A nonconvex optimization approach to spectral estimation

Weilin Li
City University of New York
Graduate Center and City College
Email: wli6@ccny.cuny.edu

1W-MINDS Seminar
February 12, 2026

Outline

Spectral estimation

Definition (Spectral estimation)

Estimate the parameters $\{(\theta_j, a_j)\}_{j=1}^s$ given data

$$\widetilde{y}(x) = \sum_{j=1}^{s} a_j e^{i\theta_j x} + \eta(x) \quad \text{for all} \quad x \in \{-m+1, \ldots, m-1\}.$$

Other names: Super-resolution, sparse spike recovery, line spectral estimation, off-the-grid compressed sensing, and spectral compressed sensing.

Terminology:

- "Frequency" $\theta_j \in \mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$
- "Amplitude" $a_j \in \mathbb{C}$ such that $|a_j| \geq 1$
- "Sparsity" s
- "Noise" η which can be deterministic or random
- "Sampling set" $\{-m+1, \ldots, m-1\}$
- "Number of samples" $2m - 1 \geq 2s$

Some applications: Direction-of-arrival, communications, line spectral estimation, system identification, and microscopy.

Rayleigh length and minimum separation

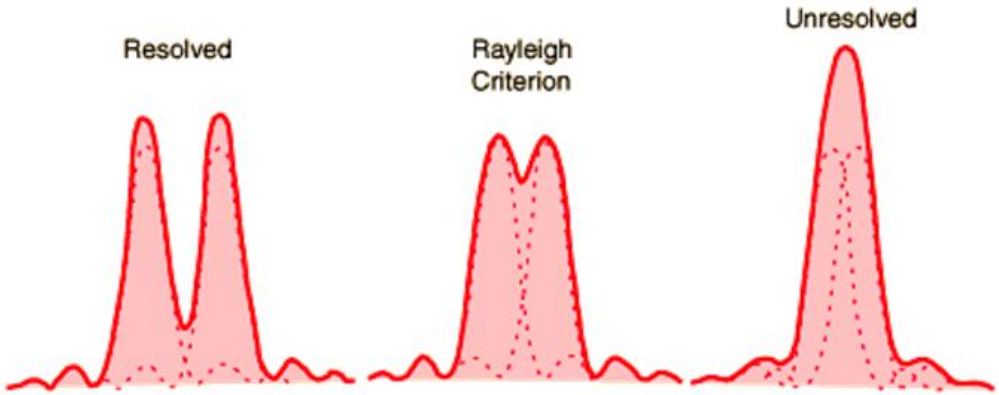


Figure: Rayleigh's criteria

Heuristic was studied rigorously in [Donoho, 1992].

- Rayleigh length $= \frac{\pi}{m}$.
- Minimum separation $\Delta := \min_{j \neq k} |\theta_j - \theta_k|$.

Two regimes

$m\Delta \gg \pi$

- Problem is very stable.
- Ample number of measurements.
- A good algorithm should estimate $\{\theta_j\}_{j=1}^s$ with accuracy proportional to the noise level with reasonable constant. In principle, stability should not depend on s or Δ, and only on m and η.
- What is the best?
  This talk!

$m\Delta \ll \pi$

- Problem is mildly to severely unstable.
- Limited number of measurements.
- In general, no algorithm can estimate $\{\theta_j\}_{j=1}^s$ with accuracy better than $Cm^{-1}(m\Delta)^{-2\lambda+2}\|\eta\|_\infty$, where $\lambda \geq 2$ is the size of largest "cluster", [Batenkov, Goldman, Yomdin, 2021].
- For arbitrary Δ and fixed s, the ESPRIT algorithm provably estimates $\{\theta_j\}_{j=1}^s$ with accuracy $C(m\Delta)^{-2\lambda+2}\|\eta\|_\infty$.
  My 1W-MINDS talks from Fall 2021.

Goals of this talk

1. Results
   Describe a new and optimal spectral estimation algorithm when $m\Delta \geq 8\pi$.

Goals of this talk

1. Results
   Describe a new and optimal spectral estimation algorithm when $m\Delta \geq 8\pi$.
2. Motivation and methodology
   Describe how this algorithm works and how it is different from other methods.

Goals of this talk

1. Results
   Describe a new and optimal spectral estimation algorithm when $m\Delta \geq 8\pi$.
2. Motivation and methodology
   Describe how this algorithm works and how it is different from other methods.
3. Generalization and philosophy
   Explain why it is a computational framework.

Outline

Main result for deterministic noise

Theorem (Fannjiang, L., Liao, 2025)

Let $m \geq 100$ *and* $\{(\theta_j, a_j)\}_{j=1}^s$ *such that* $m\Delta \geq 8\pi$ *and* $\|a\|_\infty \leq 10$.
Let $p \in [1, \infty]$ *and* η *such that* $\|\eta\|_p \leq cm^{1/p}$ *for a small enough absolute* $c > 0$.
Gradient-MUSIC converges at a linear rate to $\{\widetilde{\theta}_j\}_{j=1}^s$ *where*

$$\max_{j=1,\ldots,s} |\theta_j - \widetilde{\theta}_j| \lesssim \frac{\|\eta\|_p}{m^{1+1/p}}.$$

Sharpness:

- Under the assumption $m\Delta \geq 8\pi$, conclusion is the minimax optimal rate in $\|\eta\|_p$ and m, for all $p \in [1, \infty]$.
- For the conclusion to hold, the separation condition $m\Delta \geq 8\pi$ is necessary, up modulo a universal constant.
- Noise assumption $\|\eta\|_p \leq cm^{1/p}$ is necessary to get non-vacuous bounds for frequencies and amplitudes, regardless of method used.

Implications and interpretations

- **Benefits of oversampling, even for deterministic noise.** (ℓ^∞ theory)
  Fix $\{(\theta_j, a_j)\}_{j=1}^s$. Suppose $\|\eta\|_\infty \le c$. The noiseless signal strength $|y(x)|$ and noise strength $|\eta(x)|$ do not depend on m. Yet, the error goes to zero,

$$\max_{j=1,\dots,s} |\theta_j - \widetilde{\theta}_j| \lesssim \frac{\|\eta\|_\infty}{m} \to 0 \quad \text{as} \quad m \to \infty.$$

Implications and interpretations

- **Benefits of oversampling, even for deterministic noise.** (ℓ^∞ theory)
  Fix $\{(\theta_j, a_j)\}_{j=1}^s$. Suppose $\|\eta\|_\infty \le c$. The noiseless signal strength $|y(x)|$ and noise strength $|\eta(x)|$ do not depend on m. Yet, the error goes to zero,

$$\max_{j=1,\dots,s} |\theta_j - \widetilde{\theta}_j| \lesssim \frac{\|\eta\|_\infty}{m} \to 0 \quad \text{as} \quad m \to \infty.$$

- **Unexpected behavior in noise-to-signal ratio.** (ℓ^2 theory)
  Fix $\{(\theta_j, a_j)\}_{j=1}^s$. Suppose $\|\eta\|_2 \le c\sqrt{m}$. The noise-to-signal ratio is

$$\gamma^2 := \frac{\|\eta\|_2^2}{\|y\|_2^2} \asymp \frac{\|\eta\|_2^2}{m}.$$

  Yet, the theory tells us that

$$\max_{j=1,\dots,s} |\theta_j - \widetilde{\theta}_j| \lesssim \frac{\|\eta\|_2}{m^{3/2}} \asymp \frac{\gamma}{m} \to 0 \quad \text{as} \quad m \to \infty.$$

Comparison to other algorithms

- **Prony's method**. Uses exactly $2s$ samples, discards the rest. Cannot utilize oversampling.
- **ESPRIT.** If $m\Delta \gtrsim 1$ and other assumptions hold, [L., Liao, Fannjiang, 2020] proved that ESPRIT has error at most

$$\frac{C\|\eta\|_2}{\sqrt{m}}.$$

- **Convex optimization methods.** If $m\Delta \gtrsim 1$ and other assumptions hold, [Candés, Fernandez-Granda, 2013 and 2014] and [Azais, De Castro, Gamboa, 2015] showed that convex methods have error at most

$$\frac{C\sqrt{\|\eta\|_2}}{m}.$$

- **Sparse Fourier transform.** [Price, Song 2015] showed that if samples are collected in the continuous interval $[-m, m]$, $m\Delta \gtrsim \log(s/\delta)$, and other assumptions hold, then w.p. at least $1-\delta$, error is at most

$$\frac{1}{m}\left(\frac{1}{m}\|\eta\|_{L^2}^2 + \delta\|a\|_2^2\right)^{1/2} \asymp \frac{\|\eta\|_{L^2}}{m^{3/2}} \quad \text{if} \quad \delta = \frac{\|\eta\|_{L^2}^2}{m\|a\|_2^2}.$$

Main result for nonstationary independent Gaussian noise

Theorem (Fannjiang, L., Liao, 2025, simplified form)

Let $m \geq 100$ *and* $\{(\theta_j, a_j)\}_{j=1}^s$ *such that* $m\Delta \geq 8\pi$ *and* $\|a\|_\infty \leq 10$.
Let $\eta \sim \mathcal{N}(0, \Sigma)$, *where* $\Sigma = diag\{\sigma(1+|k|)^{2r}\}_{k=-m+1}^{m-1}$ *for some* $r \in (-\frac{1}{2}, \frac{1}{2})$.
With probability at least $1 - o(m)$, *Gradient-MUSIC converges at a linear rate to* $\{\widetilde{\theta}_j\}_{j=1}^s$ *such that*

$$\max_j |\theta_j - \widetilde{\theta}_j| \lesssim \frac{\sigma\sqrt{\log(m)}}{m^{3/2-r}}.$$

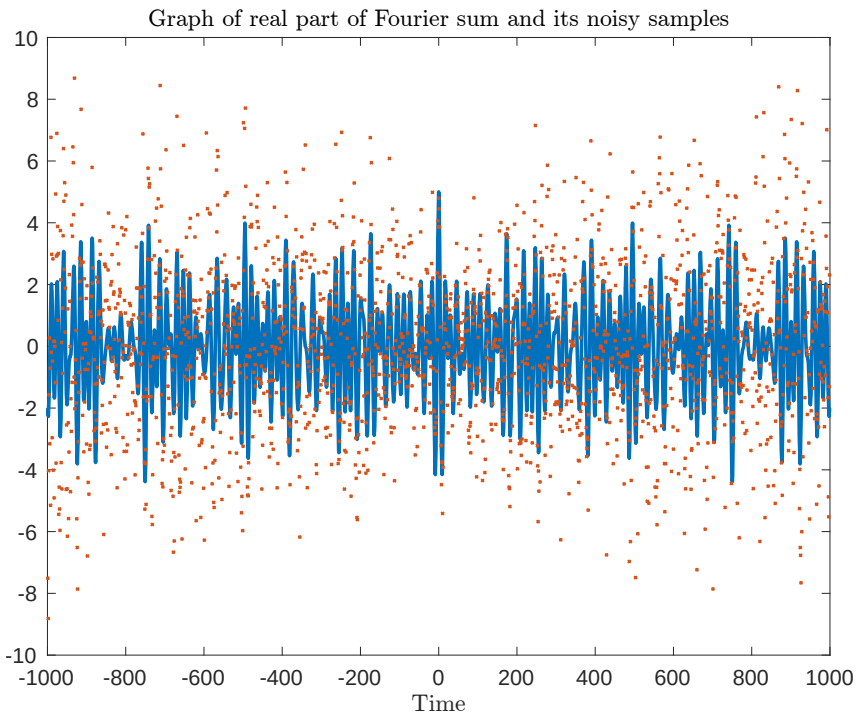
Remarks:

- For $r = 0$, this rate matches the Cramer-Ráo lower bound derived in [Stoica, Nehorai, 1989] up to $\log$ factors. ESPRIT is also optimal for this [Z. Ding et. al., 2024].
- For $r \in [0, \frac{1}{2})$, Gradient-MUSIC matches performance of nonlinear least squares (with good initialization), latter due to [L. Ying, 2025]. Both Gradient-MUSIC and NLS fail for $r \geq \frac{1}{2}$.

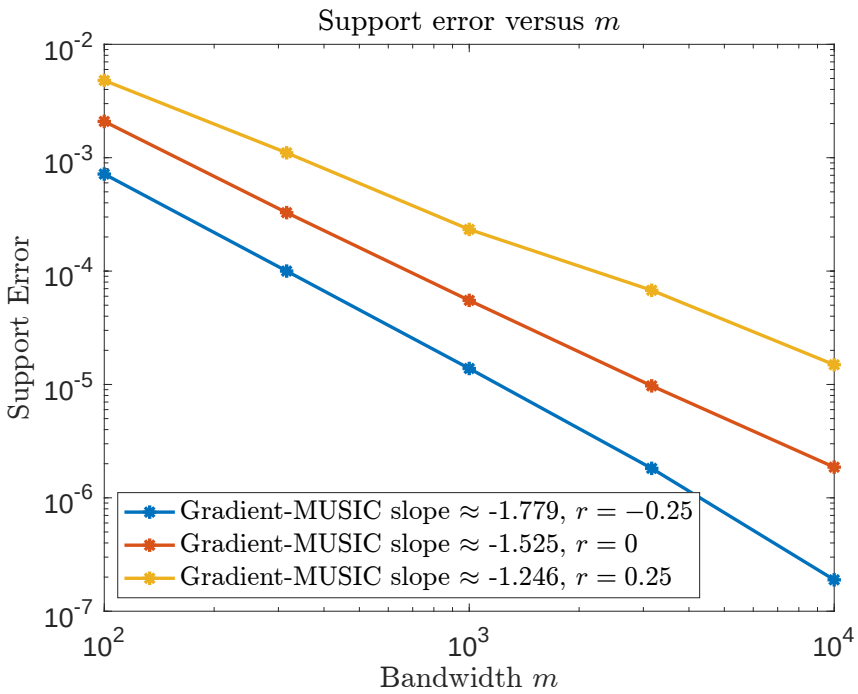
A surprising (?) consequence

Let $\eta \sim \mathcal{N}(0, \Sigma)$, where $\Sigma = \text{diag}\{\sigma(1+|k|)^{2r}\}_{k=-m+1}^{m-1}$ for $r = \frac{1}{4}$.
Theorem states that

$$\max_j |\theta_j - \widetilde{\theta}_j| \lesssim \frac{\sigma\sqrt{\log(m)}}{m^{5/4}}.$$



(a) The true function and its noisy samples.



(b) Verification of theorem for various r.

Outline

Classical MUSIC

Multiple emitter location and **signal** parameter estimation

R **Schmidt** - IEEE transactions on antennas and propagation, 1986 - ieeexplore.ieee.org

... This report is concerned first with the **multiple** emitter aspect of this problem and second with the generality of solution. A description is given of the **multiple signal classification** (MUSIC) ...

☆ Save 99 Cite Cited by 20462 Related articles All 9 versions

1. Let $\widetilde{U}$ be the leading s-dimensional left singular space of the $m \times m$ Hankel matrix generated from data $\widetilde{y}$.
2. Create the *MUSIC function* $\widetilde{q}\colon \mathbb{T} \to [0,1]$ by

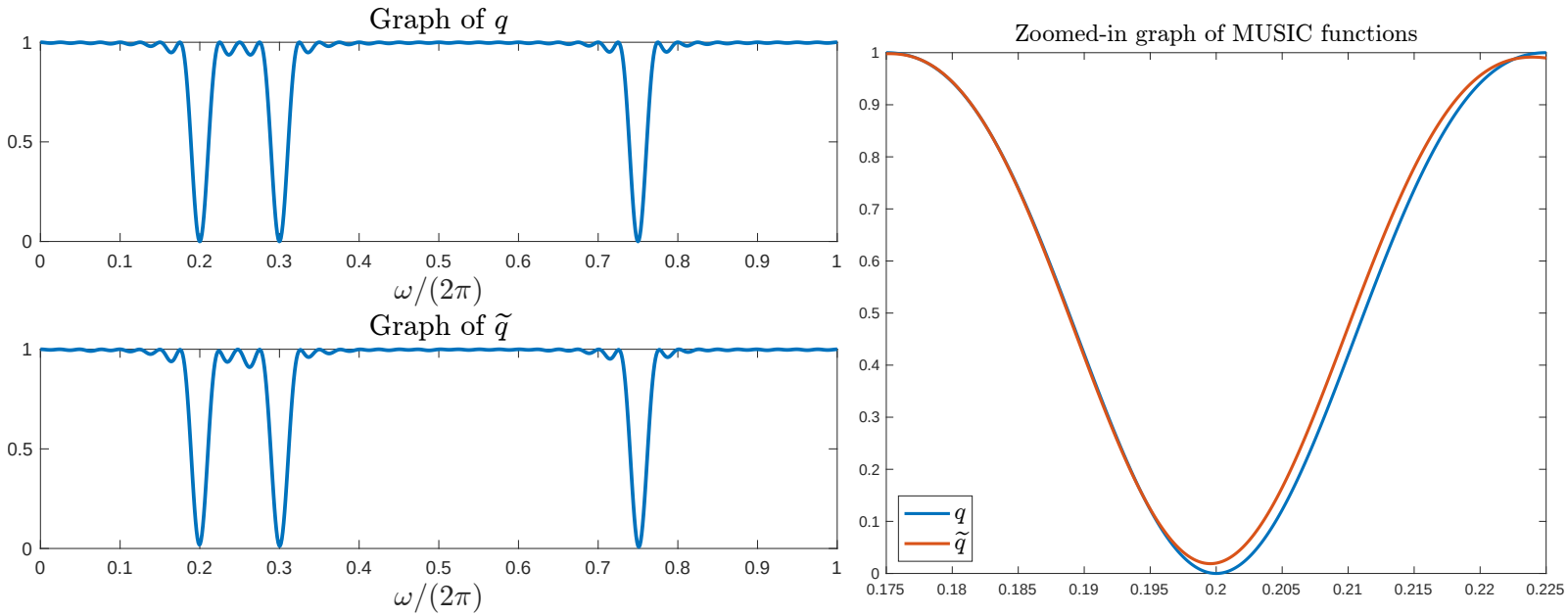
$$\widetilde{q}(\omega) := \left\| (I - P_{\widetilde{U}})\phi_\omega \right\|_2^2, \quad \text{where} \quad \phi_\omega = \frac{1}{\sqrt{m}} \left[e^{ik\omega} \right]_{k=0,\ldots,m-1}.$$

3. Find the s smallest discrete local minima of $\widetilde{q}$, denoted $\{\widetilde{\theta}_j\}_{j=1}^s$, via a fine grid search.

Plot of MUSIC function

Heuristic property of MUSIC function: The noiseless MUSIC function q has $\{\theta_j\}_{j=1}^s$ as its zero set and it appears that $\widetilde{\theta}_j \approx \theta_j$.

Classical MUSIC is expensive: Evaluation of $\widetilde{q}$ at a single $\omega \in \mathbb{T}$ requires $O(ms)$ operations. Need to evaluate $\widetilde{q}$ on a fine grid.



(a) Plots of q and $\widetilde{q}$.

(b) Zoomed in plot.

Figure: Let $m = 40$, $\theta_1 = 2\pi(0.2)$, $\theta_2 = 2\pi(0.3)$, and $\theta_3 = 2\pi(0.75)$. Graphs of the MUSIC functions q and $\widetilde{q}$ where i.i.d Gaussian noise was used.

Gradient-MUSIC methodology

Classical MUSIC: evaluates $\widetilde{q}$ on a *thin* grid to find minima $\{\widetilde{\theta}_j\}_{j=1}^s$.
Gradient-MUSIC: evaluates $\widetilde{q}$ on a *coarse* grid, thresholds to find initialization, and runs gradient descent to find minima $\{\widetilde{\theta}_j\}_{j=1}^s$.

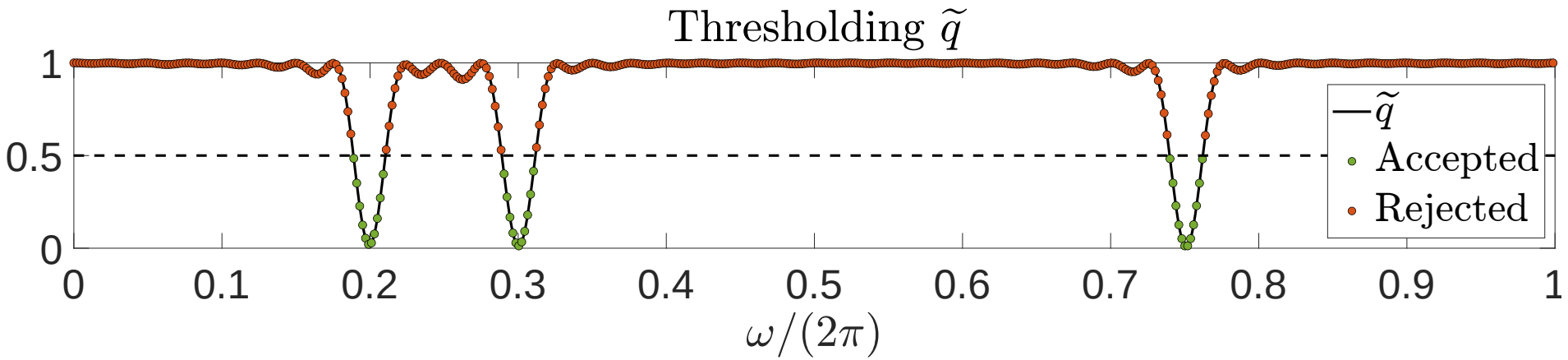


Figure: Graph of $\widetilde{q}$, its values on a uniform grid of width $1/(2m)$, and the set of accepted and rejected points are shown in green and red, respectively.

Comparison of computational complexity

Classical MUSIC

- Suffers from discretization due to use of grid.
- High accuracy $\rightarrow$ fine grid $\rightarrow$ computationally expensive.
- To get error ε/m, cost is $O(sm\varepsilon^{-1}\log(m/\varepsilon))$ via FFT.

Gradient-MUSIC

- Does not suffer from discretization.
- Coarse grid has width $1/(2m)$, reduces cost of evaluation, while gradient descent converges at a linear rate.
- To get error ε/m, cost is $O\left(sm\log(m) + s^2 m\log(1/\varepsilon)\right)$ via FFT.

Matlab `svds`	Gradient-MUSIC	Classical MUSIC
0.1505	0.5030	105.4096

Table: Runtime of each function in seconds. Here $m = 1000$ and the max time is reported over 10 trials.

Main challenges with MUSIC-type algorithms

Approximation and stability (pertains to both classical and Gradient- MUSIC)

- Why does the MUSIC function even have at least s minima?
- Why choose the smallest s local minima of $\widetilde{q}$, as opposed to other ones?
- Do $\{\widetilde{\theta}_j\}_{j=1}^s$ even approximate $\{\theta_j\}_{j=1}^s$?

Numerical and computational (pertains to Gradient- MUSIC)

- Why does thresholding on a coarse grid find suitable initialization?
- Why does gradient descent converge and at what rate?
- How to pick parameters such as number of iterations and step size?

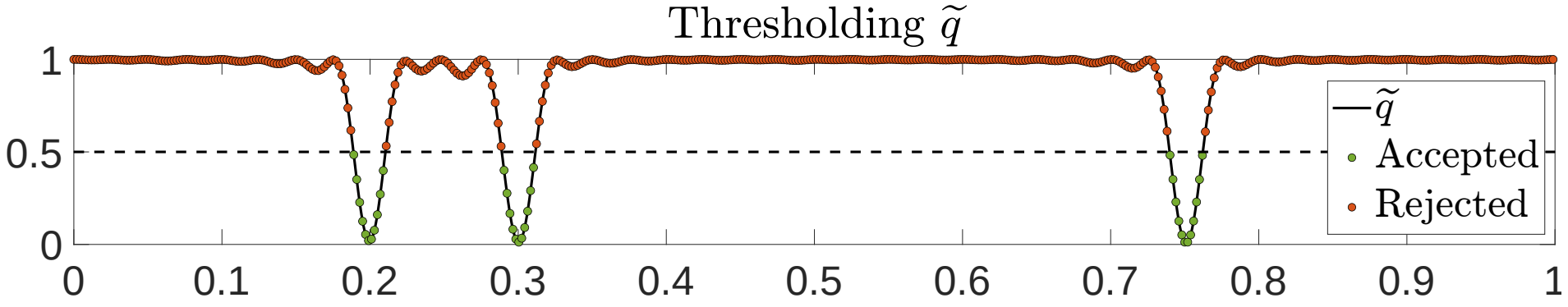
Remark: Prior results [Liao, Fannjiang, 2016] and [L., Liao 2021] analyze stability of MUSIC function, which does not imply anything about stability of its local minima. For example, $f(x) = \varepsilon^2|x|$ and $g(x) = \varepsilon^2|x - \varepsilon^{-1}|$ satisfy $\|f - g\|_{L^\infty(\mathbb{R})} \leq \varepsilon$, yet their global (and only) minima are ε^{-1} apart.

Geometric analysis of the MUSIC function

Theorem (Fannjiang, L., Liao, 2025)

For any $m \geq 100$, $\{\theta_j\}_{j=1}^s$, *and subspace* $\widetilde{U}$ *such that* $m\Delta \geq 8\pi$ *and* $\delta := \|P_U - P_{\widetilde{U}}\| \leq 0.01$, *the following hold.*

1. *(Desirable minima are close to true parameters)* $\max_j |\widetilde{\theta}_j - \theta_j| \leq \frac{7\delta}{m}$.
2. *(Local convexity)* $0.0271\, m^2 \leq \widetilde{q}''(t) \leq 0.269\, m^2$ *whenever* $|t - \widetilde{\theta}_j| \leq \frac{\pi}{3m}$.
3. *(Derivative control)* $\widetilde{q}'(t) \geq 0.0306\, m$ *for all* $t \in [\widetilde{\theta}_j + \frac{\pi}{3m}, \widetilde{\theta}_j + \frac{4\pi}{3m}]$ *and* $\widetilde{q}'(t) \leq -0.0306\, m$ *for all* $t \in [\widetilde{\theta}_j - \frac{4\pi}{3m}, \widetilde{\theta}_j - \frac{\pi}{3m}]$.
4. *(Desirable minima are deep)* $\max_j \widetilde{q}(\widetilde{\theta}_j) \leq \delta^2$.
5. *(Undesirable minima are shallow)* $\widetilde{q}(t) \geq 0.529$ *if* $|t - \widetilde{\theta}_j| \geq \frac{4\pi}{3m}$ *for all* j.



Glimpse into proof

Let W_j be $s-1$ dimensional subspace of U orthogonal to $\phi(\theta_j)$ and

$$d_m(\omega) = \frac{1}{m}\frac{\sin(m\omega/2)}{\sin(\omega/2)}.$$

Proof centers around using the formula: for each j and ω,

$$q(\omega) = 1 - |d_m(\omega - \theta_j)|^2 + \|P_{W_j}\phi_\omega\|_2^2.$$

When $m\Delta \geq 8\pi$ and $|\omega - \theta_j| \lesssim \frac{1}{m}$, the $1 - |d_m(\omega - \theta_j)|^2$ term is dominant.

Glimpse into proof

Let W_j be $s-1$ dimensional subspace of U orthogonal to $\phi(\theta_j)$ and

$$d_m(\omega) = \frac{1}{m}\frac{\sin(m\omega/2)}{\sin(\omega/2)}.$$

Proof centers around using the formula: for each j and ω,

$$q(\omega) = 1 - |d_m(\omega - \theta_j)|^2 + \|P_{W_j}\phi_\omega\|_2^2.$$

When $m\Delta \geq 8\pi$ and $|\omega - \theta_j| \lesssim \frac{1}{m}$, the $1 - |d_m(\omega - \theta_j)|^2$ term is dominant.

Other steps:

- Have to do local analysis for q', q'', and q''' as well.
- Have to lower bound $q(\omega)$ for ω far away from each θ_j.
- Transfer results to $\widetilde{q}$ through perturbation arguments.
- Intermediate value (or fixed point) theorem to establish existence of $\{\widetilde{\theta}_j\}_{j=1}^s$.
- Argue that $\{\widetilde{\theta}_j\}_{j=1}^s$ has the claimed properties.

Comparison to traditional nonconvex optimization

1. **Nonlinear least squares and maximum likelihood estimation**
   There are $2s$ many unknown parameters $\{(\theta_j, a_j)\}_{j=1}^{s}$, so NLS and MLE would create an objective function of $2s$ variables and find the best ones that minimize an error functional.

   [Traonmilin et. al., 2023 and 2024] showed that basin of attraction to global minimum for NLS (with random Fourier samples) is like a ball with radius Δ.

   As s increases $\rightarrow$ higher dimensional $\rightarrow$ volume of balls shrink $\rightarrow$ basin of attraction to global minimum shrinks.

Comparison to traditional nonconvex optimization

1. **Nonlinear least squares and maximum likelihood estimation**
   There are $2s$ many unknown parameters $\{(\theta_j, a_j)\}_{j=1}^{s}$, so NLS and MLE would create an objective function of $2s$ variables and find the best ones that minimize an error functional.

   [Traonmilin et. al., 2023 and 2024] showed that basin of attraction to global minimum for NLS (with random Fourier samples) is like a ball with radius Δ.

   As s increases $\rightarrow$ higher dimensional $\rightarrow$ volume of balls shrink $\rightarrow$ basin of attraction to global minimum shrinks.

2. **Classical and Gradient- MUSIC**
   MUSIC function $\widetilde{q}$: $\mathbb{T} \rightarrow \mathbb{R}$ is always a single variable function regardless of s.

   Geometric analysis shows that basin of attraction to smallest s local minima are intervals of size $\frac{1}{m}$. Also shows that $\widetilde{q}$ is large away from basins.

   Initialization can be found by evaluating $\widetilde{q}$ on a grid with mesh norm $\asymp \frac{1}{m}$ and thresholding. Hence initialization can be found easily and efficiently!

Outline

Multidimensional Spectral estimation

Definition (Spectral estimation)

Estimate the parameters $\{(\theta_j, a_j)\}_{j=1}^s$ given data

$$\widetilde{y}(x) = \sum_{j=1}^{s} a_j e^{i\theta_j \cdot x} + \eta(x) \quad \text{for all} \quad x \in X_\star.$$

Terminology:

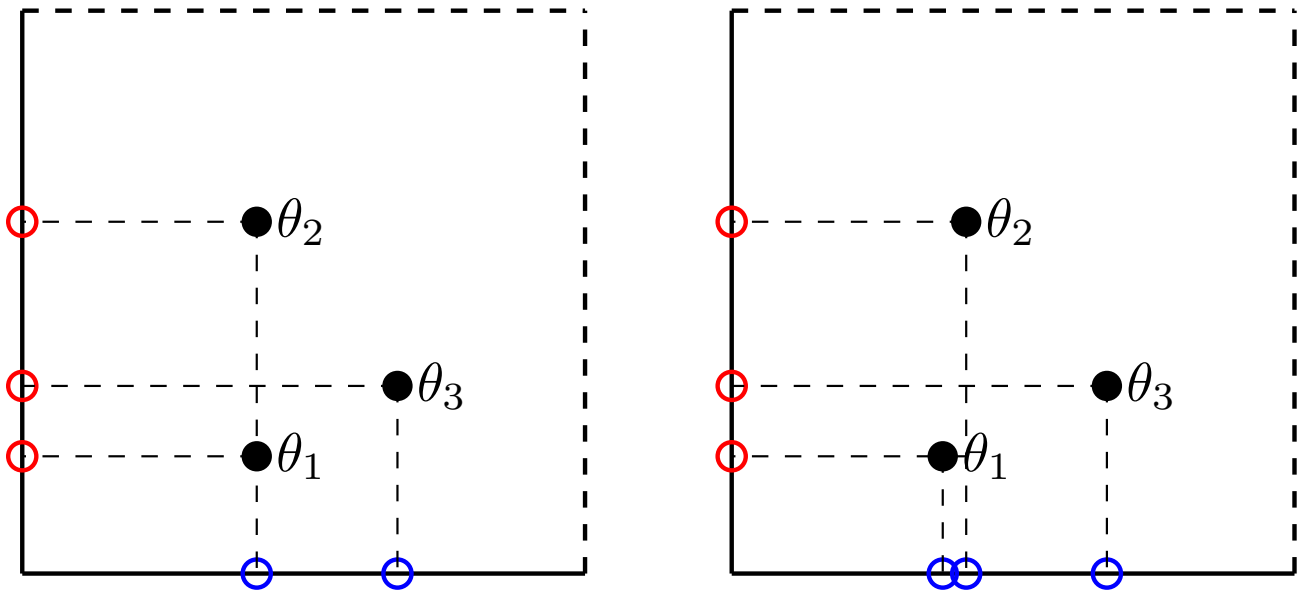
- "Frequency" $\theta_j \in \mathbb{R}^d$
- "Amplitude" $a_j \in \mathbb{C}$ such that $|a_j| \geq 1$
- "Sparsity" s
- "Noise" η which can be deterministic or random
- "Sampling set" $X_\star \subseteq \mathbb{R}^d$

A common roadblock in higher dimensions

Projection based methods: If $X_\star$ is a continuous set such as a ball, if $x = tu$ for a direction $u \in \mathbb{S}^{d-1}$, then

$$\sum_{j=1}^{s} a_j e^{i\theta_j \cdot x} = \sum_{j=1}^{s} a_j e^{it(\theta_j \cdot u)}.$$

Can interpret as one-dimensional samples for new set of 1D nodes $\{(\theta_j \cdot u)\}_{j=1}^{s}$.



Some multidimensional algorithms

- **Projection based methods.** Project onto given directions [Cuyt et. al., 2020], [Mhaskar, Kitimoon, Raj., 2025], or use random directions and reject bad ones, [Chen, Moitra, 2021].
- **Multidimensional ESPRIT**. Creates at least d shift directions and solves an approximate simultaneous diagonalization problem. Implicitly performs projections, and suffers from similar difficulties as projection methods. Various ways of addressing this issue: [Haardt, Roemer, Del Galdo, 2008], [Sahnoun, Usevich, Comon, 2017], [Andersson, Carlsson, 2018], and more.
- **Multidimensional MUSIC**. Algorithm works in the same way as 1D case, but requires a fine grid search in d-dimensions, [Liao, 2015].
- **Multidimensional Prony's method.** [Kunis et. al., 2016] and [Sauer, 2017].
- **Convex methods.** [De Castro et. al., 2016], [Poon, Keriven, Peyré, 2023].

Multi-dimensional Gradient-MUSIC

Works the same way in higher dimensions. Compute a MUSIC function $\widetilde{q}\colon \mathbb{T}^d \to [0,1]$, evaluate it on a coarse grid to find good initialization, and use gradient descent.

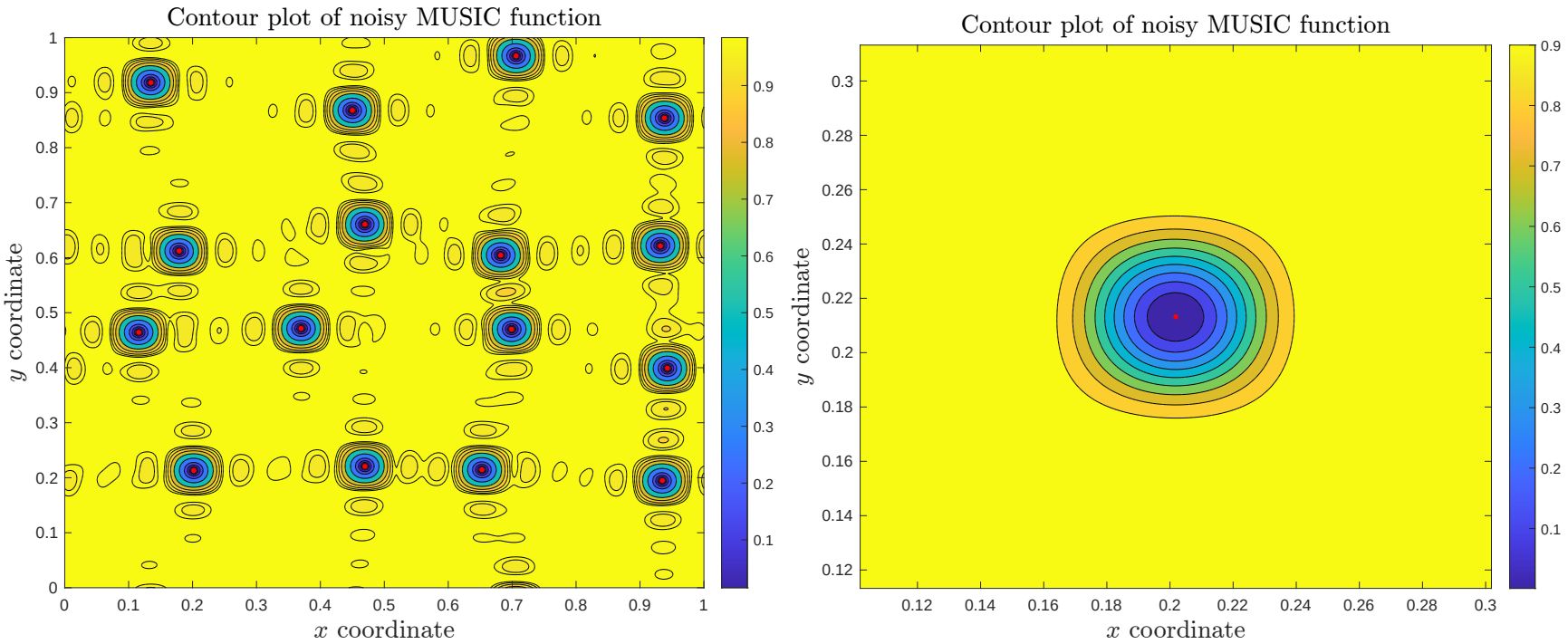


Figure: Example of a MUSIC function. Red dots represent the true parameters $\{\theta_\ell\}_{\ell=1}^{16}$ that are separated by at least $1/8$ and $\{a_\ell\}_{\ell=1}^{16}$ is a randomly generated sequence of ± 1. Samples were collected on $X_\star = \{-20, \ldots, 20\}^2$, and corrupted by i.i.d. normally distributed noise with mean zero and variance one. The smallest local minima of the MUSIC function closely approximate the true $\{\theta_\ell\}_{\ell=1}^{16}$.

Main result for samples in a cube

Theorem (Fannjiang, L., 2026, simplified version)

Let $d \geq 2$, $m \geq 1$, and $X_\star = Q_{2m} \cap \mathbb{Z}^d$.
For any $\{\theta_\ell\}_{\ell=1}^s \subseteq \mathbb{T}^d$ and $\{a_j\}_{j=1}^s$ such that $\|a\|_\infty \lesssim 1$ and $m\Delta \gtrsim_d 1$:

- *If $\|\eta\|_p \lesssim_d m^{d/p}$ for some $p \in [1, \infty]$, then Gradient-MUSIC produces iterates which converge at a linear rate to $\{\widetilde{\theta}_\ell\}_{\ell=1}^s$ such that*

$$\max_\ell \left|\theta_\ell - \widetilde{\theta}_\ell\right| \lesssim_d \frac{\|\eta\|_p}{m^{1+d/p}}.$$

Main result for samples in a cube

Theorem (Fannjiang, L., 2026, simplified version)

Let $d \geq 2$, $m \geq 1$, and $X_\star = Q_{2m} \cap \mathbb{Z}^d$.
For any $\{\theta_\ell\}_{\ell=1}^s \subseteq \mathbb{T}^d$ and $\{a_j\}_{j=1}^s$ such that $\|a\|_\infty \lesssim 1$ and $m\Delta \gtrsim_d 1$:

- *If $\|\eta\|_p \lesssim_d m^{d/p}$ for some $p \in [1, \infty]$, then Gradient-MUSIC produces iterates which converge at a linear rate to $\{\widetilde{\theta}_\ell\}_{\ell=1}^s$ such that*

$$\max_\ell \left|\theta_\ell - \widetilde{\theta}_\ell\right| \lesssim_d \frac{\|\eta\|_p}{m^{1+d/p}}.$$

- *Let $\eta(x) = g(x)(1+|x|)^r$ for $x \in X_\star$, where the values of g are i.i.d. $\mathcal{N}(0, \sigma^2)$ random variables and $r \in (-d/2, d/2)$. With probability at least $1 - o(m)$, Gradient-MUSIC produces iterates which converge at a linear rate to $\{\widetilde{\theta}_\ell\}_{\ell=1}^s$ such that*

$$\max_\ell \left|\theta_\ell - \widetilde{\theta}_\ell\right| \lesssim_d \frac{\sigma\sqrt{\log(m)}}{m^{d/2-r+1}}.$$

Computational complexity of classical vs Gradient- MUSIC

For the sampling set $X_\star = Q_{2m} \cap \mathbb{Z}^d$, which has $(4m+1)^d$ elements.

Classical MUSIC

- Suffers from discretization due to use of grid.
- High accuracy $\rightarrow$ fine grid $\rightarrow$ computationally expensive.
- Bad scaling in dimension: need $(m/\varepsilon)^d$ grid points to get mesh ε/m.
- To get accuracy ε/m, cost is $O_d(sm^d\varepsilon^{-d}\log(m/\varepsilon))$ via FFT.

Gradient-MUSIC

- Does not suffer from discretization.
- Coarse grid has width $\asymp_d 1/m$, reduces cost of evaluation, while gradient descent converges at a linear rate.
- Need $O_d(m^d)$ grid points to get mesh $\asymp_d 1/m$.
- To get accuracy ε/m, cost is $O_d\left(sm^d\log(m) + s^2m^d\log(1/\varepsilon)\right)$ via FFT.

As a computational framework

Canonical extension and generalization to

- Arbitrary sampling set $X_\star \subseteq \mathbb{R}^d$, can be continuous or discrete.
- Arbitrary domain $\Omega \subseteq \mathbb{R}^d$ such that $\theta_j \in \Omega$.
- Arbitrary noise η.

1. (Importance sampling). Pick a subset $X \subseteq X_\star$ and measure ν on X.
2. (Denoising). Find a projection operator $P_{\widetilde{U}}$ which approximates the true projection P_U onto the span of $\{e^{i\theta_j \cdot x}\}_{j=1}^s$ on $L^2_\nu(X)$.
3. (MUSIC function). Define the *MUSIC function* $\widetilde{q}\colon \Omega \to [0,1]$ by

$$\widetilde{q}(\omega) := \left\|(I - P_{\widetilde{U}})\phi_\omega\right\|_2^2, \quad \text{where} \quad \phi_\omega(x) = \frac{1}{\sqrt{\nu(X)}} e^{ix\cdot\omega} \in L^2_\nu(X).$$

4. (Initialization and optimization) Find the s smallest discrete local minima of $q_{\widetilde{U}}$, denoted $\{\widetilde{\theta}_j\}_{j=1}^s$, via evaluation on a coarse grid and gradient descent.

An associated measurement kernel

Measurement kernel:

$$K(\omega, \omega') = \int_X \overline{\phi_\omega(x)} \phi_{\omega'}(x)\, d\nu(x) = \frac{1}{\nu(X)} \int_X e^{ix\cdot(\omega'-\omega)}\, d\nu(x).$$

Meta-theorem: If K satisfies certain properties, then Gradient-MUSIC provably succeeds and

$$\max_\ell |\theta_\ell - \widetilde{\theta}_\ell| \lesssim_d \frac{\sqrt{\mathsf{trace}(H)}}{\lambda_d(H)} \|P_U - P_{\widetilde{U}}\|, \quad \text{where} \quad H = -\nabla^2 K(0).$$

An associated measurement kernel

Measurement kernel:

$$K(\omega,\omega') = \int_X \overline{\phi_\omega(x)}\phi_{\omega'}(x)\,d\nu(x) = \frac{1}{\nu(X)}\int_X e^{ix\cdot(\omega'-\omega)}\,d\nu(x).$$

Meta-theorem: If K satisfies certain properties, then Gradient-MUSIC provably succeeds and

$$\max_\ell |\theta_\ell - \widetilde{\theta}_\ell| \lesssim_d \frac{\sqrt{\mathsf{trace}(H)}}{\lambda_d(H)}\|P_U - P_{\widetilde{U}}\|, \quad \text{where} \quad H = -\nabla^2 K(0).$$

Examples of the kernel K:

- Dirichlet kernel $d_m(\omega) = \dfrac{1}{2m+1}\dfrac{\sin(\pi\omega/2)}{\sin(\omega/2)}$ if $X = \{-m,\dots,m\}$.
- Square Dirichlet kernel $D_m = d_m \otimes \cdots \otimes d_m$ if $X = \{-m,\dots,m\}^d$.
- Weighted Bessel kernel $\dfrac{1}{|B_1|}\dfrac{J_{d/2}(2\pi m|\omega|)}{(m|\omega|)^{d/2}}$ if $X = B_m$.
- Difference of weighted Bessel kernels if $X = B_m \setminus B_r$.
- Random Fourier features kernel for randomly chosen X w.r.t. to an absolutely continuous ν on B_m.

Outline

Summary of talk

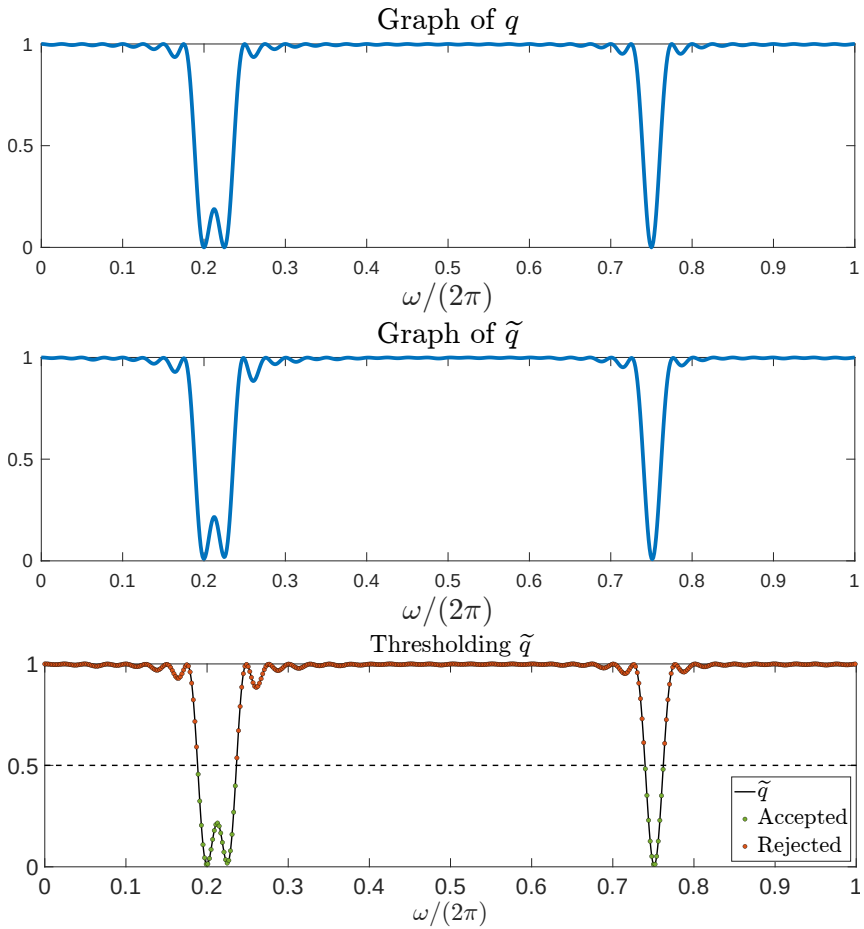
1. Gradient-MUSIC is optimal for spectral estimation when $m\Delta \geq 8\pi$ and under natural deterministic and random noise models.
2. It is a rigorous nonconvex optimization algorithm that is justified by exploiting special properties of the MUSIC function as an optimization landscape.
3. Gradient-MUSIC has a canonical extension to higher dimensions and uses the same mechanisms. There is an unifying kernel structure.
4. It is efficient finding since initialization only requires a coarse grid while gradient iterations are cheap, even in multiple dimensions.
5. Classical and Gradient- MUSIC output the same results, but the latter is always more efficient. The gap widens in higher dimensions.
6. Gradient-MUSIC is flexible, e.g., compatible with general sampling sets, randomized numerical linear algebra, and more sophisticated nonconvex optimization methods.

References

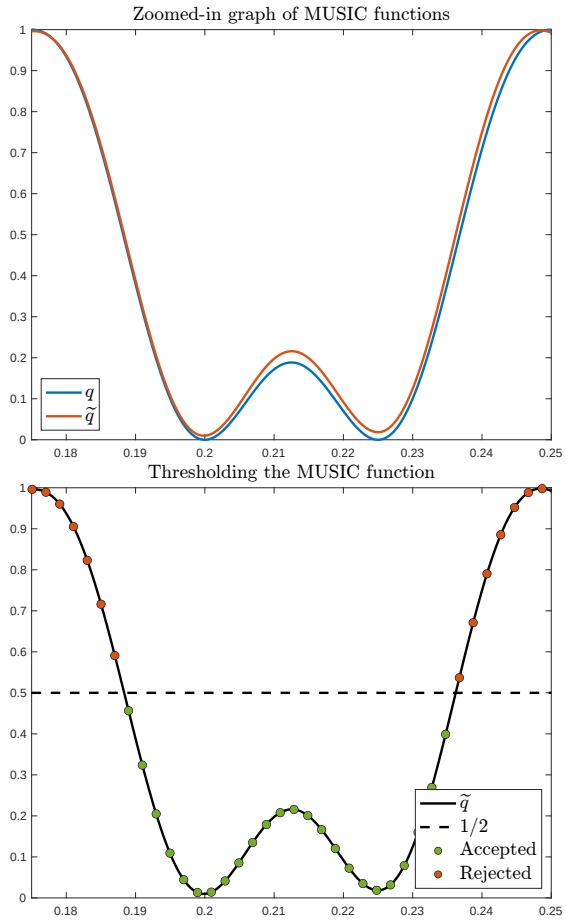
1. **One-dimensional setting**
   Albert Fannjiang, Weilin Li, and Wenjing Liao. Optimality of Gradient-MUSIC for Spectral Estimation. 60 pages, preprint, arXiv:2504.06842
2. **Multi-dimensional setting**
   Albert Fannjiang and Weilin Li. Multidimensional Gradient-MUSIC: A Global Nonconvex Optimization Framework for Optimal Resolution. https://arxiv.org/abs/2603.27379
3. **Application to structured approximation**
   Albert Fannjiang and Weilin Li. Structured Approximation of Toeplitz Matrices and Subspaces. 21 pages, preprint, arXiv:2511.17239

Thank you!

Gradient-MUSIC for super-resolution



(a) Plots of q and $\widetilde{q}$.



(b) Zoomed in plot.

A quick word on amplitudes

Approximate amplitudes $\{\widetilde{a}_j\}_{j=1}^s$ can be found by solving a system of equations. We proved that

$$\max_j |a_j - \widetilde{a}_j| \lesssim \sqrt{s}\, m \max_j |\theta_j - \widetilde{\theta}_j|.$$

Examples for $d = 1$:

- For $\|\eta\|_\infty \lesssim 1$, we get

$$\max_j |a_j - \widetilde{a}_j| \lesssim \sqrt{s}\, \|\eta\|_\infty.$$

- For $\eta \sim \mathcal{N}(0, \sigma^2 I)$, we get

$$\max_j |a_j - \widetilde{a}_j| \lesssim \sigma \sqrt{\frac{s \log(m)}{m}}.$$

Natural question: Are there situations (larger noise models) where frequencies $\{\theta_j\}_{j=1}^s$ can be approximated but not amplitudes $\{a_j\}_{j=1}^s$?