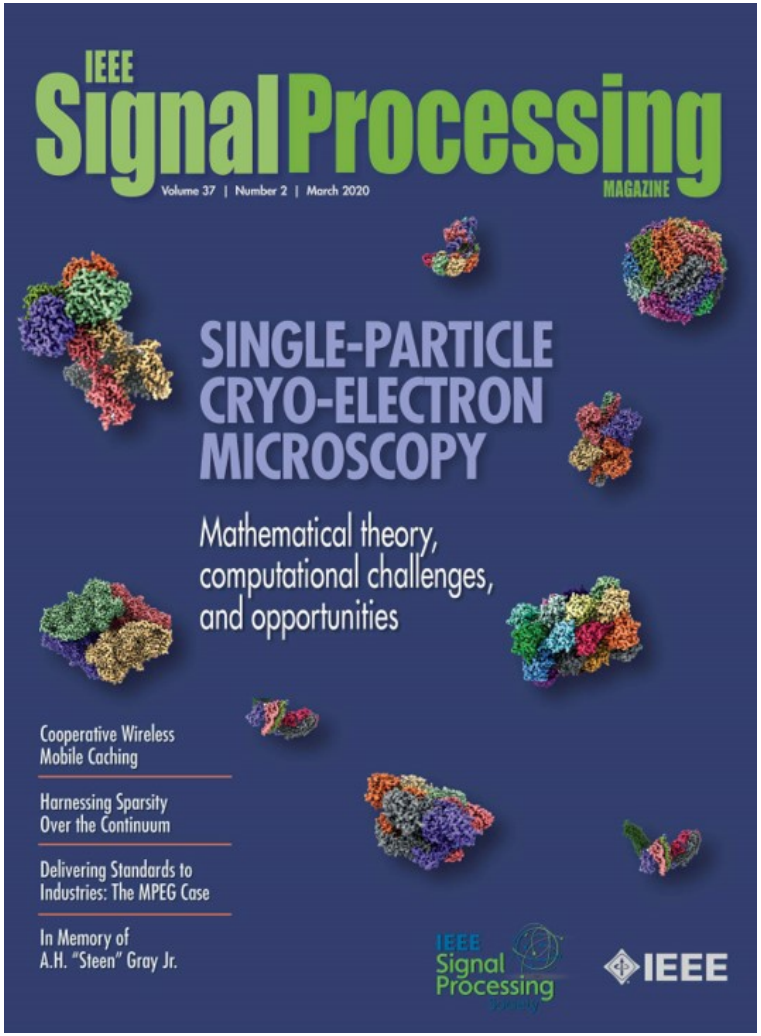
Taken from the Electron Microscopy Data Bank public repository.

The resolution revolution



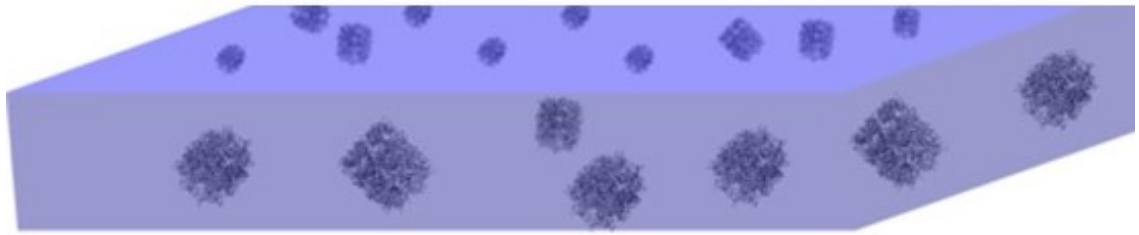
https://www.nobelprize.org/prizes/chemistry/2017/press-release/

Recent survey

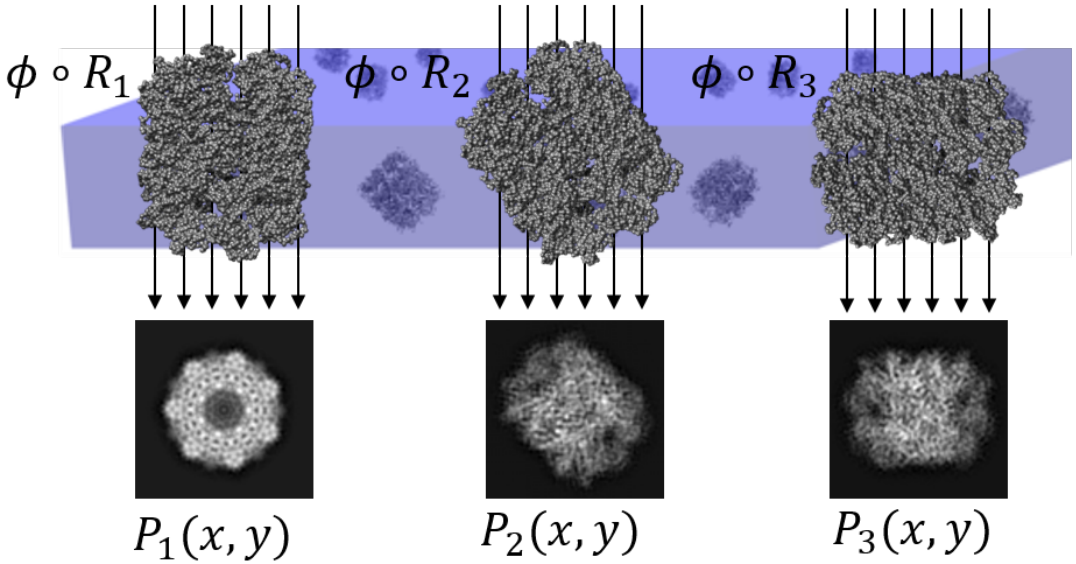


Bendory, Bartesaghi, and Singer. "Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities." IEEE signal processing magazine 37.2 (2020): 58-76.

Mathematical model of cryo-EM



Mathematical model of cryo-EM



$$P_i = \text{projection}(\text{rotation}(\phi)) + \text{noise}$$

The cryo-EM problem: Estimate 3-D structure ϕ from $P_1, \ldots, P_n$, while the 3-D rotations are unknown and the SNR is low (say, $1/100$).

Multi-reference alignment

Let $\mathbb{X}$ be a vector space and G be a group acting on $\mathbb{X}$. Suppose we have n measurements of the form

$$y_i = T(g_i \circ x) + \varepsilon_i, \quad i = 1, \dots, n,$$

where

- x is an unknown element of $\mathbb{X}$;
- $g_1, \dots, g_n$ are unknown elements of G;
- $\circ$ is the action of G on $\mathbb{X}$;
- $T : \mathbb{X} \to \mathbb{Y}$ is a linear operator;
- $\mathbb{Y}$ is the (finite-dimensional) measurement space;
- $\varepsilon_i's$ are independent noise terms.

Multi-reference alignment

Let $\mathbb{X}$ be a vector space and G be a group acting on $\mathbb{X}$. Suppose we have n measurements of the form

$$y_i = T(g_i \circ x) + \varepsilon_i, \quad i = 1, \dots, n,$$

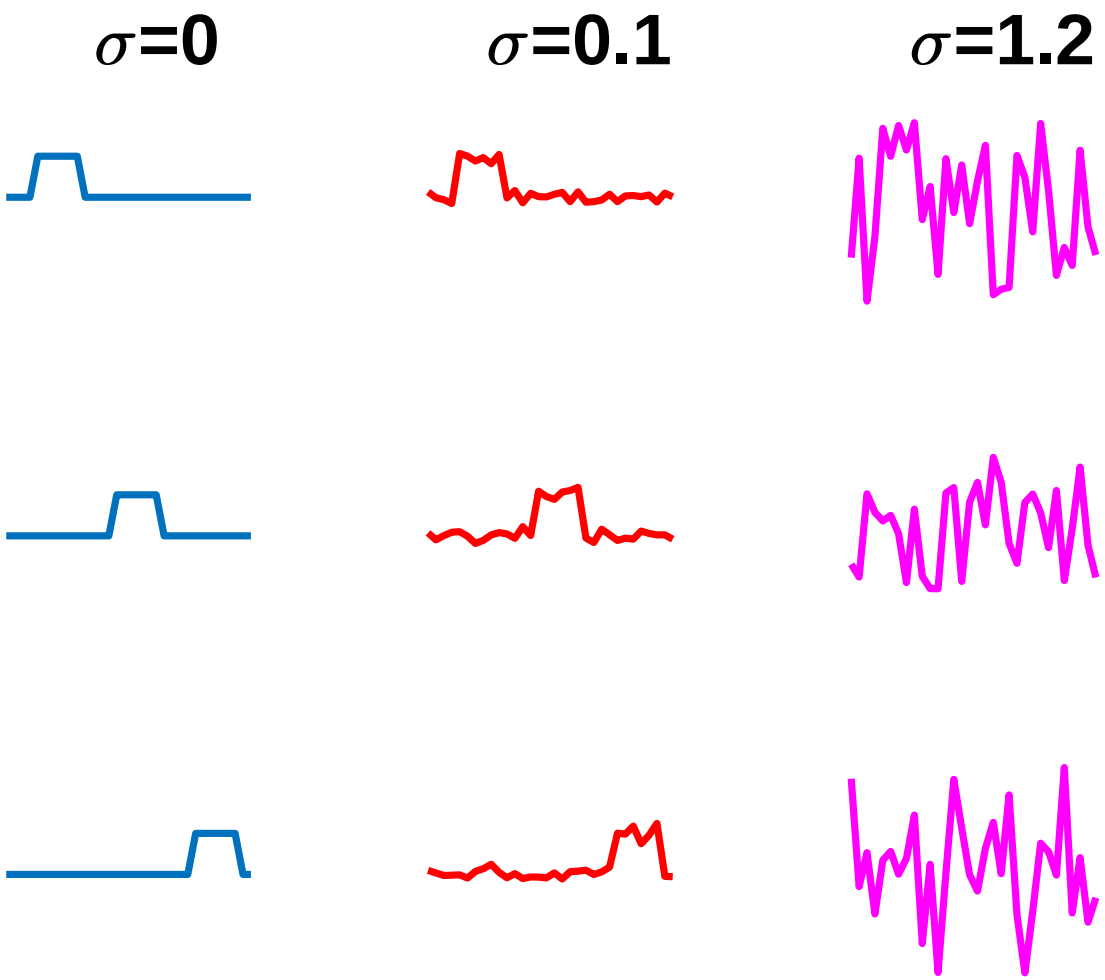
where

- x is an unknown element of $\mathbb{X}$;
- $g_1, \dots, g_n$ are unknown elements of G;
- $\circ$ is the action of G on $\mathbb{X}$;
- $T : \mathbb{X} \to \mathbb{Y}$ is a linear operator;
- $\mathbb{Y}$ is the (finite-dimensional) measurement space;
- $\varepsilon_i's$ are independent noise terms.

Our goal it to estimate the orbit

$$Gx = \{g \circ x | g \in G\}.$$

Example: 1-D discrete MRA



Estimation in the high SNR regime

- Recall that we wish to estimate the orbit of $x \in \mathbb{X}$ from

$$y_i = T(g_i \circ x) + \varepsilon_i, \quad g_i \in G,\ i = 1, \ldots, n.$$

Estimation in the high SNR regime

- Recall that we wish to estimate the orbit of $x \in \mathbb{X}$ from

$$y_i = T(g_i \circ x) + \varepsilon_i, \quad g_i \in G,\ i = 1, \ldots, n.$$

- If the g_i were known, then the task of recovering x would reduce to a classical linear inverse problem, for which many effective techniques exist.

Estimation in the high SNR regime

- Recall that we wish to estimate the orbit of $x \in \mathbb{X}$ from

$$y_i = T(g_i \circ x) + \varepsilon_i, \quad g_i \in G,\ i = 1, \ldots, n.$$

- If the g_i were known, then the task of recovering x would reduce to a classical linear inverse problem, for which many effective techniques exist.

- Therefore, the problem reduces to estimating the group elements $g_1, \ldots, g_n$ from the observations $y_1, \ldots, y_n$.

Estimation in the high SNR regime

- Recall that we wish to estimate the orbit of $x \in \mathbb{X}$ from

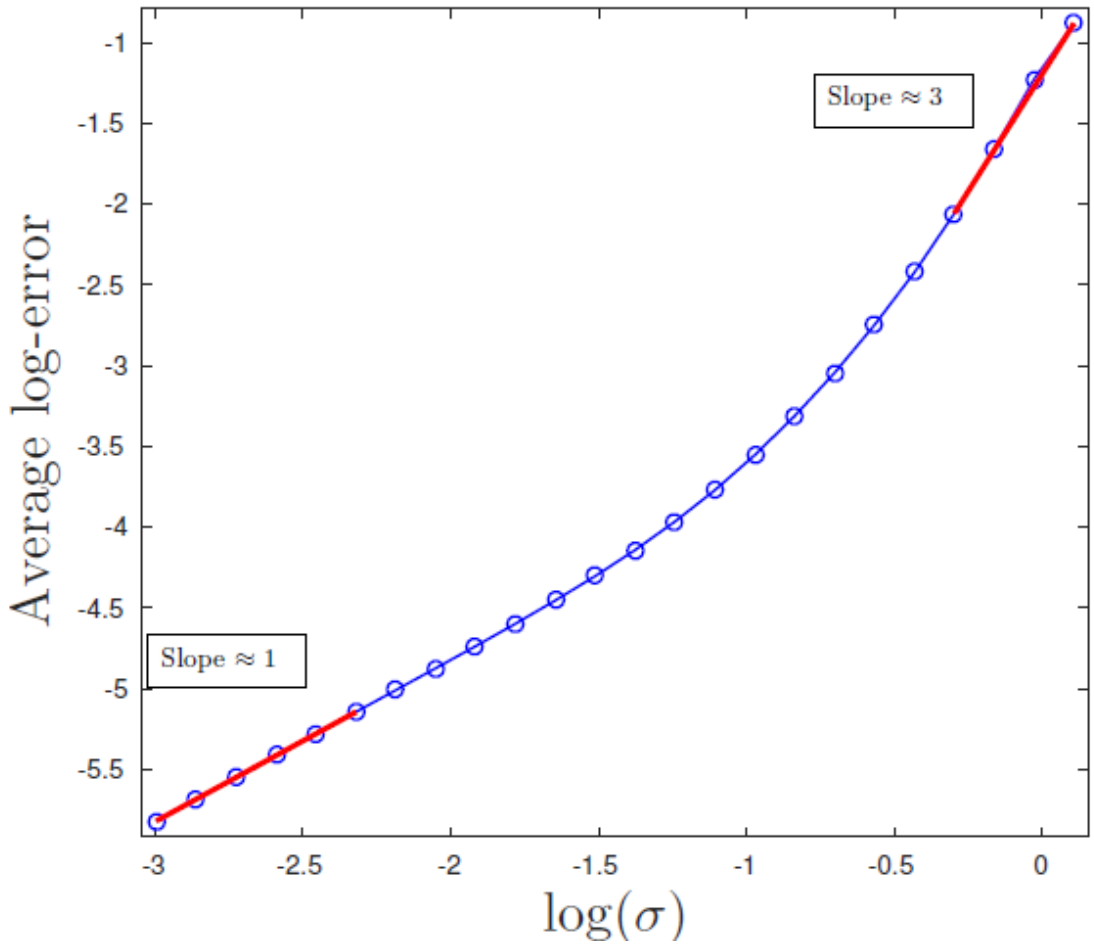
$$y_i = T(g_i \circ x) + \varepsilon_i, \quad g_i \in G,\ i = 1, \ldots, n.$$

- If the g_i were known, then the task of recovering x would reduce to a classical linear inverse problem, for which many effective techniques exist.

- Therefore, the problem reduces to estimating the group elements $g_1, \ldots, g_n$ from the observations $y_1, \ldots, y_n$.

- The leading methodology to estimate the group elements is called group synchronization, see for example [Singer, '11], [Boumal, '16], [Bandeira et al., '17].

High vs. low SNR



Abbe, Bendory, Leeb, Pereira, Sharon, and Singer. "Multireference alignment is easier with an aperiodic translation distribution." IEEE Transactions on Information Theory 65, no. 6 (2018): 3565-3584.

Sample complexity in the low SNR regime

- In the low SNR regime $n, \sigma \to \infty$ (fixed dimension L), estimating the group elements accurately is challenging.

Sample complexity in the low SNR regime

- In the low SNR regime $n, \sigma \to \infty$ (fixed dimension L), estimating the group elements accurately is challenging.

- Remarkably, is was shown that one can estimate the signal even if the group elements cannot be estimated.

Sample complexity in the low SNR regime

- In the low SNR regime $n, \sigma \to \infty$ (fixed dimension L), estimating the group elements accurately is challenging.

- Remarkably, is was shown that one can estimate the signal even if the group elements cannot be estimated.

- In particular, it was shown that if $\bar{d}$ is the lowest degree moment that determines an orbit uniquely, then $n = \omega(\sigma^{2\bar{d}})$ is a necessary condition for accurate recovery [Abbe et al., '18], [Perry et al., '19].

Sample complexity in the low SNR regime

- In the low SNR regime $n, \sigma \to \infty$ (fixed dimension L), estimating the group elements accurately is challenging.

- Remarkably, is was shown that one can estimate the signal even if the group elements cannot be estimated.

- In particular, it was shown that if $\bar{d}$ is the lowest degree moment that determines an orbit uniquely, then $n = \omega(\sigma^{2\bar{d}})$ is a necessary condition for accurate recovery [Abbe et al., '18], [Perry et al., '19].

- Therefore, the question of sample complexity boils down to identifying $\bar{d}$ for a given MRA setup; it may depend on the vector space $\mathbb{X}$, the group G, the linear operator T, and the distribution of group elements.

Some results

- 1-D discrete MRA

Some results

- 1-D discrete MRA

  - if the group elements are drawn from a uniform distribution, then $\bar{d} = 3$. Thus, $n = \omega(\sigma^6)$. [Bendory et al, '17], [Perry et al., '19]

Some results

- 1-D discrete MRA
  - if the group elements are drawn from a uniform distribution, then $\bar{d} = 3$. Thus, $n = \omega(\sigma^6)$. [Bendory et al, '17], [Perry et al., '19]
  - if the group elements are drawn from almost any non-uniform distribution, then $\bar{d} = 2$. Thus, $n = \omega(\sigma^4)$. [Abbe et al., '17]

Some results

- 1-D discrete MRA
  - if the group elements are drawn from a uniform distribution, then $\bar{d} = 3$. Thus, $n = \omega(\sigma^6)$. [Bendory et al, '17], [Perry et al., '19]
  - if the group elements are drawn from almost any non-uniform distribution, then $\bar{d} = 2$. Thus, $n = \omega(\sigma^4)$. [Abbe et al., '17]
- For cryo-EM with a uniform distribution over $SO(3)$ (under some simplifying assumptions), $\bar{d} = 3$. Thus, $n = \omega(\sigma^6)$. [Bandeira et al., '17]

More examples (partial list)

- MRA in 2-D [Ma et al., '19], [Janco and Bendory, '21]
- MRA with projection [Bandeira et al., '17]
- Heterogeneous MRA [Bandeira et al., '17; Boumal et al., '18]
- unprojected cryo-EM [Fan et al, '21; Liu and Moitra, '21]
- dihedral MRA [Bendory et al., '21]
- MRA with dilations [Hirn and Little, '19]
- MRA with the rigid motion group [Bendory et al., '21]
- sparse MRA [Ghosh, Rigollet, '21; Bendory et al. '21]
- learning a rigid body [Bandeira et al., '17; Pumir et al., '21]
- low-rank covariance estimation under unknown translations [Landa and Shkolnisky '21]

Computational consequences

- The method of moments achieves the optimal estimation rate (we've implemented the method for cryo-EM experimental datasets [Levin et al., '18], [Bendory et al., '18], [Sharon et al., '20]).

Computational consequences

- The method of moments achieves the optimal estimation rate (we've implemented the method for cryo-EM experimental datasets [Levin et al., '18], [Bendory et al., '18], [Sharon et al., '20]).
- In practice, expectation-maximization usually outperforms the method of moments.

Computational consequences

- The method of moments achieves the optimal estimation rate (we've implemented the method for cryo-EM experimental datasets [Levin et al., '18], [Bendory et al., '18], [Sharon et al., '20]).
- In practice, expectation-maximization usually outperforms the method of moments.
- In the low SNR regime, matching all the moments is equivalent to maximizing the likelihood function [Katsevich and Bandeira, '21].

Computational consequences

- The method of moments achieves the optimal estimation rate (we've implemented the method for cryo-EM experimental datasets [Levin et al., '18], [Bendory et al., '18], [Sharon et al., '20]).
- In practice, expectation-maximization usually outperforms the method of moments.
- In the low SNR regime, matching all the moments is equivalent to maximizing the likelihood function [Katsevich and Bandeira, '21].
- In some MRA models, we conjecture the existence of computation-statistical gaps: these are regimes in which the underlying statistical problem is information-theoretically possible although no efficient algorithm exists [Bandeira et al., '17],[Boumal et al., '18],[Bendory et al., '20], [Bendory et al., '21].

Sample complexity in high dimensions

- For 1-D MRA when $n, L, \sigma \to \infty$ (with a Gaussian prior), the sample complexity is not determined by the moments but by the ratio

$$\alpha = L/(\sigma^2 \log L).$$

Romanov, Bendory, and Ordentlich. "Multi-reference alignment in high dimensions: sample complexity and phase transition." SIAM Journal on Mathematics of Data Science 3.2 (2021): 494-523.

Sample complexity in high dimensions

- For 1-D MRA when $n, L, \sigma \to \infty$ (with a Gaussian prior), the sample complexity is not determined by the moments but by the ratio

$$\alpha = L/(\sigma^2 \log L).$$

- When $\alpha > 2$ the impact of the unknown circular shifts on the sample complexity is minor, and the problem is almost as easy as estimating a signal in additive white Gaussian noise.

Romanov, Bendory, and Ordentlich. "Multi-reference alignment in high dimensions: sample complexity and phase transition." SIAM Journal on Mathematics of Data Science 3.2 (2021): 494-523.

Sample complexity in high dimensions

- For 1-D MRA when $n, L, \sigma \to \infty$ (with a Gaussian prior), the sample complexity is not determined by the moments but by the ratio

$$\alpha = L/(\sigma^2 \log L).$$

- When $\alpha > 2$ the impact of the unknown circular shifts on the sample complexity is minor, and the problem is almost as easy as estimating a signal in additive white Gaussian noise.

- In sharp contrast, when $\alpha \leq 2$, the problem is significantly harder and the sample complexity grows substantially quicker with σ^2.

Romanov, Bendory, and Ordentlich. "Multi-reference alignment in high dimensions: sample complexity and phase transition." SIAM Journal on Mathematics of Data Science 3.2 (2021): 494-523.

Take-home message

- Many exciting open computational challenges in MRA (in information theory, machine learning, signal processing, statistics, algebra, etc.)

Take-home message

- Many exciting open computational challenges in MRA (in information theory, machine learning, signal processing, statistics, algebra, etc.)

- Theoretical and algorithmic results in MRA may have consequences for cryo-EM:

Take-home message

- Many exciting open computational challenges in MRA (in information theory, machine learning, signal processing, statistics, algebra, etc.)

- Theoretical and algorithmic results in MRA may have consequences for cryo-EM:

  - Reconstructing small molecular structures [Bendory et al., '18]

Take-home message

- Many exciting open computational challenges in MRA (in information theory, machine learning, signal processing, statistics, algebra, etc.)

- Theoretical and algorithmic results in MRA may have consequences for cryo-EM:

  - Reconstructing small molecular structures [Bendory et al., '18]

  - Reconstructing with fewer observations (in progress)

Take-home message

- Many exciting open computational challenges in MRA (in information theory, machine learning, signal processing, statistics, algebra, etc.)

- Theoretical and algorithmic results in MRA may have consequences for cryo-EM:

  - Reconstructing small molecular structures [Bendory et al., '18]

  - Reconstructing with fewer observations (in progress)

- Cryo-EM is an alluring example of a challenging data science problem, whose solution will have an immediate impact on all humankind.

References I

Emmanuel Abbe, Tamir Bendory, William Leeb, João M Pereira, Nir Sharon, and Amit Singer.

Multireference alignment is easier with an aperiodic translation distribution.

IEEE Transactions on Information Theory, 65(6):3565–3584, 2018.

Emmanuel Abbe, João M Pereira, and Amit Singer.

Estimation in the group action channel.

In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 561–565. IEEE, 2018.

Afonso S Bandeira, Ben Blum-Smith, Joe Kileel, Amelia Perry, Jonathan Weed, and Alexander S Wein.

Estimation under group actions: recovering orbits from invariants.

arXiv preprint arXiv:1712.10163, 2017.

Afonso S Bandeira, Yutong Chen, Roy R Lederman, and Amit Singer.

Non-unique games over compact groups and orientation estimation in cryo-EM.

Inverse Problems, 36(6):064002, 2020.

Tamir Bendory, Alberto Bartesaghi, and Amit Singer.

Single-particle cryo-electron microscopy: Mathematical theory, computational challenges, and opportunities.

IEEE signal processing magazine, 37(2):58–76, 2020.

Tamir Bendory, Nicolas Boumal, William Leeb, Eitan Levin, and Amit Singer.

Toward single particle reconstruction without particle picking: Breaking the detection limit.

arXiv preprint arXiv:1810.00226, 2018.

Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer.

Bispectrum inversion with application to multireference alignment.

IEEE Transactions on signal processing, 66(4):1037–1050, 2017.

References II

Tamir Bendory and Dan Edidin.

Toward a mathematical theory of the crystallographic phase retrieval problem.
SIAM Journal on Mathematics of Data Science, 2(3):809–839, 2020.

Tamir Bendory, Dan Edidin, William Leeb, and Nir Sharon.

Dihedral multi-reference alignment.
arXiv preprint arXiv:2107.05262, 2021.

Tamir Bendory, Ariel Jaffe, William Leeb, Nir Sharon, and Amit Singer.

Super-resolution multi-reference alignment.
arXiv preprint arXiv:2006.15354, 2020.

Tamir Bendory, Oscar Mickelin, and Amit Singer.

Sparse multi-reference alignment: sample complexity and computational hardness.
arXiv preprint arXiv:2109.11656, 2021.

Nicolas Boumal.

Nonconvex phase synchronization.
SIAM Journal on Optimization, 26(4):2355–2377, 2016.

Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer.

Heterogeneous multireference alignment: A single pass approach.
In *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2018.

Zhou Fan, Roy R Lederman, Yi Sun, Tianhao Wang, and Sheng Xu.

Maximum likelihood for high-noise group orbit estimation and single-particle cryo-EM.
arXiv preprint arXiv:2107.01305, 2021.

References III

Subhro Ghosh and Philippe Rigollet.

Multi-reference alignment for sparse signals, uniform uncertainty principles and the beltway problem.
arXiv preprint arXiv:2106.12996, 2021.

Matthew Hirn and Anna Little.

Wavelet invariants for statistically robust multi-reference alignment.
arXiv preprint arXiv:1909.11062, 2019.

Noam Janco and Tamir Bendory.

An accelerated expectation-maximization for multi-reference alignment.
arXiv preprint arXiv:2105.07372, 2021.

Anya Katsevich and Afonso Bandeira.

Likelihood maximization and moment matching in low SNR gaussian mixture models.
arXiv preprint arXiv:2006.15202, 2020.

Boris Landa and Yoel Shkolnisky.

Multi-reference factor analysis: low-rank covariance estimation under unknown translations.
Information and Inference: A Journal of the IMA, 10(3):773–812, 2021.

Eitan Levin, Tamir Bendory, Nicolas Boumal, Joe Kileel, and Amit Singer.

3d ab initio modeling in cryo-EM by autocorrelation analysis.
In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 1569–1573. IEEE, 2018.

Allen Liu and Ankur Moitra.

How to decompose a tensor with group structure.
arXiv preprint arXiv:2106.02680, 2021.

References IV

Chao Ma, Tamir Bendory, Nicolas Boumal, Fred Sigworth, and Amit Singer.

Heterogeneous multireference alignment for images with application to 2D classification in single particle reconstruction.
IEEE Transactions on Image Processing, 29:1699–1710, 2019.

Amelia Perry, Jonathan Weed, Afonso S Bandeira, Philippe Rigollet, and Amit Singer.

The sample complexity of multireference alignment.
SIAM Journal on Mathematics of Data Science, 1(3):497–517, 2019.

Thomas Pumir, Amit Singer, and Nicolas Boumal.

The generalized orthogonal Procrustes problem in the high noise regime.
Information and Inference: A Journal of the IMA, 10(3):921–954, 2021.

Elad Romanov, Tamir Bendory, and Or Ordentlich.

Multi-reference alignment in high dimensions: sample complexity and phase transition.
SIAM Journal on Mathematics of Data Science, 3(2):494–523, 2021.

Nir Sharon, Joe Kileel, Yuehaw Khoo, Boris Landa, and Amit Singer.

Method of moments for 3D single particle ab initio modeling with non-uniform distribution of viewing angles.
Inverse Problems, 36(4):044003, 2020.

Amit Singer.

Angular synchronization by eigenvectors and semidefinite programming.
Applied and computational harmonic analysis, 30(1):20–36, 2011.