Analysis and Stochastics

Pan Jing Bin

Honours Year Project Final Oral Presentation

Outline

Algebraically Structured Model

General setting: Let $\theta \in \mathbb{R}^d$ be an **unknown** vector (also known as a **signal**). Consider two independent sources of corruptions on θ.

$$P_\theta \sim G\theta + \sigma\xi \tag{1}$$

1. **Additive Gaussian noise:**

$$\xi \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d).$$

   Used to model many small, independent sources of randomness.

2. **Random rotation:** G is drawn **uniformly** via the Haar measure from a compact subgroup $\mathcal{G}$ of the orthogonal group $O(d)$ given by

$$O(d) := \left\{ \boldsymbol{A} \in \mathrm{Mat}_{d\times d}(\mathbb{R}) \ : \ \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}_d = \boldsymbol{A}^T\boldsymbol{A} \right\}.$$

Question

Given independent samples $X_1, \cdots, X_n$ drawn according to the probability distribution (1), recover back the vector θ.

Motivation: Cryogenic Electron Microscopy

1. First introduced in a seminal paper by Bandeira, Rigollet and Weed in 2017.
2. Motivated by recent advancements in a molecular imaging technique in chemistry known as **cryogenic electron microscopy (cryo-EM)**.
3. Excerpt from the 2017 Nobel Prize in chemistry press release :

PRESS RELEASE

4 October 2017

The Nobel Prize in Chemistry 2017

The Royal Swedish Academy of Sciences has decided to award the Nobel Prize in Chemistry 2017 to

Jacques Dubochet
University of Lausanne, Switzerland

Joachim Frank
Columbia University, New York, USA

Richard Henderson
MRC Laboratory of Molecular Biology, Cambridge, UK

"for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution"

Motivation: Cryogenic Electron Microscopy

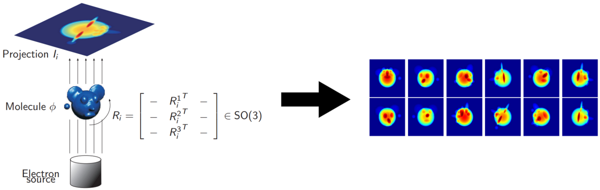


Figure: Taken from [2]. Each projection corresponds to some unknown rotation of the unknown molecule

Freezing the molecules introduce a **large amount of noise** and **randomly rotates the molecule**. But the noise level can be **reduced** with improvements in technology.

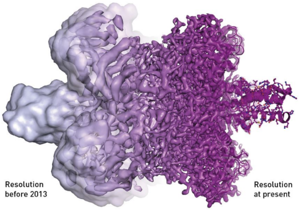


Figure: Taken from https://www.nobelprize.org/prizes/chemistry/2017/popular-information/. The resolution of cryo-EM have drastically improved since 2013

Algebraically Structured Model

Setup:

$$X_i = G_i\theta + \sigma\xi_i$$

where $\theta \in \mathbb{R}^d$, $G_i \sim \mathrm{Haar}(\mathcal{G})$ and $\xi_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_d)$.

Low signal-to-noise ratio: Assume $\dfrac{\|\theta\|}{\sigma} \leq 1$ and $K^{-1} \leq \|\theta\| \leq K$ for some universal constant K.

Problem

How does the **number of samples** needed to estimate θ depend on σ **asymptotically?**

1. The distribution of the G_i's is not necessarily uniform in cryo-EM, but can always be **reduced** to a uniform distribution.

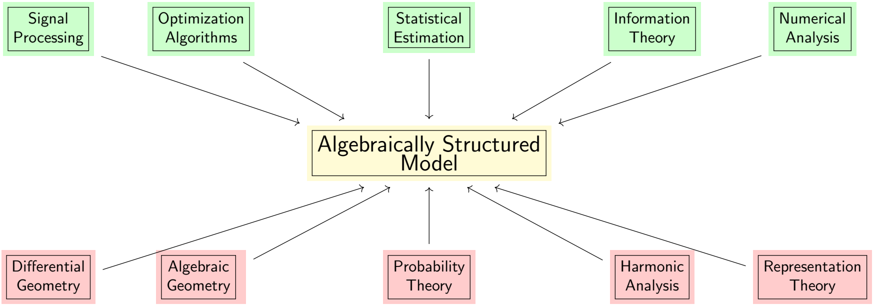
$$X_1, X_2, \cdots, X_n \longrightarrow H_1X_1, H_2X_2, \cdots, H_nX_n, \qquad H_i \sim \mathrm{Haar}(\mathcal{G}).$$

2. Vectors lying in the same $\mathcal{G}$-orbit define **identical** probability distributions, hence we can only hope to recover θ **up to $\mathcal{G}$-orbit**. Define the **invariant distance**

$$\rho(\tilde{\theta}, \theta) := \min_{G \in \mathcal{G}} \|\tilde{\theta} - G\theta\|.$$

Landscape of the Algebraically Structured Model

Tools from many different fields have been brought in to study the model



and the list continues to **grow** over time.

Kullback-Leibler Divergence

Setup:

$$P_\theta \sim G\theta + \sigma\xi.$$

The **Kullback-Leibler Divergence** between two probability distributions P_θ and P_ϕ (with densities f_θ and f_ϕ respectively) is defined to be

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) := \int_{\mathbb{R}^d} f_\theta(x) \log \frac{f_\theta(x)}{f_\phi(x)}\, dx.$$

In general, the **larger** the KL divergence, the **easier** it is to distinguish between the two distributions. Many powerful **passages**

$$\text{Performance of estimators} \longleftrightarrow D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$$

have already been established.

1. If $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \lesssim \sigma^{-k}$, then sampling complexity $\gtrsim \sigma^k$ for **any** estimator;
2. If $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \gtrsim \sigma^{-k}$, then an estimator with sampling complexity $\lesssim \sigma^k$ exist in many cases.

Kullback-Leibler Divergence

Setup:

$$P_\theta \sim G\theta + \sigma\xi.$$

The bulk of the work is to control **how large** $D_{\mathrm{KL}}(P_\theta \parallel P_\phi)$ is.

We have an explicit formula for the density function

$$f_\theta(x) = \frac{1}{\sigma^d (2\pi)^{d/2}} \mathbb{E}_G \left[\exp\left(-\frac{1}{2\sigma^2} \|x - G\theta\|^2 \right) \right],$$

which in turn gives us an explicit formula for the KL divergence

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) = \frac{1}{2\sigma^2} \left(\|\phi\|^2 - \|\theta\|^2 \right) + \mathbb{E}_\xi \left[\log \frac{\mathbb{E}_G \left[\exp\left(\frac{1}{\sigma^2} (\theta + \sigma\xi)^T G\theta \right) \right]}{\mathbb{E}_G \left[\exp\left(\frac{1}{\sigma^2} (\theta + \sigma\xi)^T G\phi \right) \right]} \right].$$

But the formula is rather **complicated**.

Moment Tensors

Given vectors $v^{(1)}, v^{(2)}, \cdots, v^{(m)} \in \mathbb{R}^d$, define the **$m$th order tensor**

$$v^{(1)} \otimes v^{(2)} \otimes \cdots \otimes v^{(m)} \in (\mathbb{R}^d)^{\otimes m} \cong \mathbb{R}^{d^m}$$

to be the **m-dimensional array** whose $(i_1, i_2, \cdots, i_m)$th entry is given by

$$\left(v^{(1)} \otimes v^{(2)} \otimes \cdots \otimes v^{(m)}\right)_{i_1 i_2 \cdots i_m} = v^{(1)}_{i_1} v^{(2)}_{i_2} \cdots v^{(m)}_{i_m}.$$

Let $\theta, \phi \in \mathbb{R}^d$ be vectors. The **mth moment tensor** of θ is defined to be

$$\mathbb{E}_G\left[(G\theta)^{\otimes m}\right] \in (\mathbb{R}^d)^{\otimes m}$$

and the **mth moment difference tensor** between θ and ϕ is defined to be

$$\Delta_m(\theta, \phi) := \mathbb{E}_G\left[(G\theta)^{\otimes m} - (G\phi)^{\otimes m}\right].$$

Moment Tensors and Kullback-Leibler Divergence

What the theorem technically says:

Banderia-Rigollet-Weed (2017)

Let $\theta, \phi \in \mathbb{R}^d$ be vectors satisfying some technical conditions. There exist universal constants $\underline{C}$ and $\overline{C}$ such that for any positive integer k,

$$\underline{C} \sum_{m=1}^{\infty} \frac{\|\Delta_m(\theta,\phi)\|^2}{(\sqrt{3}\sigma)^{2m} m!} \leq D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \leq \overline{C} \left(\sum_{m=1}^{k-1} \frac{\|\Delta_m(\theta,\phi)\|^2}{\sigma^{2m} m!} + \frac{\|\theta\|^{2k-2} \rho(\theta,\phi)^2}{\sigma^{2k}} \right).$$

What the theorem means **in practice:**

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \approx \sigma^{-2k}$$

where k is the **smallest** positive integer such that $\Delta_k(\theta,\phi) \neq 0$.

In the multi-reference alignment model: $k = 2$ or 3.

Challenges in the Algebraically Structured Model

Setup: $P_\theta \sim G\theta + \sigma\xi$.

1. **Highly noisy observations:**

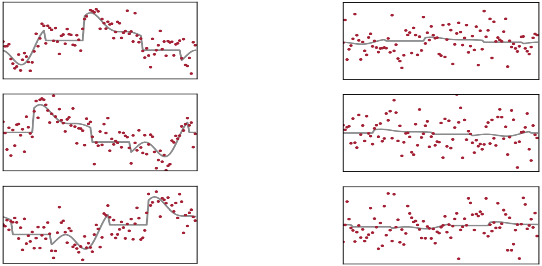


Figure: Figure taken from [1]. The column on the left represents the low noise regime and the column on the right represents the high noise regime.

The random rotation G and the Gaussian noise ξ are **deeply entangled**.

2. **Complexity of the orthogonal group:** Many of the quantities (e.g. moment tensors, KL divergence) involve an expectation over a subgroup of the orthogonal group, which is **complicated**.

Multi-reference Alignment

General setting for Multi-reference Alignment:

$$P_\theta \sim R\theta + \sigma\xi$$

where R is drawn uniformly from the subgroup $\mathcal{R}$ defined by

$$\mathcal{R} := \left\{ R_\ell \ : \ 1 \le \ell \le d \right\} \cong \mathbb{Z}/d\mathbb{Z}$$

and the action of R_ℓ is given by

$$R_\ell\big((\theta_1, \cdots, \theta_d)\big) := \big(\theta_{1+\ell}, \theta_{2+\ell}, \cdots, \theta_{d+\ell}\big).$$

and the indices are taken modulo d.

The multi-reference alignment model is much more **well-understood** because the moment tensors admit much **simpler** descriptions.

$$\mathbb{E}_R\big[(R\theta)^{\otimes m}\big] = \frac{1}{d}\sum_{\ell=1}^{d}(R_\ell\theta)^{\otimes m}.$$

The Discrete Fourier Transform

The **discrete Fourier transform (DFT)** of a vector $\theta \in \mathbb{R}^d$ is given by

$$\hat{\theta}_k := \frac{1}{\sqrt{d}} \sum_{j=1}^{d} e^{\frac{2\pi ijk}{d}} \theta_j, \qquad 1 \leq k \leq d.$$

After passing through the passageway

$$\mathbb{R}^d \xleftrightarrow{\text{DFT}} \left\{\theta \in \mathbb{C}^d \ : \ \theta_j = \overline{\theta}_{-j} \ \ \forall j\right\},$$

explicit formulas for the moment tensors can be obtained:

$$\mathbb{E}_R\left[(\widehat{R\theta})\right]_{k_1} = \begin{cases} \hat{\theta}_0 & \text{if } k_1 \equiv 0 \pmod{d}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\mathbb{E}_R\left[(\widehat{R\theta})^{\otimes 2}\right]_{k_1 k_2} = \begin{cases} |\hat{\theta}_{k_1}|^2 & \text{if } k_1 + k_2 \equiv 0 \pmod{d}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\vdots$$

$$\mathbb{E}_R\left[(\widehat{R\theta})^{\otimes m}\right]_{k_1 \cdots k_m} = \begin{cases} \hat{\theta}_{k_1}\hat{\theta}_{k_2} \cdots \hat{\theta}_{k_m} & \text{if } k_1 + k_2 + \cdots + k_m \equiv 0 \pmod{d}, \\ 0 & \text{otherwise.} \end{cases}$$

Moment Matching

$$\mathbb{E}_R\left[(\widehat{R\theta})^{\otimes m}\right]_{k_1\cdots k_m} = \begin{cases} \hat{\theta}_{k_1}\hat{\theta}_{k_2}\cdots\hat{\theta}_{k_m} & \text{if } k_1+k_2+\cdots+k_m \equiv 0 \pmod{d}, \\ 0 & \text{otherwise.} \end{cases}$$

The Fourier domain is the **natural setting** for the multi-reference alignment model.

Key Idea: Let $\mathcal{S} \subseteq \mathbb{R}^d$ denote the space of all possible signals and let $k \in \mathbb{Z}_{\geq 1}$.

1. **Upper bound:** If there exist two signals $\theta, \phi \in \mathcal{S}$ lying in different orbits such that
$$\mathbb{E}_R\left[(\widehat{R\theta})^{\otimes m}\right] = \mathbb{E}_R\left[(\widehat{R\phi})^{\otimes m}\right]$$
for all $1 \leq m \leq k-1$, then $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \lesssim \sigma^{-2k}$;
2. **Lower bound:** If for all signals $\theta, \phi \in \mathcal{S}$ lying in different orbits, there exists $1 \leq m \leq k-1$ such that
$$\mathbb{E}_R\left[(\widehat{R\theta})^{\otimes m}\right] \neq \mathbb{E}_R\left[(\widehat{R\phi})^{\otimes m}\right],$$
then $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \gtrsim \sigma^{-2k+2}$.

Moment Matching in the Worst Case

No restrictions:

$$\hat{\phi}_j = \begin{cases} 1 & \text{if } j \in \{\pm 1\} \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \hat{\theta}_j = \begin{cases} e^{i\delta} & \text{if } j = 1 \\ e^{-i\delta} & \text{if } j = -1 \\ 0 & \text{otherwise.} \end{cases}$$

for some specially chosen quantity δ.

Banderia-Rigollet-Weed (2017)

If $\mathcal{S} = \mathbb{R}^d$, then sampling complexity $\gtrsim \sigma^{2d}$ (which is **really bad**).

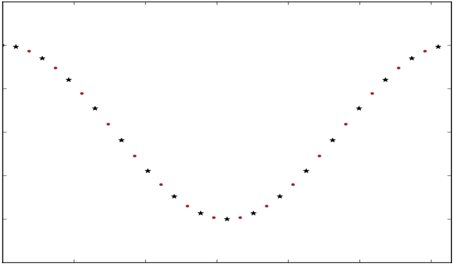


Figure: Figure taken from [1]. The signal ϕ is represented by the red dots and the signal θ is represented by the black stars.

Moment Matching for Generic Signals

Signals having full Fourier support: It is well-known in signal processing that the orbit of θ can be completely recovered from

$$\textbf{DC:} \qquad \mathbb{E}_R\left[(\widehat{R\theta})\right]_i = \begin{cases} \hat{\theta}_0 & \text{if } i \equiv 0 \pmod{d}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\textbf{Power spectrum:} \qquad \mathbb{E}_R\left[(\widehat{R\theta})^{\otimes 2}\right]_{ij} = \begin{cases} |\hat{\theta}_i|^2 & \text{if } i + j \equiv 0 \pmod{d}, \\ 0 & \text{otherwise.} \end{cases}$$

$$\textbf{Bispectrum:} \qquad \mathbb{E}_R\left[(\widehat{R\theta})^{\otimes 3}\right]_{ijk} = \begin{cases} \hat{\theta}_i\hat{\theta}_j\hat{\theta}_k & \text{if } i + j + k \equiv 0 \pmod{d}, \\ 0 & \text{otherwise.} \end{cases}$$

But not from its **first two** moment tensors alone.

$$\hat{\theta} := \left(\hat{\theta}_{-d/2}, \cdots, \hat{\theta}_{-1}, \hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_{d/2}\right)$$

$$\downarrow$$

$$\hat{\phi} := \left(\hat{\theta}_{-d/2}, \cdots, e^{-i\delta}\hat{\theta}_{-1}, \hat{\theta}_0, e^{i\delta}\hat{\theta}_1, \cdots, \hat{\theta}_{d/2}\right).$$

Sampling Complexity for Generic Signals

Perry-Weed-Bandeira-Rigollet-Singer (2017)

For signals having full Fourier support, sampling complexity $\approx \sigma^6$.

A significant improvement from σ^{2d}, but still **pretty bad** in many practical applications.

Question

For which class of signals can a better sampling complexity be obtained?

Sparse Multi-reference Alignment

Sparse signal: Most of the coefficients of the signal are zero.

For a vector $\theta \in \mathbb{R}^d$, consider the multiset

$$\mathcal{D}(\theta) := \left\{ i - j \pmod{d} \ : \ \theta_i, \theta_j \neq 0 \right\} \subseteq \mathbb{Z}/d\mathbb{Z}.$$

If each element in $\mathcal{D}(\theta)$ appears with multiplicity 1, then θ is said to be **collision-free**.

The class of collision-free signals are very well-behaved **locally**.

Ghosh-Rigollet (2021)

If θ, ϕ are collision-free signals and $\rho(\theta, \phi)$ is sufficiently small, then

$$\left\| \mathbb{E}_R\left[(R\theta)^{\otimes 2}\right] - \mathbb{E}_R\left[(R\phi)^{\otimes 2}\right] \right\| \gtrsim \rho(\theta, \phi).$$

We (locally) have $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \approx \sigma^{-4}$.

Maximum Likelihood Estimation

1. If $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \lesssim \sigma^{-k}$, then sampling complexity $\gtrsim \sigma^k$ for **any** estimator;
2. If $D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \gtrsim \sigma^{-k}$, then an estimator with sampling complexity $\lesssim \sigma^k$ exist in many cases.

Question

When does (2) hold for the **maximum likelihood estimator (MLE)?**

$$\tilde{\theta}_n := \operatorname*{argmax}_{\phi \in \mathbb{R}^d} \sum_{i=1}^{n} \log f_\phi(X_i)$$

Banderia-Rigollet-Weed (2017)

Let $\mathcal{L}$ be a subspace of $\mathbb{R}^d$. Suppose that for any $\theta, \phi \in \mathcal{L}$,

$$D_{\mathrm{KL}}(P_\theta \parallel P_\phi) \gtrsim \sigma^{-2k}.$$

Then we have the following upper bound for the rate of estimation of the restricted MLE that is **uniform in θ:**

$$\mathbb{E}_\theta\left[\rho(\tilde{\theta}_n, \theta)\right] \lesssim \frac{\sigma^k}{\sqrt{n}}.$$

Maximum Likelihood Estimation for Sparse Signals

What we already have:

Ghosh-Rigollet (2021)

Suppose that for all ϕ is a **neighbourhood** U of θ,

$$D_{\mathsf{KL}}(P_\theta \parallel P_\phi) \gtrsim \sigma^{-2k}.$$

Then we have the following upper bound for the rate of estimation of the restricted MLE that holds **pointwise**

$$\mathbb{E}_\theta\left[\rho(\tilde{\theta}_n, \theta)\right] \lesssim_\theta \frac{\sigma^k}{\sqrt{n}}.$$

With **finer analysis**, we hope to obtain uniform rates of estimation for the class of collision-free signals as well.

Bibliography

[1] Afonso S Bandeira, Jonathan Niles-Weed, and Philippe Rigollet. Optimal rates of estimation for multi-reference alignment. *Mathematical Statistics and Learning*, 2(1):25–75, 2020.

[2] Amit Singer. Mathematics for cryo-electron microscopy. In *Proceedings of the International Congress of Mathematicians: Rio de Janeiro 2018*, pages 3995–4014. World Scientific, 2018.

The End

Thank you for your attention.

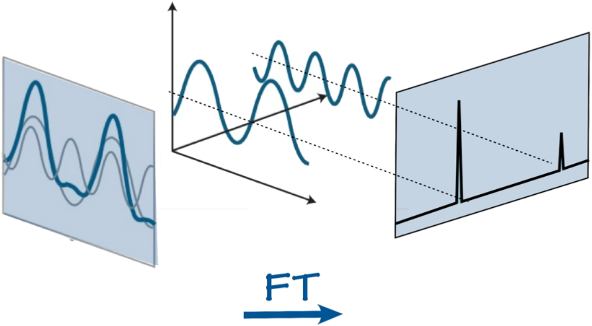


Figure: Image taken from https://mriquestions.com/fourier-transform-ft.html