PRINCETON UNIVERSITY

Pattern Theory seminar series – Brown University

Fundamental Limits in Multireference Alignment

Nir Sharon

PACM, Princeton University

October 11, 2017

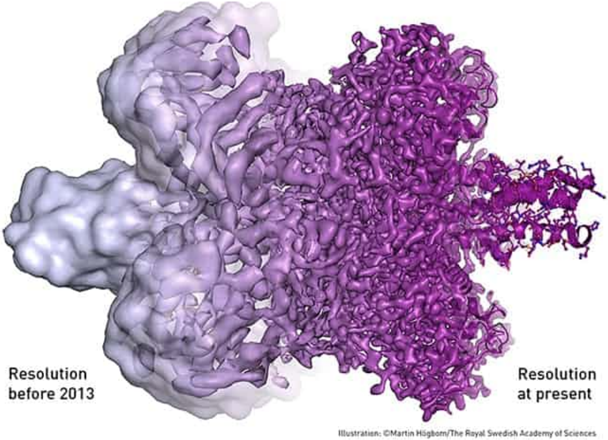
Agenda

1. The Problem of MultiReference Alignment (MRA)
2. Moments Approach and Theoretical Guarantees
3. Numerical Examples and Conclusions

Table of contents

Motivational Section – Cryo-EM in a Nutshell

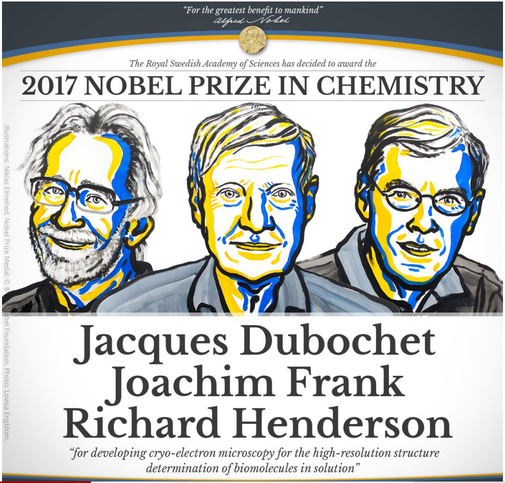
Appear in Olena Shmahalo/Quanta Magazine; source: Martin Högborn/The Royal Swedish Academy of Sciences:



The "resolution revolution" (Werner Kühlbrandt, 2014)

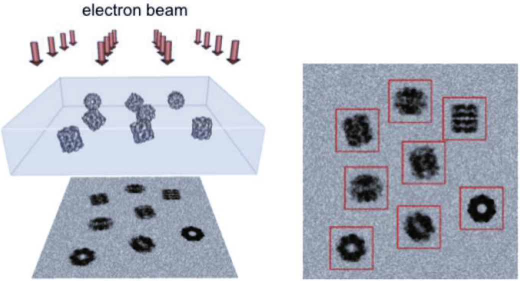
Motivational Section – Cryo-EM in a Nutshell

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



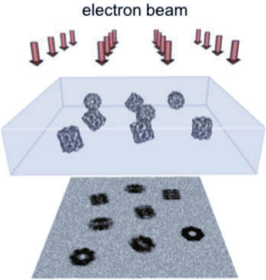
Motivational Section – Cryo-EM in a Nutshell

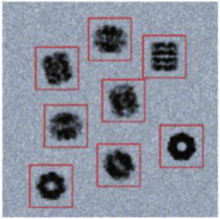
The process:



Motivational Section – Cryo-EM in a Nutshell

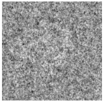
The process:





Main computational challenges:

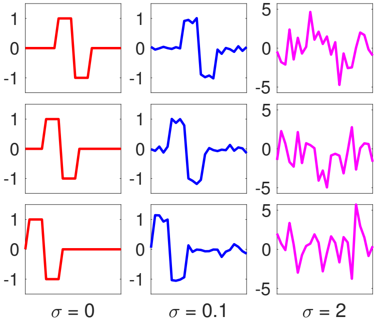
1. High level of noise.
2. Unknown viewing directions.



The Problem of Multi-Reference Alignment (MRA)

Formulation: Estimating a signal $x \in \mathbb{R}^L$, up to shifting, from its noisy circularly–translated copies

$$y_j = R_{r_j} x + \varepsilon_j, \quad j = 1, \dots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I).$$

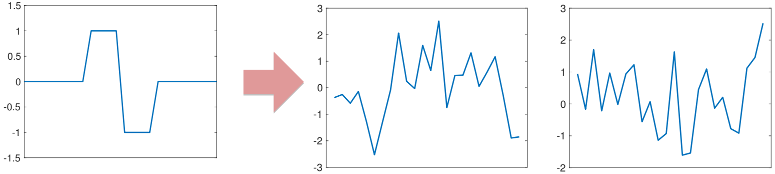


Applications: radar, image registration, structural biology

The Sample Complexity of the Problem

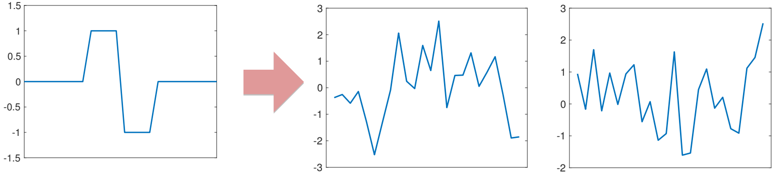
The Sample Complexity of the Problem

Data:



The Sample Complexity of the Problem

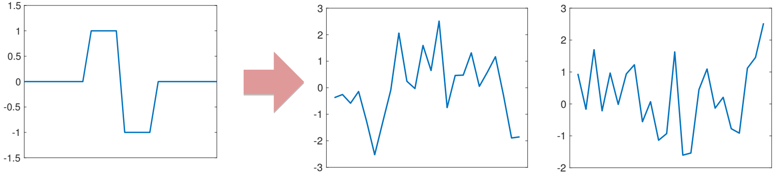
Data:



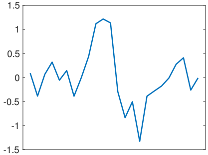
1. **Can we recover the signal for any level of noise?**
2. **How many samples do we need for attaining a certain accuracy?**

The Sample Complexity of the Problem

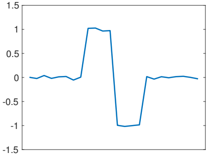
Data:



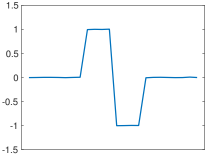
Estimations:



$N = 1,000$



$N = 100,000$



$N = 10,000,000$

The Role of Translations in MRA

Recall the model:

$$y_j = R_{r_j}x + \varepsilon_j, \quad j = 1, \ldots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I).$$

- The translations are the latent/hidden variables of the problem.

The Role of Translations in MRA

Recall the model:

$$y_j = R_{r_j} x + \varepsilon_j, \quad j = 1, \ldots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I).$$

- The translations are the latent/hidden variables of the problem.
- Given the translations, we can estimate

$$\tilde{x} = \frac{1}{N} \sum_{j=1}^{N} R_{r_j}^{-1} y_j.$$

Therefore, estimating the translation reduces the problem significantly.

The Role of Translations in MRA

Recall the model:

$$y_j = R_{r_j} x + \varepsilon_j, \quad j = 1, \ldots, N, \quad \varepsilon_j \sim \mathcal{N}(0, \sigma^2 I).$$

- The translations are the latent/hidden variables of the problem.
- Given the translations, we can estimate

$$\tilde{x} = \frac{1}{N} \sum_{j=1}^{N} R_{r_j}^{-1} y_j.$$

  Therefore, estimating the translation reduces the problem significantly.

1. **Should we find the translations to estimate the signal?**
2. **Can we recover the signal without explicit translations?**

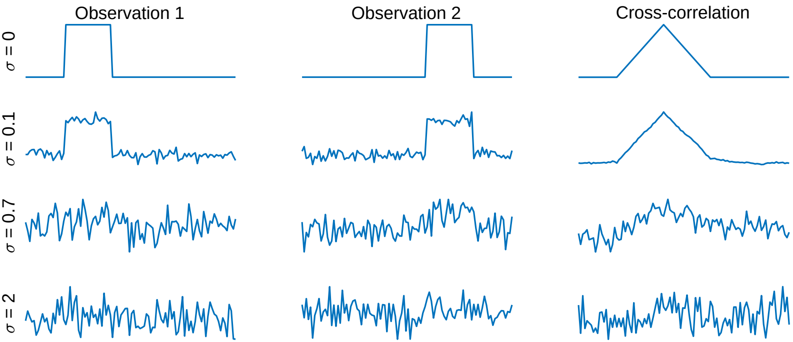
Low vs High Level of Noise

Low vs High Level of Noise

- Low level of noise $\rightarrow$ Explicitly estimating the translations.

Low vs High Level of Noise

- Low level of noise $\rightarrow$ Explicitly estimating the translations.



An example – alignment via cross-correlation

Low vs High level of noise Regimes

- Low level of noise $\rightarrow$ Explicitly estimating the translations.

- High level of noise $\rightarrow$ Keeping translations behind the scene:

Low vs High level of noise Regimes

- Low level of noise $\rightarrow$ Explicitly estimating the translations.

- High level of noise $\rightarrow$ Keeping translations behind the scene:
  - Invariant features – the info is encompassed in mean, power spectrum, and bispectrum.
  - EM – a classical statistical method with state-of-the-art performance for MRA.

Low vs High level of noise Regimes

- Low level of noise $\rightarrow$ Explicitly estimating the translations.
- High level of noise $\rightarrow$ Keeping translations behind the scene:
- The sample complexity tradeoff,

$$N \gtrsim \sigma^2 \longleftrightarrow N \gtrsim \sigma^6$$

Table of contents

Statistical Moments Approach – Including the Distribution

Statistical Moments Approach – Including the Distribution

- The first moment of the data is

$$\mathbb{E}[y] = C_x \rho = x * \rho,$$

where C_x is a circulant matrix and ρ is the distribution of translations.

Statistical Moments Approach – Including the Distribution

- The first moment of the data is

$$\mathbb{E}\left[y\right] = C_x \rho = x * \rho,$$

where C_x is a circulant matrix and ρ is the distribution of translations.

- The second moment of the data is

$$\mathbb{E}\left[yy^T\right] = C_x D_\rho C_x^T + \sigma^2 I,$$

where D_ρ is a diagonal matrix with ρ on its diagonal.

Statistical Moments Approach – Including the Distribution

- The first moment of the data is

$$\mathbb{E}\left[y\right] = C_x \rho = x * \rho,$$

where C_x is a circulant matrix and ρ is the distribution of translations.

- The second moment of the data is

$$\mathbb{E}\left[yy^T\right] = C_x D_\rho C_x^T + \sigma^2 I,$$

where D_ρ is a diagonal matrix with ρ on its diagonal.

Proposition

Assume the DFT of ρ satisfies $\hat{\rho}[k] \neq 0$ for some k, where k and L are coprime. If the DFT of x is non-vanishing, then it is uniquely determined (up to translation) from the first two moments of the data.

Statistical Moments Approach – Including the Distribution

- The first moment of the data is

$$\mathbb{E}\left[y\right] = C_x \rho = x * \rho,$$

where C_x is a circulant matrix and ρ is the distribution of translations.

- The second moment of the data is

$$\mathbb{E}\left[yy^T\right] = C_x D_\rho C_x^T + \sigma^2 I,$$

where D_ρ is a diagonal matrix with ρ on its diagonal.

- Assume $\left|\hat{x}[k]\right| = 1$, we have

$$C_x D_\rho C_x^T = C_x D_\rho C_x^{-1},$$

and x is recovered whenever ρ has a distinct entry.

Resolving Distinct Eigenvalue

- The success of the spectral algorithm is contingent on a distinct entry in ρ. Can we somehow guarantee it?

Resolving Distinct Eigenvalue

- The success of the spectral algorithm is contingent on a distinct entry in ρ. Can we somehow guarantee it?
- The answer: Yes! by randomly reshuffle the samples!

Resolving Distinct Eigenvalue

- The success of the spectral algorithm is contingent on a distinct entry in ρ. Can we somehow guarantee it?

- The answer: Yes! by randomly reshuffle the samples!

Proposition

Let ρ be a non-periodic vector on the simplex and let θ be a random probability density function on the simplex. Then, all entries of $\rho * \theta$ are distinct with probability 1.

Resolving Distinct Eigenvalue

- The success of the spectral algorithm is contingent on a distinct entry in ρ. Can we somehow guarantee it?
- The answer: Yes! by randomly reshuffle the samples!

Proposition

Let ρ be a non-periodic vector on the simplex and let θ be a random probability density function on the simplex. Then, all entries of $\rho * \theta$ are distinct with probability 1.

A conclusion: the spectral algorithm provides a constructive solution for any nonperiodic ρ and achieves a sample complexity of $N \gtrsim \sigma^4$ (equivalently of order $1/\mathrm{SNR}^2$).

The Spectral Algorithm in Full

The Spectral Algorithm in Full

1. Estimate moments and power spectrum
   1. $\mu \leftarrow \frac{1}{N}\sum_{j=1}^{N} y_j$
   2. $M \leftarrow \frac{1}{N}\sum_{j=1}^{N} y_j y_j^T - \sigma^2 I,$ where $M \approx C_x D_\rho C_x^T$
   3. $P_x \leftarrow \frac{1}{N}\sum_{j=1}^{N} |Fy_j|^2 - \sigma^2 L\mathbf{1}$

The Spectral Algorithm in Full

1. Estimate moments and power spectrum
   1. $\mu \leftarrow \frac{1}{N}\sum_{j=1}^{N} y_j$
   2. $M \leftarrow \frac{1}{N}\sum_{j=1}^{N} y_j y_j^T - \sigma^2 I,$ where $M \approx C_x D_\rho C_x^T$
   3. $P_x \leftarrow \frac{1}{N}\sum_{j=1}^{N} |F y_j|^2 - \sigma^2 L\mathbf{1}$
2. Normalize the signal
   1. $p \leftarrow (P_x)^{-1/2}$
   2. $Q \leftarrow F^{-1} D_p F$
   3. $\widetilde{M} \leftarrow Q\ M\ Q^{-1}$ Here $\widetilde{M} \approx C_{x'} D_\rho C_{x'}^T$ with $C_{x'} = C_{x'}^T$

The Spectral Algorithm in Full

1. Estimate moments and power spectrum
   1. $\mu \leftarrow \frac{1}{N}\sum_{j=1}^{N} y_j$
   2. $M \leftarrow \frac{1}{N}\sum_{j=1}^{N} y_j y_j^T - \sigma^2 I,$ where $M \approx C_x D_\rho C_x^T$
   3. $P_x \leftarrow \frac{1}{N}\sum_{j=1}^{N} |Fy_j|^2 - \sigma^2 L\mathbf{1}$
2. Normalize the signal
   1. $p \leftarrow (P_x)^{-1/2}$
   2. $Q \leftarrow F^{-1} D_p F$
   3. $\widetilde{M} \leftarrow Q\ M\ Q^{-1}$ Here $\widetilde{M} \approx C_{x'} D_\rho C_{x'}^T$ with $C_{x'} = C_{x'}^T$
3. Extract and scale solutions
   1. $v \leftarrow \mathrm{UniqEig}(\widetilde{M})$
   2. $\tilde{v} \leftarrow F^{-1}\left((P_x)^{1/2} \odot Fv\right)$ Reset the Fourier modulus
   3. $x \leftarrow \left(\mathrm{Sum}(\mu)/\,\mathrm{Sum}(\tilde{v})\right)\tilde{v}$ Scale by the first moment
   4. $\rho \leftarrow C_x^{-1}\mu$

Estimation Error – Spectral Algorithm

Consider $\sigma \to \infty$ and suppose we have $\sigma_1 \leq \sigma_2 \leq \ldots$, such that $\sigma_n \to \infty$. Also, let $N_1 \leq N_2 \leq \ldots$. For each n, we draw observations $y_1, \ldots, y_{N_n}$ at noise level σ_n.

Estimation Error – Spectral Algorithm

Consider $\sigma \to \infty$ and suppose we have $\sigma_1 \leq \sigma_2 \leq \ldots$, such that $\sigma_n \to \infty$. Also, let $N_1 \leq N_2 \leq \ldots$. For each n, we draw observations $y_1, \ldots, y_{N_n}$ at noise level σ_n.

Theorem

Let ρ be periodic. Then, for sufficiently small $t > 0$:

$$\mathbb{P}\left[\min_s \|R_s \hat{x} - x\| \geq t\right] \leq C_1 \exp\left\{-C_2 \frac{N_n}{\sigma_n^4} t\right\},$$

where $C_1 = C_1(x, \rho, L)$ and $C_2 = C_2(x, \rho, L)$ are finite, positive constants.

Estimation Error – Spectral Algorithm

Consider $\sigma \to \infty$ and suppose we have $\sigma_1 \leq \sigma_2 \leq \dots$, such that $\sigma_n \to \infty$. Also, let $N_1 \leq N_2 \leq \dots$. For each n, we draw observations $y_1, \dots, y_{N_n}$ at noise level σ_n.

Theorem

Let ρ be periodic. Then, for sufficiently small $t > 0$:

$$\mathbb{P}\left[\min_s \|R_s\hat{x} - x\| \geq t\right] \leq C_1 \exp\left\{-C_2 \frac{N_n}{\sigma_n^4} t\right\},$$

where $C_1 = C_1(x, \rho, L)$ and $C_2 = C_2(x, \rho, L)$ are finite, positive constants.

Therefore, if $N_n \geq K \log(n)\sigma_n^4$ for a sufficiently large constant K, then the error of $\hat{x}_n$ converges to 0 almost surely as $n \to \infty$.

The Periodicity is Tight

Is non-periodicity a real property of the problem or maybe just an artifact of the construction?

The Periodicity is Tight

Is non-periodicity a real property of the problem or maybe just an artifact of the construction?

Proposition

Let $\ell < L/2$ be a divisor of $L > 1$. Suppose that ρ is a periodic distribution with period of ℓ. Then, for a given real signal x_1 with non-vanishing DFT, there exists a different real signal x_2 (which is not a translation of x_1) such that both signals have the same first and second moments.

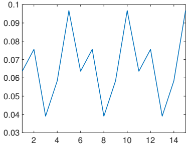
The Periodicity is Tight

Is non-periodicity a real property of the problem or maybe just an artifact of the construction?

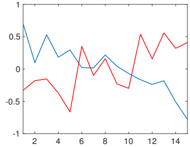
Proposition

Let $\ell < L/2$ be a divisor of $L > 1$. Suppose that ρ is a periodic distribution with period of ℓ. Then, for a given real signal x_1 with non-vanishing DFT, there exists a different real signal x_2 (which is not a translation of x_1) such that both signals have the same first and second moments.

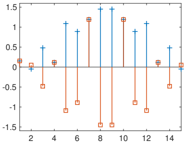
A visual example:



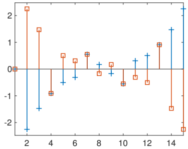
Periodic dist.



Two signals



Real Fourier



Imag. Fourier

Lower Bound

Let $\phi_x(\hat{x}) := \operatorname{argmin}_{z \in \{R_\ell \hat{x}\}_{\ell \in \mathbb{Z}_L}} \|z - x\|$ and define Minimal Square Error (MSE) to be

$$\mathsf{MSE} = \frac{1}{\|x\|^2} \mathrm{E}\left[\|\phi_x(\hat{x}) - x\|^2 \right].$$

Lower Bound

Let $\phi_x(\hat{x}) := \text{argmin}_{z \in \{R_\ell \hat{x}\}_{\ell \in \mathbb{Z}_L}} \|z - x\|$ and define Minimal Square Error (MSE) to be

$$\text{MSE} = \frac{1}{\|x\|^2} \mathrm{E}\left[\|\phi_x(\hat{x}) - x\|^2 \right].$$

Theorem

Assume that x is not a constant vector. If $\hat{x}$ is an asymptotically unbiased estimator of x, that is $\phi_x(\hat{x}) \to x$, then

$$MSE \geq \mathcal{O}\left(\frac{1}{SNR^2} \right).$$

Moreover, if ρ is periodic, with a period $\ell < \frac{L}{2}$, then

$$MSE \geq \mathcal{O}\left(\frac{1}{SNR^3} \right).$$

Dependency on L via Spike Covariance Model

In the spiked model, given a rank r matrix X, we observe

$$Y = X + G \in \mathbb{R}^{L \times N}, \quad g_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Usually, it assumes that $L = L(N)$ and $L/N \to \gamma > 0$ as $N \to \infty$.

Dependency on L via Spike Covariance Model

In the spiked model, given a rank r matrix X, we observe

$$Y = X + G \in \mathbb{R}^{L\times N}, \quad g_{ij} \overset{iid}{\sim} \mathcal{N}(0,\sigma^2).$$

Usually, it assumes that $L = L(N)$ and $L/N \to \gamma > 0$ as $N \to \infty$.
Let λ be the top eigenvalue of XX^T/N. Then, the phase transition at

$$\lambda_{critical} = \sigma^2\sqrt{\gamma}.$$

means that for $\lambda > \lambda_{critical}$ there is a non-trivial correlation between top eigenvectors of YY^T/N and XX^T/N.

Dependency on L via Spike Covariance Model

In the spiked model, given a rank r matrix X, we observe

$$Y = X + G \in \mathbb{R}^{L \times N}, \quad g_{ij} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2).$$

Usually, it assumes that $L = L(N)$ and $L/N \to \gamma > 0$ as $N \to \infty$.
Let λ be the top eigenvalue of XX^T/N. Then, the phase transition at

$$\lambda_{critical} = \sigma^2 \sqrt{\gamma}.$$

means that for $\lambda > \lambda_{critical}$ there is a non-trivial correlation between top eigenvectors of YY^T/N and XX^T/N.
In MRA it is equal to

$$N \geq \frac{L\sigma^4}{\|x\|^4 (\max \rho)^2} = \frac{L}{(\max \rho)^2} \frac{1}{\mathrm{SNR}^2}.$$

Table of contents

Two additional algorithms – LS optimization

The least square algorithm aim to minimize

$$\min_{\tilde{x}\in\mathbb{R}^L,\tilde{\rho}\in\Delta^L} \|\hat{M}_y - C_{\tilde{x}} D_{\tilde{\rho}} C_{\tilde{x}}^T\|_{\mathrm{F}}^2 + \lambda\|\hat{\mu}_y - C_{\tilde{x}}\tilde{\rho}\|_2^2,$$

where $\lambda > 0$ is a predefined parameter.

In low SNR regime, the variance of the first estimator is proportional to σ^2 and the variance of the second is proportional to $3L\sigma^4$. Therefore, we set $\lambda = \frac{1}{L(1+3\sigma^2)}$.

We solve the problem with a gradient-based method.

Two additional algorithms – adapted EM

By the model of the MRA,

$$p(y,\ell|x,\rho) = \prod_{j=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}\left\|R_{r_j}x-y_j\right\|^2} \rho[r_j].$$

The log-likelihood function is then given, up to a constant, by

$$\log L(x,\rho|y,r) = \sum_{j=1}^{N} \left\{ \log \rho[r_j] - \frac{1}{2\sigma^2} \left\| R_{r_j}x - y_j \right\|^2 \right\}.$$

To maximize the expectation of the marginal log-likelihood we use the iteration

$$x_{k+1} = \frac{1}{N} \sum_{j=1}^{N} \sum_{\ell=0}^{L-1} w_k^{\ell,j} R_\ell^{-1} y_j \quad \text{and} \quad \rho_{k+1}[\ell] = \frac{1}{N} \sum_{k=1}^{N} \frac{w_k^{\ell,j}}{\sum_{\ell'=0}^{L-1} w_k^{\ell',j}}.$$

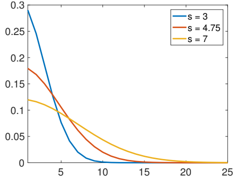
for weights which are iteratively updated according to the data.

$$w_k^{\ell,j} = \mathbb{P}[r_j = \ell | y, x_k, \rho_k].$$

EM: Standard vs Adapted

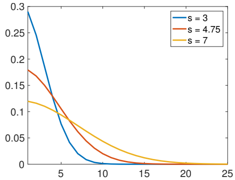
EM: Standard vs Adapted

A family of distributions

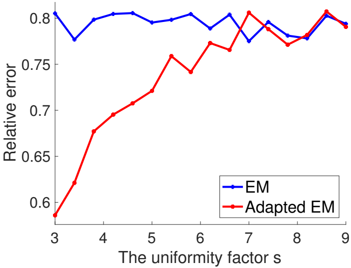


EM: Standard vs Adapted

A family of distributions



A comparison of standard EM and our adapted version for non-uniform distributions,

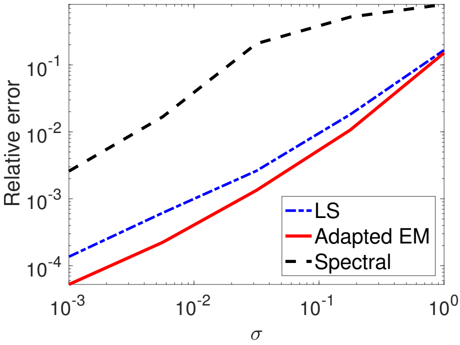


Comparing the Different Methods

The relative error as a function of varying level of noise, σ.

Comparing the Different Methods

The relative error as a function of varying level of noise, σ.



Numerical Error Rates via LS Optimization

We expect the (unsquared) error to be

$$\min_{s \in (Z)^L} \|R_s \hat{x} - x\| \leq \frac{\sigma^d}{\sqrt{N}},$$

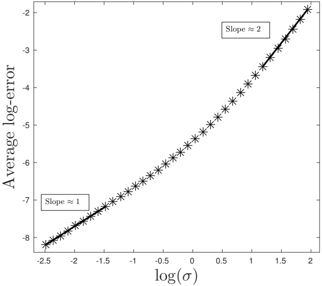
where $d = 1$ for low level of noise (via alignment) whereas $d = 2$ for high level of noise (as implied by the spectral algorithm).

Numerical Error Rates via LS Optimization

We expect the (unsquared) error to be

$$\min_{s \in (Z)^L} \| R_s \hat{x} - x \| \leq \frac{\sigma^d}{\sqrt{N}},$$

where $d = 1$ for low level of noise (via alignment) whereas $d = 2$ for high level of noise (as implied by the spectral algorithm).



Numerical evidence to the error rates: least squares fitting in a log-log plot.

Concluding Remarks

- We can solve the MRA problem in regimes of high level of noise by using low order statistics and simultaneously targeting the signal and the distribution.

- The sample complexity of the problem is $N \gtrsim \sigma^4$ for aperiodic distributions and $N \gtrsim \sigma^6$ for periodic distributions. That makes the aperiodic case easier!

- Numerical examples confirm the theory and show the advantage one gets of including the distribution into the estimator model.

Collaborators from PACM

Emmanuel Abbe

Tamir Bendory

Will Leeb

João Pereira

Amit Singer

Th-th-th-that's all folks!

Thank you

Reference:

E. Abbe, T. Bendory, W. Leeb, J. a. Pereira, N. Sharon, and A. Singer. Multireference alignment is Easier with aperiodic translation distributions. 2017.