February 2024

POPILS seminar, INSA Lyon

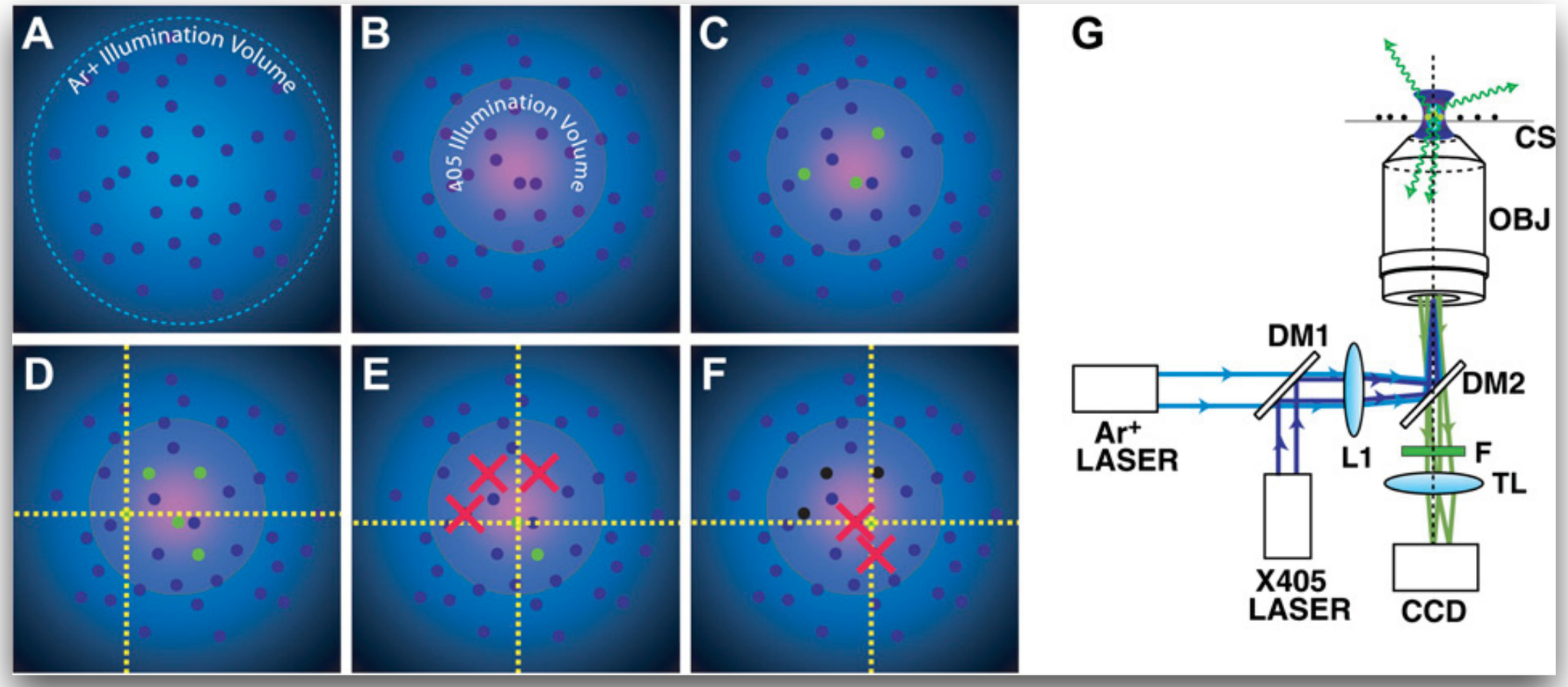
Convex regularization on measures for GMMs

Based on joint works with C. Marteau (ICJ), S. Gadat (TSE), R. Gribonval (INRIA, ENSL),
J.M. Azaïs & F. Gamboa (Toulouse 3), C. Maugis (INSA Toulouse),
D. Henrion & J.B. Lasserre (LAAS), N. Jouvin (INRAE),
C. Boyer (P6), J. Salmon (Univ. Montpellier)

Super-Resolution

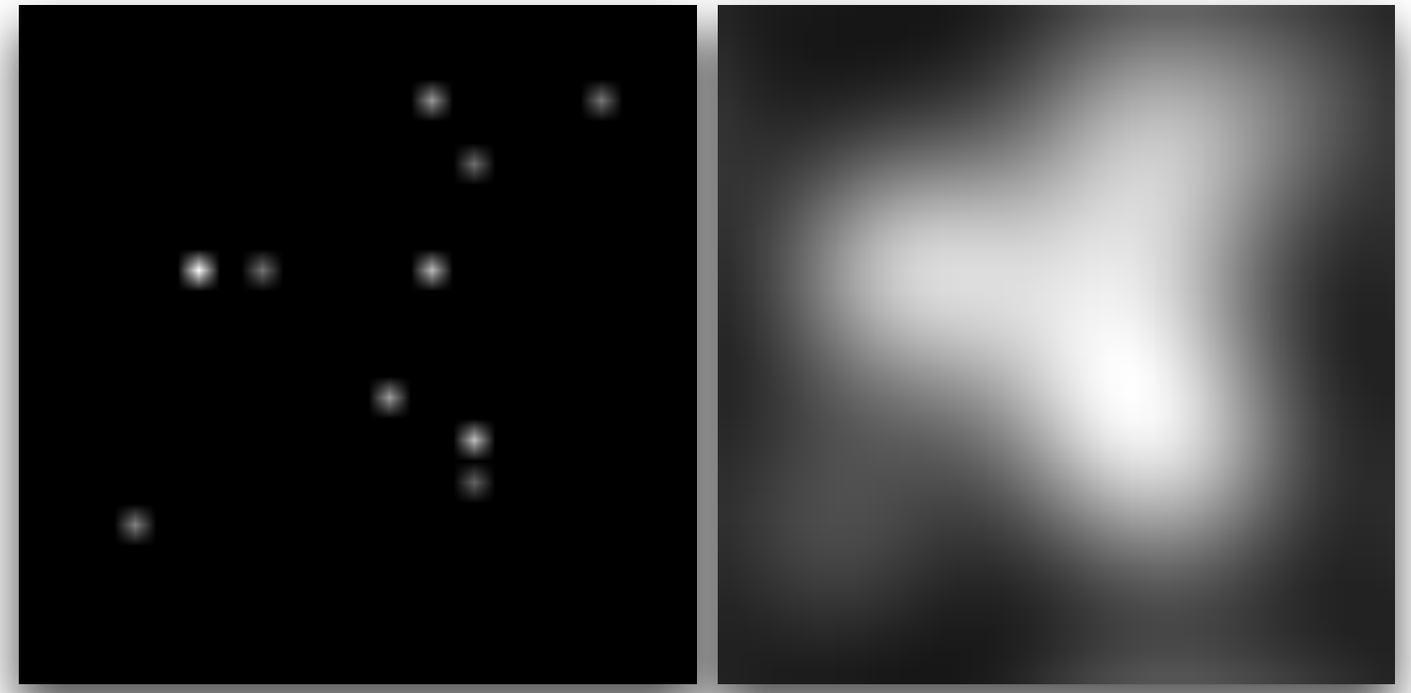
POPILS 2024

Super-resolution: Fluorescence microscopy



- S. Hess, T. Girirajan, M. Mason, Ultra-High Resolution Imaging by Fluorescence Photoactivation Localization Microscopy, *Biophysical Journal* (2004).

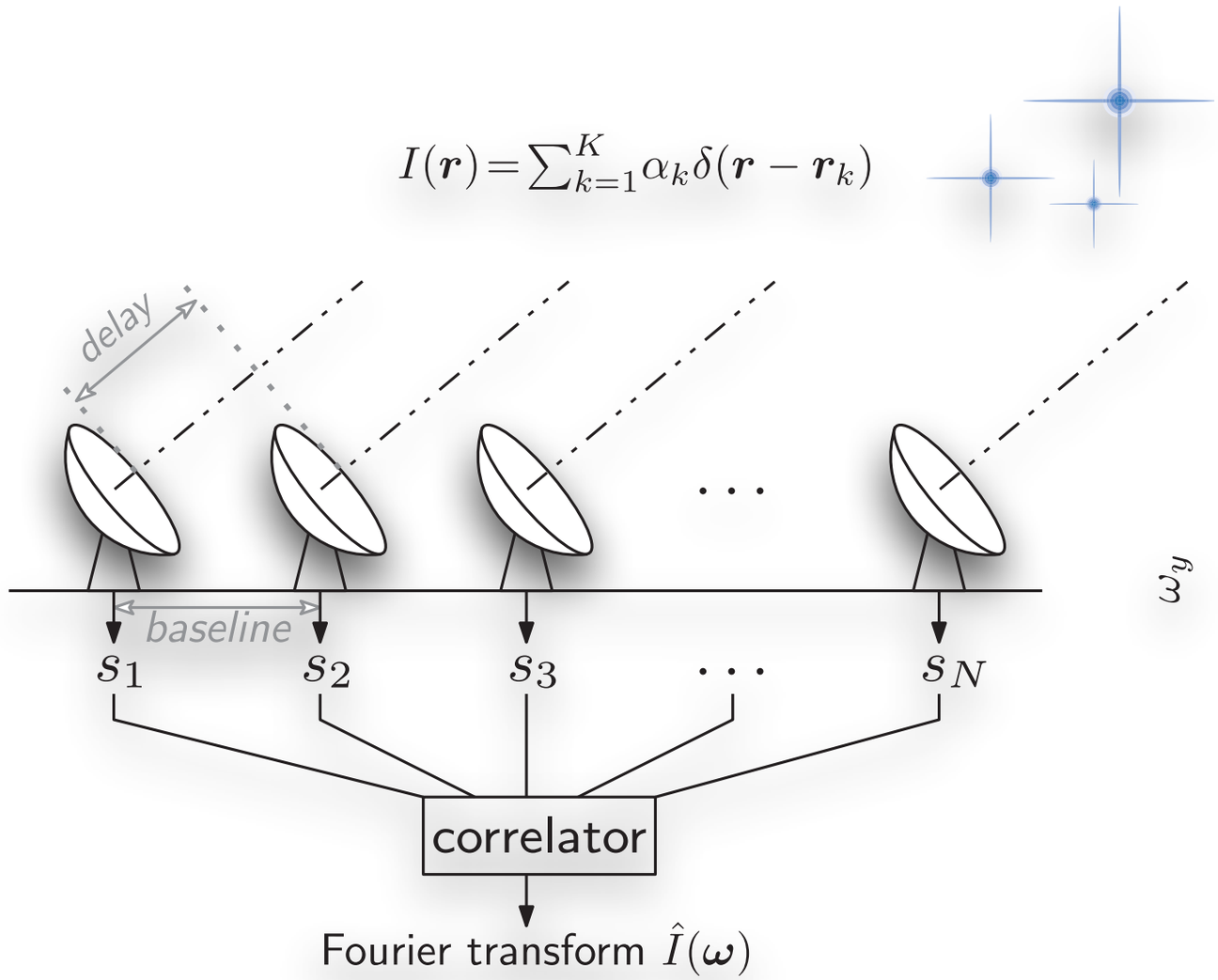
Super-Resolution: Heat source localization



(a) Heat source u_0 (b) $\mathrm{A}u_0 = f$

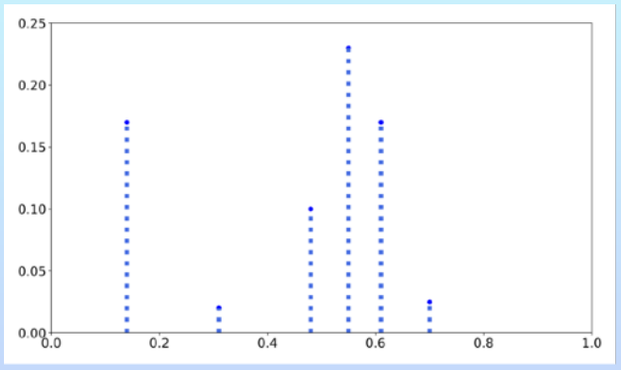
- Y. Li, S. Osher, R. Tsai, Heat Source Identification based on L1 Constrained Minimization, *Inverse Problems and Imaging* (2014).

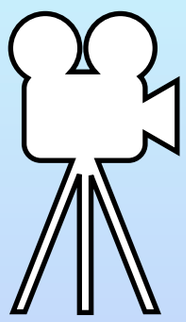
Super-Resolution: Radioastronomy

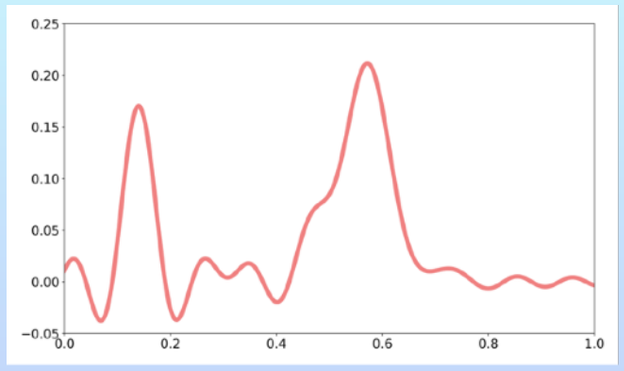


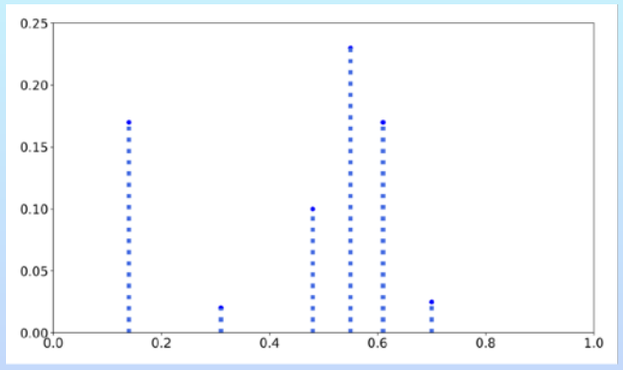
- H. Pan, T. Blu, M. Vetterli, Towards Generalized FRI Sampling With an Application to Source Resolution in Radioastronomy, *IEEE trans. on Signal Processing* (2017).

L1 recovery









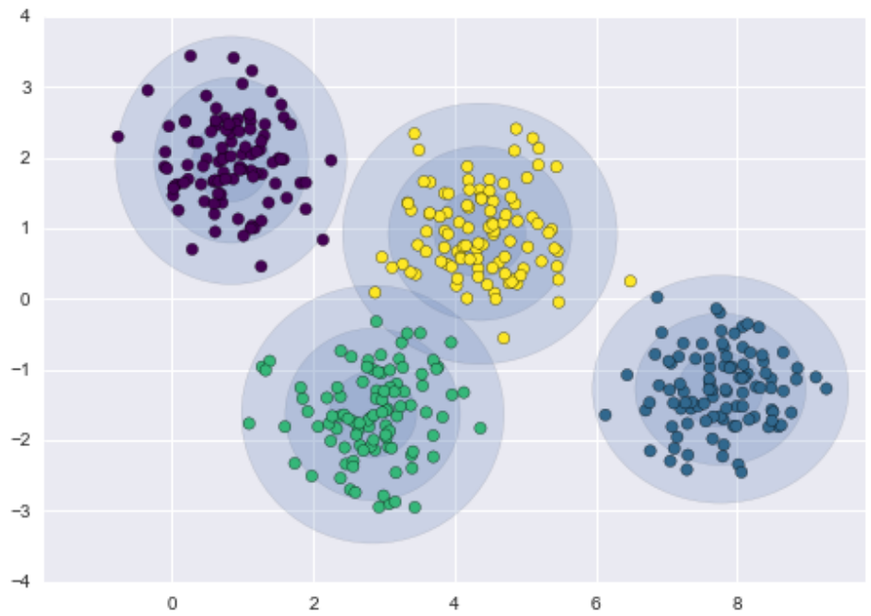
Discrete Measure

Perfect Recovery

$\mu^\star$ $\quad \Phi \quad$ $\mathbf{Y} = \Phi(\mu^\star)$ $\quad \min_{\mu\,:\,\Phi(\mu)=\mathbf{Y}} \|\mu\|_1$

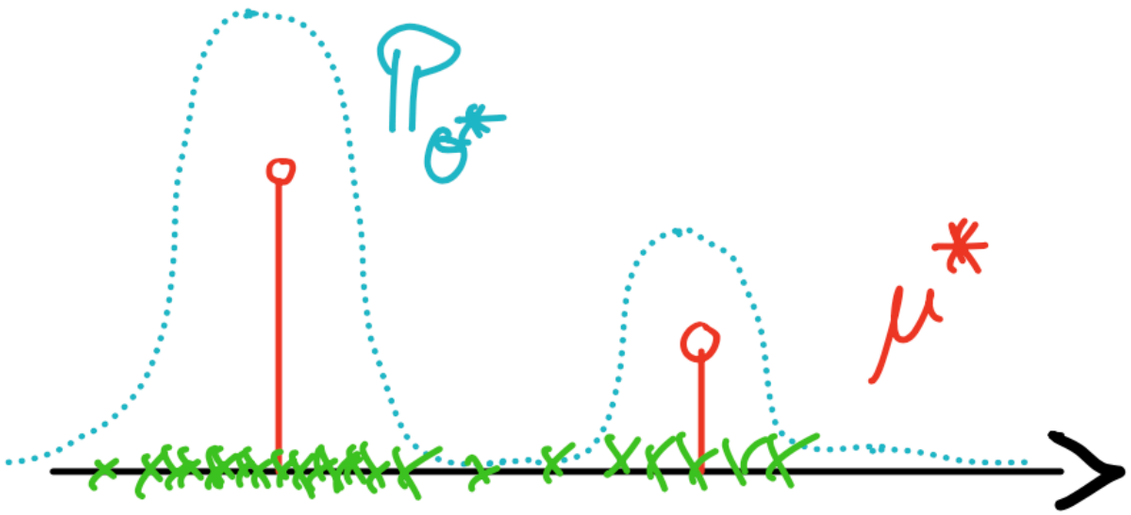
Gaussian Mixture Models

Gaussian Mixture Models (GMMs)



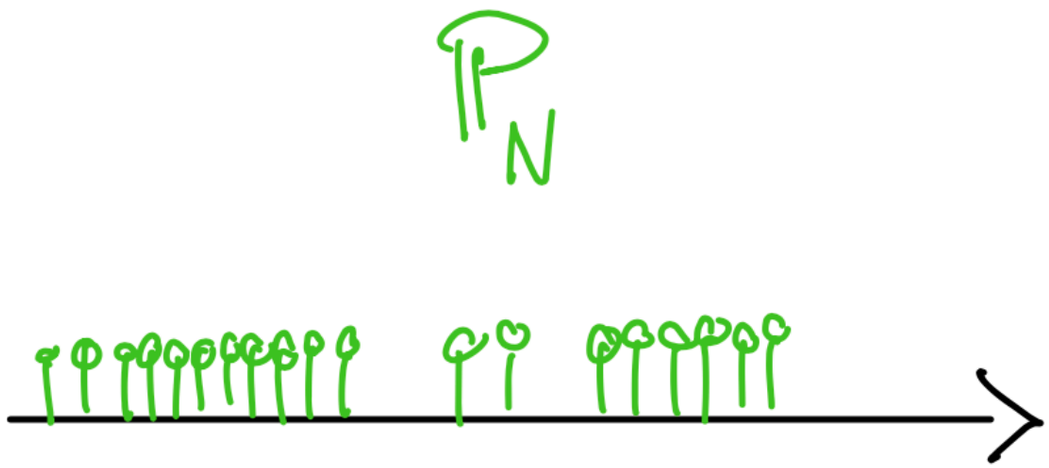
- Data: $X_1, \ldots, X_N \sim \mathbb{P}_\theta$
- Mixture model: $\frac{\mathrm{d}\mathbb{P}_\theta}{\mathrm{dLeb}} = \sum_{j=1}^{s} w_i \psi_{t_i}$
- Gaussian density: $\frac{\mathrm{d}\mathcal{N}(t,\mathrm{Id}_d)}{\mathrm{dLeb}} =: \psi_t$
- Parameters: $\theta = \left\{ s;\ w_1, \ldots w_s \in \mathbb{R};\ t_1, \ldots t_s \in \mathcal{X} \right\}$
- Assumption: $\mathcal{X}$ compact set of $\mathbb{R}^d$

Expectation-Maximization (EM)



- MLE is a **non-convex** program
- Usually solved by **EM**
- **s** is assumed to be known
- Convergence of EM to global maximum can be tedious

- Empirical measure: $\mathbb{P}_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{X_i}$
- Target measure: $\mu^\star = \sum_{j=1}^{s^\star} w_j^\star \delta_{t_j^\star}$

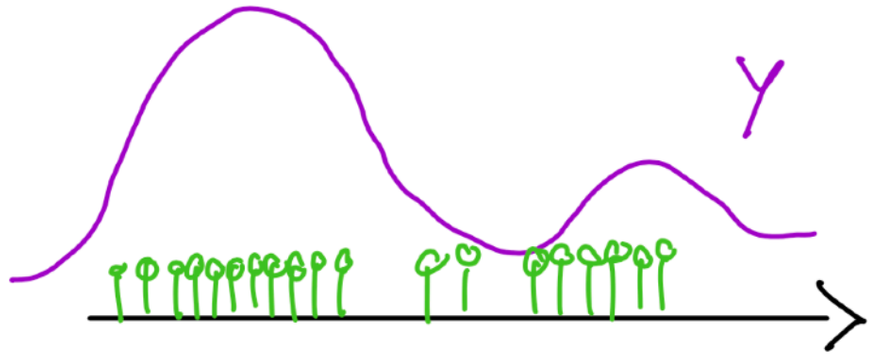


- Inverse Problem: convolution + sampling

$$\mu^\star \longrightarrow \frac{\mathrm{d}\mathbb{P}_{\theta^\star}}{\mathrm{dLeb}} = \mu^\star \star \psi \longrightarrow \mathbb{P}_N$$

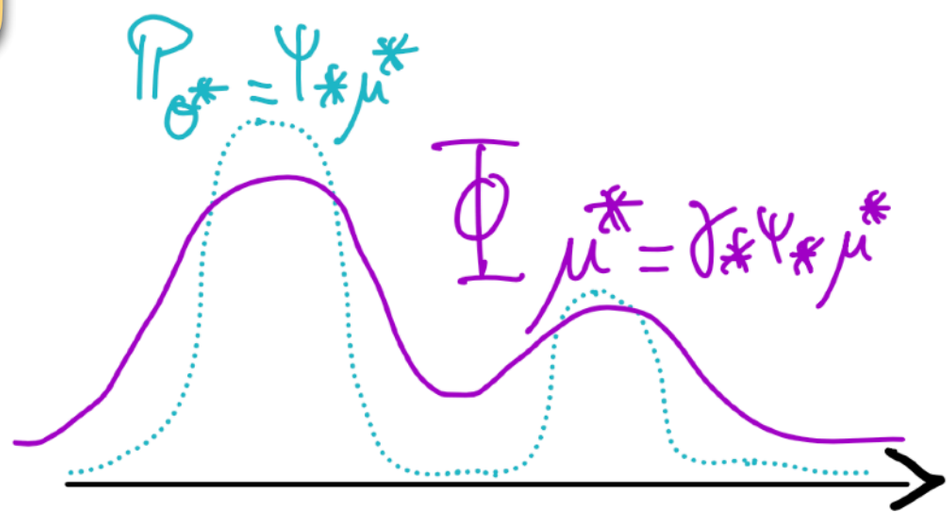
Key idea: kernel estimator

- Consider a kernel $\gamma(t-s)$ with bandwidth σ_γ
- In this talk, γ such that $\mathcal{F}\gamma(\omega) = \mathbb{1}_{\|\omega\|_\infty \leq m}$ with bandwidth $\sigma_\gamma = 1/m$
- Let $\mathbb{H}$ be the RKHS associated to γ
- Set $\varphi_t := \gamma \star \psi_t$ and $\Phi\mu = \displaystyle\int_{\mathcal{X}} \varphi_t \mathrm{d}\mu(t)$



- Kernel embedding
  - Data embedding: $Y = \gamma \star \mathbb{P}_N$
  - Measure embedding: $\Phi\mu$

both in $\mathbb{H}$



- Check that $\mathbb{E}Y = \Phi\mu^\star$
- BLASSO: $\min_\mu \left\{ \frac{1}{2}\|Y - \Phi\mu\|_{\mathbb{H}}^2 + \lambda\|\mu\|_1 \right\}$

Functional analysis

- Observation: $\mathbf{Y} \in \mathbb{H}$ with $\mathbb{H}$ **separable** Hilbert space;

- Measures: $(\mathcal{X}, \mathbf{d}_{\mathcal{X}})$ compact metric space, $(\mathcal{M}(\mathcal{X}), \|\cdot\|_1)$ space of measures with bounded TV norm;

$$(\mathcal{M}(\mathcal{X}), \|\cdot\|_1) = (\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)^\star$$

$$\langle \mu, f \rangle_{\mathcal{M}(\mathcal{X}),\mathcal{C}(\mathcal{X})} = \int_{\mathcal{X}} f \mathrm{d}\mu$$

- Features: $\varphi : x \in \mathcal{X} \mapsto \varphi_x \in \mathbb{H}$ **Assumption:** φ is continuous

- $\forall h \in \mathbb{H},\ \forall \mu \in \mathcal{M}(\mathcal{X}), \Phi : \mu \in \mathcal{M}(\mathcal{X}) \mapsto \int_{\mathcal{X}} \varphi_x \mathrm{d}\mu(x)$

- Well-defined and bounded linear;

$$\langle \Phi\mu \,,\, h \rangle_{\mathbb{H}} = \int_{\mathcal{X}} \langle \varphi_x \,,\, h \rangle_{\mathbb{H}} \, \mathrm{d}\mu(x) = \langle \mu \,,\, \Phi^\star h \rangle_{\mathcal{M},\mathcal{C}}$$

- $\Phi^\star : (\mathbb{H}, \|\cdot\|_{\mathbb{H}}) \to \underline{(\mathcal{C}(\mathcal{X}), \|\cdot\|_\infty)}$ bounded linear;

Interlude: A general method

Remark

- Many instances of ML problems can be formulated as an inverse problem

$$\text{Recover } \mu^\star \text{ from } \mathbf{Y} \sim \mathbb{P}_{\mathbf{\Phi}\mu^\star}$$

where

$$\mathbf{\Phi} : \mu \in \mathcal{M} \mapsto \mathbb{H}$$

with $\mathbb{H}$ a Hilbert space and $\mathcal{M}$ the space of signed regular Borel measures on some metric space $\mathcal{X}$.

- The sought after target measure $\mu^\star$ can be described by few parameters

$$\mu^\star \simeq \sum_{k=1}^{K^\star} a_k^\star \delta_{x_k^\star}$$

namely, it is a discrete measure with $a_k^\star \in \mathbb{R}$ and $x_k^\star \in \mathcal{X}$.

Some inverse problems

Consider a metric space $(\mathcal{X}, \mathbf{d}_{\mathcal{X}})$ of *predictors* and a space $\mathcal{Y}$ of *observations*. Given n observations $\mathbf{Y}_n := \{y_1, \dots, y_n\} \in \mathcal{Y}^{\otimes n}$, we would like to recover a target $\mu^\star$ in $\mathcal{M}(\mathcal{X})$, the space of signed measures on $\mathcal{X}$, from these n observations.

Assume that the target is discrete $\mu^\star = \sum_{k=1}^{K^\star} a_k^\star \delta_{x_k^\star}$ and we have 3 scenarii:

$$\mathbf{Y}_n \text{ i.i.d } \mathbf{\Phi}\mu^\star, \qquad (S_1 : \text{ Sampling})$$

$$\mathbf{Y} = \mathbf{\Phi}\mu^\star, \quad (n = 1) \qquad (S_2 : \text{ Functional inference})$$

$$\mathbf{Y}_n = \mathbf{\Phi}_n\mu^\star + \mathbf{e}_n\,, \qquad (S_3 : \text{ Noisy linear measurements})$$

where $a_k \in \mathbb{R}$, $x_k^\star \in \mathcal{X}$, $\mathbf{\Phi}$ is linear from $\mathcal{M}(\mathcal{X})$ to (probability) distributions (S_1 and S_2), $\mathbf{\Phi}_n$ is linear from $\mathcal{M}(\mathcal{X})$ to $\mathcal{Y}^{\otimes n}$ (S_3), and $\mathbf{e}_n$ some noise.

$$\mu^\star \longrightarrow \mathbf{\Phi}_n\mu^\star \longrightarrow \mathbf{\Phi}_n\mu^\star + \mathbf{e}_n \longrightarrow \hat{\mu}_n \qquad (S_3)$$

Some inverse problems

[**P1**] (Mixtures): With scenario S_1, the observation y_k is sampled from a mixture density $f^\star$, and target $\mu^\star$ is a mixing law where $a_k^\star$ are mixture weights and $x_k^\star$ are mixture parameters:

$$f^\star = \sum_{k=1}^{K^\star} a_k^\star \varphi_{x_k^\star}\,, \quad \text{and} \quad \mathbf{\Phi}\mu := \int_{\mathcal{X}} \varphi_x \mathrm{d}\mu(x)\,,$$

where φ_x denotes a parametric density function with parameters x.

[**P2**] (Continuous Sparse): With scenario S_3, the observation y_k is a noisy linear measurement of $\mu^\star$:

$$y_k = (\mathbf{\Phi}_n \mu^\star)_k + e_k \quad \text{and} \quad (\mathbf{\Phi}_n \mu)_k := \int_{\mathcal{X}} \psi_k \mathrm{d}\mu\,, \tag{1}$$

where ψ_k is some known bounded function.

Some inverse problems

[**P3**] (Two-layer neural networks): With scenario S_3, one observes a couple (y_k, z_k) of input data z_k and response y_k as a linear measurement of $\mu^\star$:

$$f^\star = \sum_{k=1}^{K^\star} a_k^\star \varphi_{x_k^\star}\,, \quad y_k = (\mathbf{\Phi}_n \mu^\star)_k + e_k \quad \text{and} \quad (\mathbf{\Phi}_n \mu)_k := \int_{\mathcal{X}} \varphi_x(z_k) \mathrm{d}\mu(x)\,,$$

where $f^\star$ is the target function, $\varphi_x(z) := \sigma(\langle x, (1, z)\rangle) = \sigma(x_1 + \sum_{j=2}^d x_j z_j)$ is the neuron outcome with activation σ and weights x at input point z.

[**P4**] (Kernel Sparse Designs): With S_2, given $f^\star$, we aim at $\mu^\star$ s.t.

$$\mathbf{F}(\mu) := \left\| \mathbf{\Phi}\mu - \int_{\mathcal{X}} \varphi_x(\cdot) f^\star(x) \mathrm{d}x \right\|_{\mathcal{F}}, \quad \text{with} \quad \mathbf{\Phi}\mu := \int_{\mathcal{X}} \varphi_x \mathrm{d}\mu(x)\,,$$

is the best approximation (for the criterion $\mathbf{F}(\mu)$) of $f^\star$, and φ_x feature map at feature input point $x \in \mathcal{X}$.

[**P5**] (Symmetric Tensors): With scenario S_3, we aim at finding a $K^\star$-rank d-way symmetric tensor $\sigma^\star$ from the noisy linear measurements y_k:

$$\sigma^\star := \sum_{k=1}^{K^\star} a_k^\star \, x_k^{\star \otimes d} = \int_{\mathcal{X}} x^{\otimes d} \mathrm{d}\mu^\star(x) \,, \quad (\mathbf{\Phi}_n \mu)_k := \int_{\mathcal{X}} \psi_k(x) \mathrm{d}\mu(x) \,,$$

with $a_k^\star > 0$, $x_k^\star$ are normalized ($\|x_k^\star\|_2 = 1$), $\psi_k(x) := \langle \tau_k \,,\, x^{\otimes d} \rangle$ for some $\tau_k^\star \in (\mathbb{R}^n)^{\otimes d}$ (e.g., the canonical basis), $\langle \cdot \,,\, \cdot \rangle$ the standard dot product of tensors, and $e_k \in \mathbb{R}$ some centered random variable. Note that $(\mathbf{\Phi}_n \mu^\star)_k = \langle \tau_k \,,\, \sigma^\star \rangle$ is some linear form evaluated at the target point $\sigma^\star$.

We consider BLASSO solutions:

$$\hat{\mu}_n \in \arg \min_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ \mathbf{F}_n(\mathbf{Y}_n, \mu) + \lambda_n |\mu| \right\}$$

where the total variation norm is $|\mu| = \sup \left\{ \int_{\mathcal{X}} f \mathrm{d}\mu \ : \ |f| \leq 1 \right\}$, the so-called "*data fitting*" term $\mathbf{F}_n(\mathbf{Y}_n, \mu)$ quantifies how much the measure μ is likely to fit the data $\mathbf{Y}_n$ (here, the smaller the better), and $\lambda_n > 0$ a tuning parameter.

Existence Theorem

We consider BLASSO solutions:

$$\hat{\mu}_n \in \arg \min_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ \mathbf{F}_n(\mathbf{Y}_n, \mu) + \lambda_n |\mu| \right\}$$

where the total variation norm is $|\mu| = \sup \left\{ \int_{\mathcal{X}} f \mathrm{d}\mu \ : \ |f| \leq 1 \right\}$, the so-called "*data fitting*" term $\mathbf{F}_n(\mathbf{Y}_n, \mu)$ quantifies how much the measure μ is likely to fit the data $\mathbf{Y}_n$ (here, the smaller the better), and $\lambda_n > 0$ a tuning parameter.

Theorem: Let $\mathbb{H}$ be separable Hilbert space and let $\mathcal{X}$ be compact metric space. Consider the problem

$$\inf_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ L(\Phi\mu) + \lambda \|\mu\|_1 \right\} \tag{1}$$

where $L : \mathbb{H} \to [0, \infty]$ is convex and lower semi-continuous. There exists $\mu^\star \in \mathcal{M}(\mathcal{X})$ solution to (1). Furthermore, if L is strictly convex, then $\Phi(\mu^\star) \in \mathbb{H}$ is unique (it does not depend on the choice of the solution $\mu^\star$).

Existence Theorem

We consider BLASSO solutions:

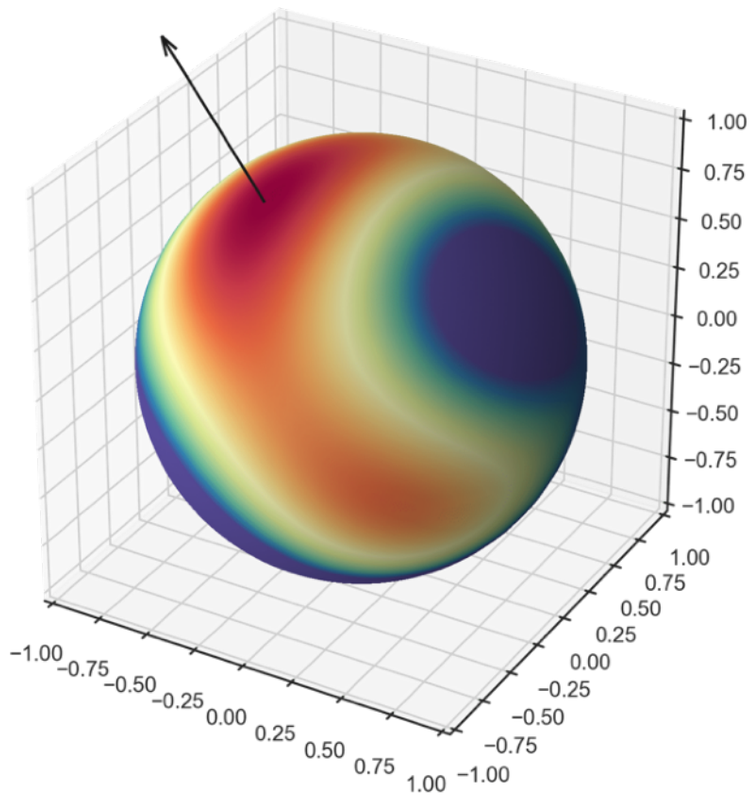
$$\hat{\mu}_n \in \arg \min_{\mu \in \mathcal{M}(\mathcal{X})} \left\{ \mathbf{F}_n(\mathbf{Y}_n, \mu) + \lambda_n |\mu| \right\}$$

where the total variation norm is $|\mu| = \sup \left\{ \int_{\mathcal{X}} f \mathrm{d}\mu \ : \ |f| \leq 1 \right\}$, the so-called "*data fitting*" term $\mathbf{F}_n(\mathbf{Y}_n, \mu)$ quantifies how much the measure μ is likely to fit the data $\mathbf{Y}_n$ (here, the smaller the better), and $\lambda_n > 0$ a tuning parameter.

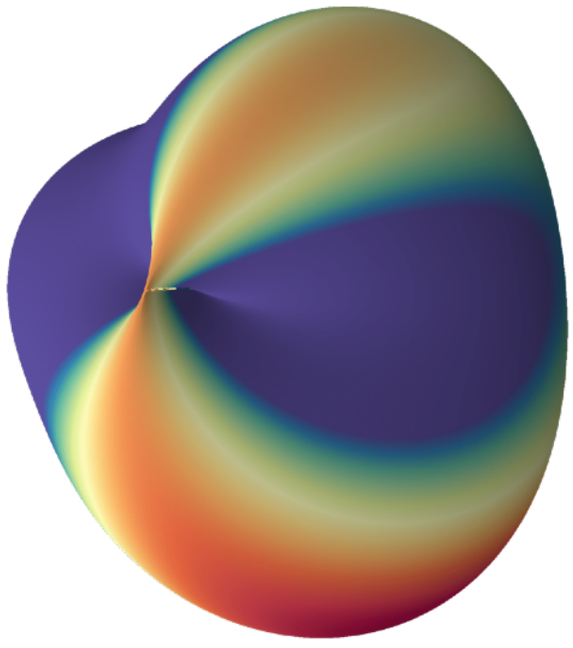
Theorem: If Φ has rank m then there exists a $m+1$-Diracs solution to (1).

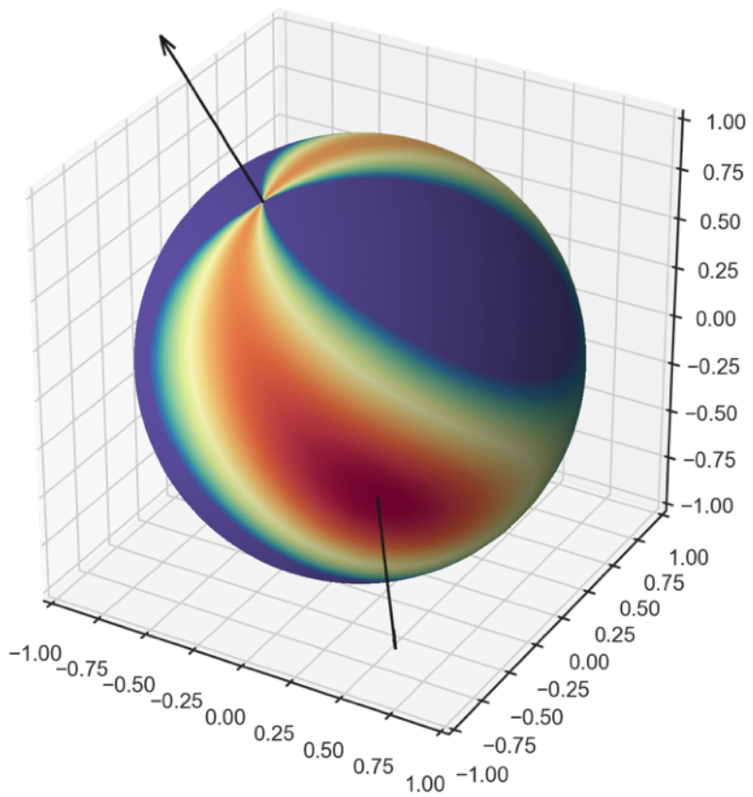
- In Sketching, the rank of Φ is at most the number of sketches;
- In Super-Resolution, the rank of Φ is finite;

Interlude: Regularization path on Tensor PCA



$x \mapsto \langle \mathbf{Y}, \varphi_x \rangle$



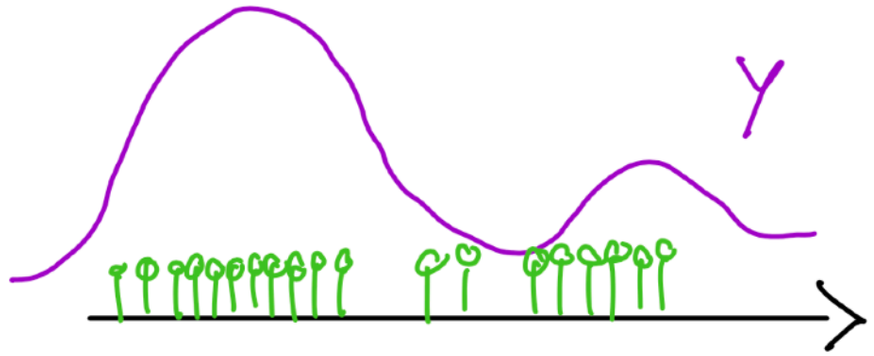


$x \mapsto \frac{\langle \mathbf{Y} - \widehat{a}_1 \widehat{x}_1^{\otimes d}, \varphi_x \rangle}{1 - \langle x, \widehat{x}_1 \rangle^d}$

Back to GMMs

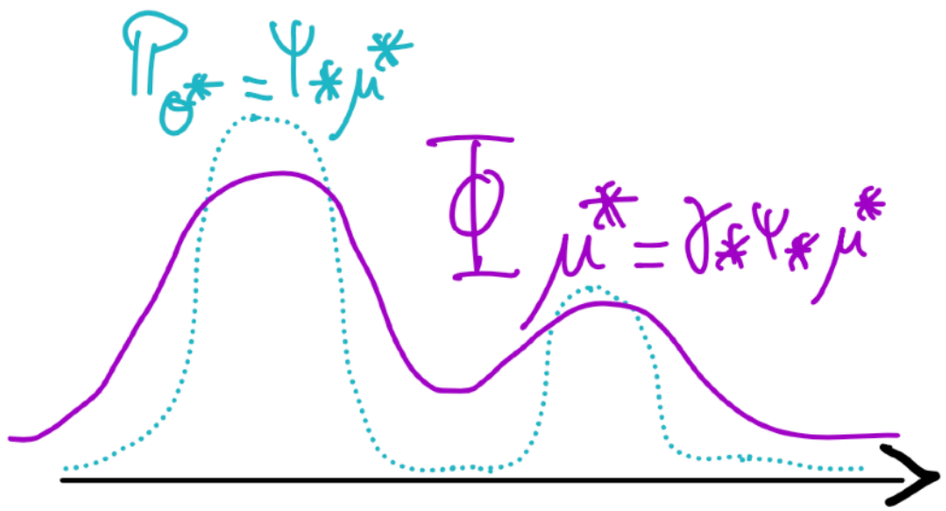
Reminder: kernel estimator

- Consider a kernel $\gamma(t-s)$ with bandwidth σ_γ
- In this talk, γ such that $\mathcal{F}\gamma(\omega) = \mathbb{1}_{\|\omega\|_\infty \leq m}$ with bandwidth $\sigma_\gamma = 1/m$
- Let $\mathbb{H}$ be the RKHS associated to γ
- Set $\varphi_t := \gamma \star \psi_t$ and $\Phi\mu = \displaystyle\int_{\mathcal{X}} \varphi_t \mathrm{d}\mu(t)$



- Kernel embedding
  - Data embedding: $Y = \gamma \star \mathbb{P}_N$
  - Measure embedding: $\Phi\mu$

both in $\mathbb{H}$



- Check that $\mathbb{E}Y = \Phi\mu^\star$
- BLASSO: $\min_\mu \left\{ \frac{1}{2}\|Y - \Phi\mu\|_{\mathbb{H}}^2 + \lambda\|\mu\|_1 \right\}$

Basis Pursuit (BP) aka Beurling Minimal Extrapolation (BME)

- BME: $\min_{\mu} \left\{ \|\mu\|_1 \ : \ \mu \text{ s.t. } \Phi\mu = \Phi\mu^\star (= \mathbb{E}Y) \right\}$

Theorem $(A) \Longleftrightarrow (B)$ where
(A) $\mu^\star$ solution to (BME);
(B) There exists $\eta \in \mathrm{Im}(\Phi^\star)$ such that

$$(i) \ \|\eta\|_\infty \leq 1 \text{ and } \eta(t_j^\star) = \mathrm{sign}(w_j^\star)\,, \ j = 1, \dots, s^\star$$

Furthermore, $\mu^\star$ is the **unique** solution to BME if (i) holds and

$$|\eta(t)| < 1\,, \ \forall t \neq t_1^\star, \dots, t_{s^\star}^\star$$

- no matter the weights $|w_j^\star|$
- η is a sub-gradient of the TV-norm at $\mu^\star$

Minimal separation and dual certificate

- Define the **minimal separation** as

$$\Delta := \min_{t_j^\star \neq t_\ell^\star} \|t_j^\star - t_\ell^\star\|_2$$

- Recall the **bandwidth** $\sigma_\gamma = 1/m$

- Resolution limit

$$\sigma_\gamma \lesssim \frac{\Delta}{\sqrt{s^\star d^3}} \qquad (\mathcal{H}_{\mathrm{res}})$$

- BLASSO: $\min_\mu \left\{ \frac{1}{2}\|Y - \Phi\mu\|_{\mathbb{H}}^2 + \lambda\|\mu\|_1 \right\}$

Theorem: BLASSO with $\lambda = \mathcal{O}\big(\frac{1}{\sqrt{\sigma_\gamma^d N}}\big)$ achieves

$$|w_j^\star - \widehat{\mu}(t_j^\star + \varepsilon B_2)| = \mathcal{O}(\lambda) \quad \text{and} \quad \widehat{\mu}\big(\mathcal{X} \setminus \cup_j \{t_j^\star + \varepsilon B_2\}\big) = \mathcal{O}(\lambda)$$

where $\varepsilon = \mathcal{O}(\sigma_\gamma)$ and B_2 unit Euclidean ball.

Minimal separation and dual certificate

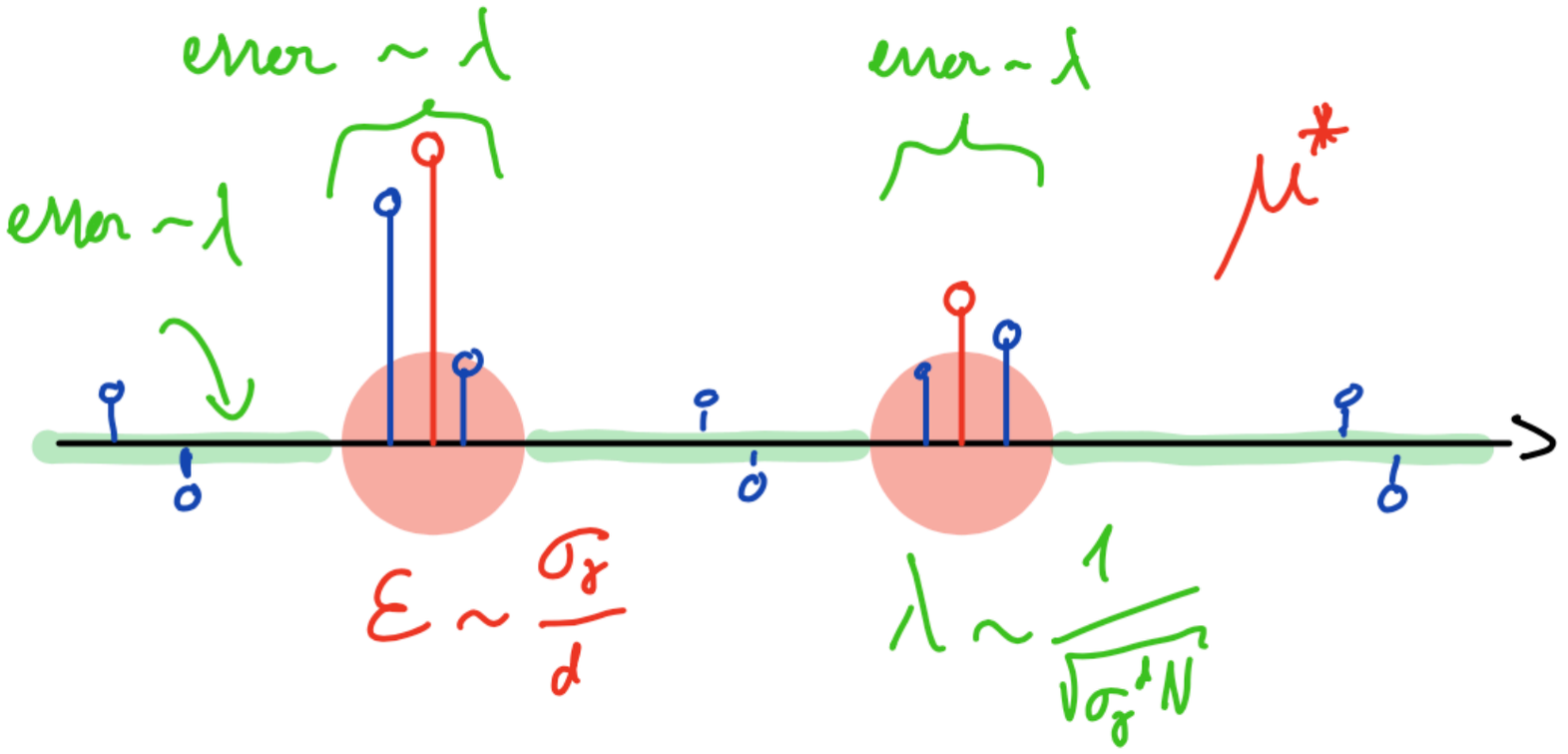
- Define the **minimal separation** as

$$\Delta := \min_{t_j^\star \neq t_\ell^\star} \|t_j^\star - t_\ell^\star\|_2$$

- Recall the **bandwidth** $\sigma_\gamma = 1/m$

- Resolution limit

$$\sigma_\gamma \lesssim \frac{\Delta}{\sqrt{s^\star d^3}} \qquad (\mathcal{H}_{\text{res}})$$



Optimization strategies for BLASSO

- **Conic Particle Gradient Descent**: optimizes a function on m parameters (x_k, a_k) (called *particles*) that can be written as $\mathbf{F}_n(\mathbf{Y}_n, \sum_{k=1}^m a_k \delta_{x_k})$ then we can study the limit $m \to \infty$ (over-parametrized) by considering the objective $\mu \mapsto \mathbf{F}_n(\mathbf{Y}_n, \mu)$ on the space of measures as in the works of Bach and Chizat, and Chizat.

- **Sliding Frank-Wolfe**: solves convex programs on weakly compact sets (e.g., closed balls of the TV-norm for the weak-$\star$ topology). This algorithm is a conditional gradient descent that may converge in a finite number of steps (Denoyelle, Duval, Peyré, and Soubies) under some conditions.

- **Kernel SoS**: Based on representation of nonnegative function based and a subsampling strategy, see Bach, Rudi, and Marteau-Ferey, and Lasserre, Magron, et al.

- **Other popular methods**: Prony-type spectral methods such as MUSIC and ESPRIT, and non-convex approaches based on greedy minimization (e.g., (COMP) and "Continuous" LARS), see Elvira, Gribonval, Soussen, and Herzet.

As of today's research

- **Sketching** BLASSO to reduce the number of measurements up to the Information Theory limit; Ongoing with R. Gribonval and N. Jouvin

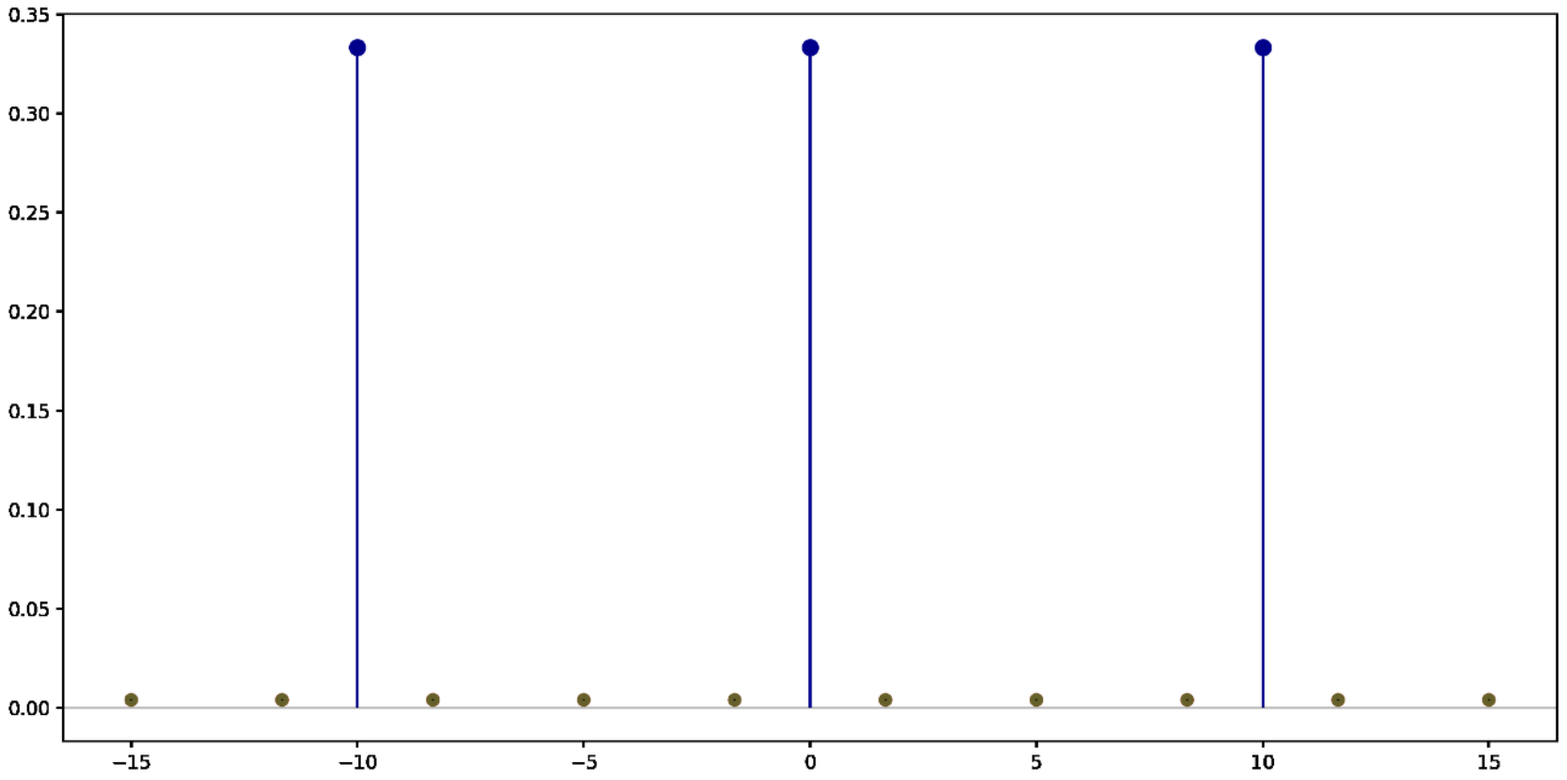
- **SGD** Conic Particle Gradient Descent; Ongoing with S. Gadat and C. Marteau

- **Logistic** BLASSO; Ongoing with Antoine SIMOES (PhD @ECL)

- **Inference** BLASSO; Ongoing with F. Dalmao (Uruguay) and JM Azaïs

- **Tensor** BLASSO dual certificate; Ongoing with C. Boyer and V. Duval

- Convolutional Kernel Networks, Dual certificates for 2 layers NN, Kernel Sparse designs, Kernel SoS... feel free to join!



Courtesy of **Nicolas Jouvin** (ex ICJ, now INRAE)

https://nicolasjouvin.github.io/

• **SGD** Conic Particle Gradient Descent; Ongo [arXiv > math > arXiv:2312.05993] eau

Informal results CPGD with particles $\nu = \sum_{\ell=1}^{p} w_\ell \delta_{t_\ell}$ consists in optimizing (non-convex) on (w_ℓ, t_ℓ) instead of ν (convex).

- Objective involves **only** the kernel $K(x, x') := \langle \varphi_x \, , \, \varphi_{x'} \rangle_{\mathbb{H}}$;
- Compatible with **Sketching** by substituting K by a random random on rank m (number of sketches);
- **Dimension reduction/Privacy:** Only $m = \mathcal{O}(ds^\star \log)$ sketches are sufficient, instead of N samples;
- Compatible with stochastic gradient descent:
  - Update a batch of particles;
  - Use a batch of sample;
  - Approximate kernel by rank-one random kernel;