Sparsity in imaging:
Sparse recovery and the Beurling LASSO

Clarice Poon
University of Bath

January, 2019

Outline

1. The sparse spikes problem
2. The BLASSO and dual certificates
3. Minimal norm certificate and support stability
4. Analysis of the minimal norm certificate
5. Recovery statements
6. Numerical algorithms

The space of Radon measures

Let $\mathcal{X} \subset \mathbb{R}^d$. The space of Radon measures $\mathcal{M}(\mathcal{X})$ is defined as the dual of

$$C_0(\mathcal{X}) \stackrel{\text{def.}}{=} \overline{\{f \in \mathcal{C}(\mathcal{X}) \,;\, f \text{ has compact support in } \mathcal{X}\}}^{\|\cdot\|_\infty}$$

endowed with the uniform norm.

$\mathcal{M}(\mathcal{X})$ is a Banach space with the dual norm

$$|\mu|(\mathcal{X}) = \sup \left\{ \operatorname{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \,;\, \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\}.$$

This is called the **total variation norm.**

The space of Radon measures

Let $\mathcal{X} \subset \mathbb{R}^d$. The space of Radon measures $\mathcal{M}(\mathcal{X})$ is defined as the dual of

$$C_0(\mathcal{X}) \stackrel{\text{def.}}{=} \overline{\{f \in \mathcal{C}(\mathcal{X}) \; ; \; f \text{ has compact support in } \mathcal{X}\}}^{\|\cdot\|_\infty}$$

endowed with the uniform norm.

$\mathcal{M}(\mathcal{X})$ is a Banach space with the dual norm

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\}.$$

This is called the **total variation norm.**

Examples:

- $\mu \stackrel{\text{def.}}{=} \sum_{j=1}^s a_j \delta_{x_j} \in \mathcal{M}(\mathcal{X})$ where $a_j \in \mathbb{C}$ and $a\delta_x$ denotes the Dirac at $x \in \mathcal{X}$ with amplitude $a \in \mathbb{C}$. Moreover, $|\mu|(\mathcal{X}) = \sum_j |a_j|$.
- If μ is such that $f = \frac{\mathrm{d}\mu}{\mathrm{d}x}$ with $f \in L^1(\mathcal{X})$, then $|\mu|(\mathcal{X}) = \|f\|_{L^1}$.

The sparse spikes problem

Aim: Recover $\mu_0 \in \mathcal{M}(\mathcal{X})$, $\mathcal{X} \subseteq \mathbb{R}^d$, from m observations, $y = \Phi\mu_0 + w$.

- $w \in \mathcal{H}$ is the additive noise
- $\Phi : \mathcal{M}(\mathcal{X}) \to \mathcal{H}$, $\Phi\mu = \int \varphi(x)\mathrm{d}\mu(x)$ with $\varphi \in \mathcal{C}(\mathcal{X}, \mathcal{H})$.
- Typically, the measure of interest is of the form $\mu_0 = \sum_{j=1}^{s} a_j \delta_{x_j}$.

Examples

Deconvolution

- $\mathcal{H} = L^2(\mathcal{X})$, $\varphi(x) = t \mapsto \psi(x-t) \in L^2(\mathcal{X})$ for some $\psi \in L^2(\mathcal{X})$.
- $(\Phi\mu)(t) = \sum_j a_j \psi(x_j - t)$.
- $\psi(t) = \exp\left(-\|t\|^2\right)$ for Gaussian deconvolution.

Sampling Fourier coefficients

- Let $\mathcal{X} = \mathbb{T}^d$, $\mathcal{H} = \mathbb{C}^m$ and $\varphi(x) = \left(e^{2\pi i \langle x,\, \omega \rangle}\right)_{\omega \in \Omega}$ where $\Omega \subset \mathbb{R}^d$ consists of m values.
- If $\Omega = \left\{k \in \mathbb{Z}^d \; ; \; |k|_\infty \leqslant f_c\right\}$, then $m = (2f_c + 1)^d$ and $\Phi\mu = \left(\sum_j a_j e^{2\pi i \langle k,\, x_j \rangle}\right)_{|k| \leqslant f_c}$.

Sampling the Laplace transform

- $\varphi(x) = t \mapsto \exp(-\langle x,\, t\rangle)$.
- $(\Phi\mu)(t) = \sum_j a_j \exp(-\langle x_j,\, t\rangle)$.

Examples

Neuron imaging (EEG/MEG)

- We want to locate point sources on some domain $\mathcal{X}$ given boundary measurements.
- Let $\mathcal{H} = L^2(\partial\mathcal{X})$, and $\varphi(x) = (\psi(x,t))_{t\in\partial\mathcal{X}}$ for some kernel ψ.
- E.g. $\psi(x,t) = \|x-t\|^{-2}$.

In **machine learning**, you may want to fit a probability distribution to some data.

- Estimate parameters $(a_i) \in \mathbb{R}^N$ and $(x_i)_{i=1}^N \in \mathcal{X}^N$ of a mixture $\sum_{i=1}^N a_i\varphi(x_i)$ of N elementary distributions.
- w accounts for the sampling scheme.

Gaussian mixture model: In a simple setup, consider recovering the means $m \in \mathbb{R}$ and standard deviation $s \in \mathbb{R}_+$ of a Gaussian mixture, i.e. $x = (m,s) \in \mathcal{X} = \mathbb{R}\times\mathbb{R}_+$ and $\varphi(x) = \frac{1}{s}e^{-(\cdot-m)^2/(2s^2)} \in \mathcal{H} = L^2(\mathbb{R})$.

Relation to the compressed sensing problem

- Note that in compressed sensing, we aim to recover an s-sparse vector $v_0 \in \mathbb{C}^N$ from m measurements of the form Av_0 where $A \in \mathbb{C}^{m\times N}$.
- There are $2s$ unknowns, since we need to locate the support and the corresponding amplitudes of v_0.
- In the sparse spikes problem, we want to recover $\mu_0 = \sum_{j=1}^{s} a_j \delta_{x_j}$. So there are still $2s$ unknown values $\{(a_j, x_j)\}_{j=1}^{s}$, however, the points $\{x_j\}_{j=1}^{s}$ are no longer constrained to a finite set of values.

Relation to the compressed sensing problem

- Note that in compressed sensing, we aim to recover an s-sparse vector $v_0 \in \mathbb{C}^N$ from m measurements of the form Av_0 where $A \in \mathbb{C}^{m \times N}$.
- There are $2s$ unknowns, since we need to locate the support and the corresponding amplitudes of v_0.
- In the sparse spikes problem, we want to recover $\mu_0 = \sum_{j=1}^{s} a_j \delta_{x_j}$. So there are still $2s$ unknown values $\{(a_j, x_j)\}_{j=1}^{s}$, however, the points $\{x_j\}_{j=1}^{s}$ are no longer constrained to a finite set of values.

Off-the-grid compressed sensing

Set $\mathcal{H} \overset{\text{def.}}{=} \mathbb{C}^m$ and Φ is a linear operator defined as follows:

- Let (Ω, Λ) be a probability space. For $\omega \in \Omega$, we have random features $\varphi_\omega \in \mathcal{C}(\mathcal{X})$.
- $\varphi(x) \overset{\text{def.}}{=} \frac{1}{\sqrt{m}} (\varphi_{\omega_k}(x))_{k=1}^{m}$.
- The sampling operator is $\Phi : \mathcal{M}(\mathcal{X}) \to \mathbb{C}^m$, $\Phi\mu \overset{\text{def.}}{=} \frac{1}{\sqrt{m}} \left(\int \varphi_{\omega_k}(x) \mathrm{d}\mu(x)\right)_{k=1}^{m}$, where $\omega_k \overset{iid}{\sim} \Lambda$.

Examples

- **Random Fourier sampling:** instead of recovering μ from $(\mathcal{F}\mu(\omega))_{|\omega|_\infty \leqslant f_c}$, recover from $(\mathcal{F}\mu(\omega_k))_{k=1}^m$ where $\mathcal{F}$ is the Fourier transform and ω_k are drawn iid from $(\llbracket -f_c, f_c \rrbracket^d, \mathrm{Unif})$. Here, $\varphi_\omega(x) = \exp\left(-\mathrm{i}2\pi x^\top \omega\right)$.
- **Sampling the Laplace transform:** Recover $\mu \in \mathcal{M}(\mathbb{R}_+^d)$ from $(\mathcal{L}\mu(\omega_k))_{k=1}^m$ where $\mathcal{L}$ is the Laplace transform and ω_k are drawn iid from $(\mathbb{R}_+^d, \Lambda_\alpha)$ where $\Lambda_\alpha(\omega) \propto \exp\left(-2\alpha^\top \omega\right)$. Here, $\varphi_\omega(x) = \exp\left(-x^\top \omega\right)$.

Density estimation with sketching

Given data on $\mathcal{T}$, estimate parameters $(a_i) \in \mathbb{R}_+^N$ and $(x_i)_{i=1}^s \in \mathcal{X}^s$ of a mixture

$$\xi(t) = \sum_{j=1}^{s} a_j \xi_{x_j}(t) = \int_{\mathcal{X}} \xi_x(t) \mathrm{d}\mu_0(x)$$

where $\mu_0 = \sum_j a_j \delta_{x_j}$ where $(\xi_x)_{x \in \mathcal{X}}$ is a family of template distributions. E.g. $x = (m, \sigma) \in \mathcal{X} = \mathbb{R} \times \mathbb{R}_+$ and $\xi_x = \mathcal{N}(m, \sigma)$.

Density estimation with sketching

Given data on $\mathcal{T}$, estimate parameters $(a_i) \in \mathbb{R}_+^N$ and $(x_i)_{i=1}^s \in \mathcal{X}^s$ of a mixture

$$\xi(t) = \sum_{j=1}^{s} a_j \xi_{x_j}(t) = \int_{\mathcal{X}} \xi_x(t) \mathrm{d}\mu_0(x)$$

where $\mu_0 = \sum_j a_j \delta_{x_j}$ where $(\xi_x)_{x\in\mathcal{X}}$ is a family of template distributions. E.g. $x = (m, \sigma) \in \mathcal{X} = \mathbb{R} \times \mathbb{R}_+$ and $\xi_x = \mathcal{N}(m, \sigma)$.

Sketching [Gribonval, Blanchard, Keriven & Traonmilin, 2017]

Typically, there is no direct access to ξ but n iid samples $(t_1, \dots, t_n) \in \mathcal{T}^n$ drawn from ξ. Moreover, since n might be very large, rather than recording this huge set of data, one could compute online a small set $y \in \mathbb{C}^m$ of m "sketches" against sketching functions $\theta_\omega(t)$:

$$y_k \overset{\text{def.}}{=} \frac{1}{n} \sum_{j=1}^{n} \theta_{\omega_k}(t_j) \approx \int_{\mathcal{T}} \theta_{\omega_k}(t)\xi(t)\mathrm{d}t = \int_{\mathcal{X}} \int_{\mathcal{T}} \theta_{\omega_k}(t)\xi_x(t)\mathrm{d}t\mathrm{d}\mu_0(x).$$

So, we are back to the sparse spikes problem with $\varphi_\omega(x) \overset{\text{def.}}{=} \int_{\mathcal{T}} \theta_{\omega_k}(t)\xi_x(t)\mathrm{d}t$. For example, if $\theta_\omega(t) = e^{\mathrm{i}\langle\omega,\, t\rangle}$, then $\varphi_\cdot(x)$ is the characterisatic function of ξ_x.

Outline

1. The sparse spikes problem
2. The BLASSO and dual certificates
3. Minimal norm certificate and support stability
4. Analysis of the minimal norm certificate
5. Recovery statements
6. Numerical algorithms

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \tag{$\mathcal{P}_\lambda(y)$}$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\}.$$

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \qquad (\mathcal{P}_\lambda(y))$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\} .$$

In the noiseless case, consider

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \qquad (\mathcal{P}_\lambda(y))$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\}.$$

In the noiseless case, consider

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

Questions:

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \qquad (\mathcal{P}_\lambda(y))$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\} .$$

In the noiseless case, consider

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

Questions:

- Under what conditions can we recover a sparse measure $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ exactly in the noiseless setting by solving $(\mathcal{P}_0(y))$?

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \qquad (\mathcal{P}_\lambda(y))$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\} .$$

In the noiseless case, consider

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

Questions:

- Under what conditions can we recover a sparse measure $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ exactly in the noiseless setting by solving $(\mathcal{P}_0(y))$?
- If μ_0 can be recovered in the noiseless setting, can it be stably recovered via $(\mathcal{P}_\lambda(y))$?

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \qquad (\mathcal{P}_\lambda(y))$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \operatorname{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\}.$$

In the noiseless case, consider

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

Questions:

- Under what conditions can we recover a sparse measure $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ exactly in the noiseless setting by solving $(\mathcal{P}_0(y))$?
- If μ_0 can be recovered in the noiseless setting, can it be stably recovered via $(\mathcal{P}_\lambda(y))$?
- The question of stability is a little more delicate here. Given $\mu_1 = \sum_j a_j \delta_{x_j}$ and $\mu_2 = \sum_j a'_j \delta_{x'_j}$, we have $|\mu_1 - \mu_2|(\mathcal{X}) = \sum_j |a_j| + |a'_j|$.

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \qquad (\mathcal{P}_\lambda(y))$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\}.$$

In the noiseless case, consider

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

Questions:

- Under what conditions can we recover a sparse measure $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ exactly in the noiseless setting by solving $(\mathcal{P}_0(y))$?
- If μ_0 can be recovered in the noiseless setting, can it be stably recovered via $(\mathcal{P}_\lambda(y))$?
- The question of stability is a little more delicate here. Given $\mu_1 = \sum_j a_j \delta_{x_j}$ and $\mu_2 = \sum_j a'_j \delta_{x'_j}$, we have $|\mu_1 - \mu_2|(\mathcal{X}) = \sum_j |a_j| + |a'_j|$.
- When do we have support stability? That is, we recover exactly s spikes and have control on error of the amplitudes and positions.

The Beurling LASSO (BLASSO)

Let us consider the following optimisation problem:

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) + \frac{1}{2\lambda} \|\Phi\mu - y\|^2 . \qquad (\mathcal{P}_\lambda(y))$$

where $\lambda > 0$ is a regularisation parameter and the total variation norm $|\mu|(\mathcal{X})$ of $\mu \in \mathcal{M}(\mathcal{X})$ is defined as

$$|\mu|(\mathcal{X}) = \sup \left\{ \mathrm{Re} \int_{\mathcal{X}} \eta(x) \mathrm{d}\mu(x) \; ; \; \eta \in C_0(\mathcal{X}), \|\eta\|_{L^\infty} \leqslant 1 \right\}.$$

In the noiseless case, consider

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = y. \qquad (\mathcal{P}_0(y))$$

Questions:

- Under what conditions can we recover a sparse measure $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ exactly in the noiseless setting by solving $(\mathcal{P}_0(y))$?
- If μ_0 can be recovered in the noiseless setting, can it be stably recovered via $(\mathcal{P}_\lambda(y))$?
- The question of stability is a little more delicate here. Given $\mu_1 = \sum_j a_j \delta_{x_j}$ and $\mu_2 = \sum_j a'_j \delta_{x'_j}$, we have $|\mu_1 - \mu_2|(\mathcal{X}) = \sum_j |a_j| + |a'_j|$.
- When do we have support stability? That is, we recover exactly s spikes and have control on error of the amplitudes and positions.
- Numerical algorithms which respect the infinite dimensional structure?

Outline

1. The sparse spikes problem
2. The BLASSO and dual certificates
3. Minimal norm certificate and support stability
4. Analysis of the minimal norm certificate
5. Recovery statements
6. Numerical algorithms

Optimality condition

Let us first remark that $|\mu|(\mathcal{X})$ is non-differentiable (just like the ℓ^1-norm is not differentiable), so we consider instead its subdifferential

$$\partial |\mu|(\mathcal{X}) \overset{\text{def.}}{=} \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; |\tilde{\mu}|(\mathcal{X}) \geqslant |\mu|(\mathcal{X}) + \int \eta \mathrm{d}(\tilde{\mu} - \mu) \right\}$$

Optimality condition

Let us first remark that $|\mu|(\mathcal{X})$ is non-differentiable (just like the ℓ^1-norm is not differentiable), so we consider instead its subdifferential

$$\partial |\mu|(\mathcal{X}) \stackrel{\text{def.}}{=} \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; |\tilde{\mu}|(\mathcal{X}) \geqslant |\mu|(\mathcal{X}) + \int \eta \mathrm{d}(\tilde{\mu} - \mu) \right\}$$

One can show that

$$\partial |\mu|(\mathcal{X}) = \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; \|\eta\|_\infty \leqslant 1 \quad \text{and} \quad \int \eta \mathrm{d}\mu = |\mu|(\mathcal{X}) \right\}.$$

Optimality condition

Let us first remark that $|\mu|(\mathcal{X})$ is non-differentiable (just like the ℓ^1-norm is not differentiable), so we consider instead its subdifferential

$$\partial |\mu|(\mathcal{X}) \stackrel{\text{def.}}{=} \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; |\tilde{\mu}|(\mathcal{X}) \geqslant |\mu|(\mathcal{X}) + \int \eta \mathrm{d}(\tilde{\mu} - \mu) \right\}$$

One can show that

$$\partial |\mu|(\mathcal{X}) = \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; \|\eta\|_\infty \leqslant 1 \quad \text{and} \quad \int \eta \mathrm{d}\mu = |\mu|(\mathcal{X}) \right\}.$$

In particular, if $\mu = \sum_j a_j \delta_{x_j}$,

$$\partial |\mu|(\mathcal{X}) = \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; \|\eta\|_\infty \leqslant 1 \quad \text{and} \quad \forall j, \; \eta(x_j) = \mathrm{sign}(a_j) \right\}.$$

and given any $\mu \in \mathcal{M}(\mathcal{X})$ and $\eta \in \partial |\mu|(\mathcal{X})$,

$$\mathrm{Supp}(\mu) \subseteq \{ x \in \mathcal{X} \; ; \; |\eta(x)| = 1 \}.$$

Optimality condition

Let us first remark that $|\mu|(\mathcal{X})$ is non-differentiable (just like the ℓ^1-norm is not differentiable), so we consider instead its subdifferential

$$\partial |\mu|(\mathcal{X}) \overset{\text{def.}}{=} \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; |\tilde{\mu}|(\mathcal{X}) \geqslant |\mu|(\mathcal{X}) + \int \eta \mathrm{d}(\tilde{\mu} - \mu) \right\}$$

One can show that

$$\partial |\mu|(\mathcal{X}) = \left\{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; \|\eta\|_\infty \leqslant 1 \quad \text{and} \quad \int \eta \mathrm{d}\mu = |\mu|(\mathcal{X}) \right\}.$$

In particular, if $\mu = \sum_j a_j \delta_{x_j}$,

$$\partial |\mu|(\mathcal{X}) = \{ \eta \in \mathcal{C}(\mathcal{X}) \; ; \; \|\eta\|_\infty \leqslant 1 \quad \text{and} \quad \forall j, \; \eta(x_j) = \operatorname{sign}(a_j) \}.$$

and given any $\mu \in \mathcal{M}(\mathcal{X})$ and $\eta \in \partial |\mu|(\mathcal{X})$,

$$\operatorname{Supp}(\mu) \subseteq \{ x \in \mathcal{X} \; ; \; |\eta(x)| = 1 \}.$$

Optimality condition

FACT: μ is a minimiser of a convex functional F if and only if $0 \in \partial F(\mu)$.
A discrete measure $\mu = \sum_j a_j \delta_{x_j}$ is a solution of $(\mathcal{P}_\lambda(y))$ iff

$$0 \in \Phi^*(\Phi\mu - y) + \lambda \partial |\mu|(\mathcal{X}).$$

That is, $\eta \overset{\text{def.}}{=} \frac{1}{\lambda}\Phi^*(y - \Phi\mu)$ satisfies $\eta \in \partial |\mu|(\mathcal{X})$, $\eta(x_j) = \operatorname{sign}(a_j)$, and $\|\eta\|_\infty \leqslant 1$.

Duality

Fenchel dual problems

The dual problem to $(\mathcal{P}_\lambda(y))$ and $\mathcal{P}_0(y)$ are $(\mathcal{D}_\lambda(y))$ and $(\mathcal{D}_0(y))$ respectively:

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle y,\, p\rangle - \frac{\lambda}{2}\|p\|^2 \tag{$\mathcal{D}_\lambda(y)$}$$

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle y,\, p\rangle. \tag{$\mathcal{D}_0(y)$}$$

Duality

Fenchel dual problems

The dual problem to $(\mathcal{P}_\lambda(y))$ and $\mathcal{P}_0(y)$ are $(\mathcal{D}_\lambda(y))$ and $(\mathcal{D}_0(y))$ respectively:

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle y,\, p\rangle - \frac{\lambda}{2}\|p\|^2 \tag{$\mathcal{D}_\lambda(y)$}$$

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle y,\, p\rangle. \tag{$\mathcal{D}_0(y)$}$$

Comments:

- Solving $(\mathcal{D}_\lambda(y))$ is equivalent to

$$\min_{\|\Phi^* p\|_\infty \leqslant 1} \left\| \frac{y}{\lambda} - p \right\|^2$$

  This is a projection onto a closed convex set and we have immediately existence and uniqueness of the dual solution.
- Here, existence of solutions to $(\mathcal{D}_0(y))$ is not guaranteed, but is true when $\mathrm{Im}(\Phi^*)$ is finite dimensional.
- We have strong duality.

Duality

Fenchel dual problems

The dual problem to $(\mathcal{P}_\lambda(y))$ and $\mathcal{P}_0(y)$ are $(\mathcal{D}_\lambda(y))$ and $(\mathcal{D}_0(y))$ respectively:

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle y,\, p\rangle - \frac{\lambda}{2}\|p\|^2 \tag{$\mathcal{D}_\lambda(y)$}$$

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle y,\, p\rangle. \tag{$\mathcal{D}_0(y)$}$$

Primal and dual solutions

1. Primal solution μ_λ to $(\mathcal{P}_\lambda(y))$ and dual solution p_λ to $(\mathcal{D}_\lambda(y))$ satisfy

$$\Phi^* p_\lambda \in \partial |\mu_\lambda|(\mathcal{X}) \quad \text{and} \quad p_\lambda = -\frac{1}{\lambda}(\Phi\mu_\lambda - y)$$

2. If $\exists$ a solution p_0 to $(\mathcal{D}_0(y))$, then it is linked to any solution μ_0 of $(\mathcal{P}_0(y))$ by

$$\Phi^* p_0 \in \partial |\mu_0|(\mathcal{X}).$$

Note in particular that $\mathrm{Supp}(\mu_\lambda) \subset \{x \in \mathcal{X}\ ;\ |\Phi^* p_\lambda(x)| = 1\}$. These dual solutions correspond precisely to dual certificates in compressed sensing.

Unique recovery

Given $X \stackrel{\text{def.}}{=} \{x_j\}_{j=1}^s$, define $\Phi_X : \mathbb{R}^s \to \mathcal{H}$ by $\Phi_X a = \sum_j a_j \varphi(x_j)$.

Theorem

Let $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ and let $y = \Phi\mu_0$. Suppose that there exists $\eta = \Phi^* p$ such that such that

- *for all $j = 1, \ldots, s$, $\eta(x_j) = \mathrm{sign}(a_j)$,*
- *for all $x \notin \{x_j\}_j$, $|\eta(x)| < 1$*
- *Φ_X is injective.*

Then, μ_0 is the unique solution to $(\mathcal{P}_0(y))$.

Unique recovery

Given $X \overset{\text{def.}}{=} \{x_j\}_{j=1}^s$, define $\Phi_X : \mathbb{R}^s \to \mathcal{H}$ by $\Phi_X a = \sum_j a_j \varphi(x_j)$.

Theorem

Let $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ *and let* $y = \Phi\mu_0$*. Suppose that there exists* $\eta = \Phi^* p$ *such that such that*

- *for all* $j = 1, \ldots, s$, $\eta(x_j) = \mathrm{sign}(a_j)$,
- *for all* $x \notin \{x_j\}_j$, $|\eta(x)| < 1$
- Φ_X *is injective.*

Then, μ_0 *is the unique solution to* $(\mathcal{P}_0(y))$.

Proof.

Since $\eta \in \partial |\mu_0| (\mathcal{X})$, μ_0 must be a primal solution and p must a a dual solution. Moreover, any solution μ of $(\mathcal{P}_0(y))$ must satisfy $\mathrm{Supp}(\mu) \subset X$. Given two solutions $\mu = \sum_j a_j \delta_{x_j}$ and $\nu = \sum_j \tilde{a}_j \delta_{x_j}$, we have

$$\Phi(\mu - \nu) = \sum_j (a_j - \tilde{a}_j)\varphi(x_j) = \Phi_X(a - \tilde{a}) = 0$$

if and only if $a_j = \tilde{a}_j$ for all j. Therefore, $\mu = \mu_0$. □

Unique recovery

Given $X \stackrel{\text{def.}}{=} \{x_j\}_{j=1}^s$, define $\Phi_X : \mathbb{R}^s \to \mathcal{H}$ by $\Phi_X a = \sum_j a_j \varphi(x_j)$.

Theorem

Let $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ *and let* $y = \Phi\mu_0$. *Suppose that there exists* $\eta = \Phi^* p$ *such that such that*

- *for all* $j = 1, \ldots, s$, $\eta(x_j) = \text{sign}(a_j)$,
- *for all* $x \notin \{x_j\}_j$, $|\eta(x)| < 1$
- Φ_X *is injective.*

Then, μ_0 *is the unique solution to* $(\mathcal{P}_0(y))$.

Definition

We say that a certificate is nondegenerate wrt $\text{sign}(a)$ and $X \stackrel{\text{def.}}{=} \{x_j\}_j$ if $\eta(x_j) = \text{sign}(a_j)$, $\eta(x) < 1$ for all $x \notin X$ and $\text{sign}(a_j)\nabla^2\eta(x_j) \prec 0$. Precise control on the nondegeneracy of η around each x_j's will lead to bounds on how closely solutions to $(\mathcal{P}_\lambda(y))$ "cluster" around $\{x_j\}_j$.

Stability

Theorem (Candès & Fernandez Granda '14, Azaïs et al '15)

Let $\mu_0 = \sum_{j=1}^{s} a_j \delta_{x_j}$ *and suppose that* $\eta = \Phi^* p \in \partial |\mu_0| (\mathcal{X})$. *Suppose that there exists* $\varepsilon, c_2, c_0 > 0$ *and* η *such that*

- $|\eta(x)| \leqslant 1 - c_2 \|x - x_i\|^2$ *for all* $x \in B_\varepsilon(x_i)$.
- $|\eta(x)| < 1 - c_0$ *for all* $x \notin \bigcup_i B_\varepsilon(x_i)$.

Then, choosing $\lambda \sim \delta / \|p\|$, *any solution* μ *to* $(\mathcal{P}_\lambda(y))$ *with* $y = \Phi\mu_0 + w$ *and* $\|w\| \leqslant \delta$ *satisfies*

$$c_0 |\mu| \left(\mathcal{X} \setminus \bigcup_i B_\varepsilon(x_i) \right) + c_2 \sum_{i=1}^{s} \int_{B_\varepsilon(x_i)} \|x - x_i\|^2 \, \mathrm{d} |\mu| (x) \lesssim \delta \|p\| .$$

Stability

Theorem (Candès & Fernandez Granda '14, Azaïs et al '15)

Let $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ *and suppose that* $\eta = \Phi^* p \in \partial |\mu_0| (\mathcal{X})$. *Suppose that there exists* $\varepsilon, c_2, c_0 > 0$ *and* η *such that*

- $|\eta(x)| \leqslant 1 - c_2 \|x - x_i\|^2$ *for all* $x \in B_\varepsilon(x_i)$.
- $|\eta(x)| < 1 - c_0$ *for all* $x \notin \bigcup_i B_\varepsilon(x_i)$.

Then, choosing $\lambda \sim \delta / \|p\|$, *any solution* μ *to* $(\mathcal{P}_\lambda(y))$ *with* $y = \Phi\mu_0 + w$ *and* $\|w\| \leqslant \delta$ *satisfies*

$$c_0 |\mu| \left(\mathcal{X} \setminus \bigcup_i B_\varepsilon(x_i) \right) + c_2 \sum_{i=1}^s \int_{B_\varepsilon(x_i)} \|x - x_i\|^2 \,\mathrm{d}\, |\mu| (x) \lesssim \delta \|p\| .$$

Remark

Suppose that $\mu = \sum_{j=1}^s \sum_k \hat{a}_{j,k} \delta_{\hat{x}_{j,k}} + \sum_j \hat{b}_k \delta_{\hat{z}_k}$ where $\{\hat{x}_{j,k}\}_k \subset B_\varepsilon(x_j)$ and $\{\hat{z}_k\}_k \subset \mathcal{X} \setminus \bigcup_j B_\varepsilon(x_j)$. Then, this theore implies that

$$c_0 \sum_k \left| \hat{b}_k \right| + c_2 \sum_j \sum_k \left| \hat{x}_{j,k} - x_j \right|^2 \left| \hat{a}_{j,k} \right| \lesssim \delta \|p\|$$

which suggest that the closer $\hat{x}_{j,k}$ is to x_j, the smaller $\left|\hat{a}_{j,k}\right|$ should be.

Proof: step 1, bounding the Bregman "distance"

Lemma (Burger & Osher '04)

Let $\mu_0 \in \mathcal{M}(\mathcal{X})$ *be such that* $\|y - \Phi\mu_0\| \leqslant \delta$ *and let* $\eta = \Phi^* p$ *be such that* $\eta \in \partial |\mu_0|(\mathcal{X})$. *Then,*

$$d^\eta(\mu, \mu_0) \stackrel{\text{def.}}{=} |\mu|(\mathcal{X}) - |\mu_0|(\mathcal{X}) - \langle \eta, \mu - \mu_0 \rangle \leqslant \frac{\delta^2}{2\lambda} + \frac{\lambda \|p\|^2}{2} + \delta \|p\|.$$

Proof.

Since μ is a minimizer,

$$\lambda |\mu|(\mathcal{X}) + \frac{1}{2}\|\Phi\mu - y\|^2 \leqslant \lambda |\mu_0|(\mathcal{X}) + \frac{1}{2}\|\Phi\mu_0 - y\|^2 \leqslant \lambda |\mu_0|(\mathcal{X}) + \frac{\delta^2}{2}.$$

So,

$$\frac{1}{2}\|\Phi\mu - y\|^2 + \lambda d^\eta(\mu, \mu_0) + \lambda\langle \eta, \mu - \mu_0 \rangle \leqslant \frac{\delta^2}{2}.$$

By recalling that $\eta = \Phi^* p$,

$$\frac{1}{2}\|\Phi\mu - y + \lambda p\|^2 + \lambda d^\eta(\mu, \mu_0) - \frac{\lambda^2 \|p\|^2}{2} + \lambda\langle p, y - \Phi\mu_0 \rangle \leqslant \frac{\delta^2}{2},$$

and by rearranging the above inequality,

$$d^\eta(\mu, \mu_0) \leqslant \frac{\delta^2}{2\lambda} + \frac{\lambda \|p\|^2}{2} + \delta \|p\|.$$

□

Proof: step 2 lower bound on $d^\eta(\mu, \mu_0)$

Choosing $\lambda \sim \delta / \|p\|$, we have $d^\eta(\mu,\mu_0) \lesssim \delta \|p\|$. The claim of Theorem 2.2 follows combining this result with the following lower bound for $d^\eta(\mu, \mu_0)$:

Lemma

Under the assumptions of Theorem 2.2, we have

$$d^\eta(\mu,\mu_0) \geqslant c_2 \sum_j \int_{B_\varepsilon(x_j)} \|x - x_j\|^2 \,\mathrm{d}\,|\mu|\,(x) + c_0\,|\mu| \left(\bigcup_i B_\varepsilon(x_i) \right).$$

Proof.

Let $\mathcal{X}^{far} \overset{\text{def.}}{=} \mathcal{X} \setminus \bigcup_i B_\varepsilon(x_i)$.

(i) $|\mu|\,(\mathcal{X}) - |\mu_0| - \langle \eta,\, \mu - \mu_0 \rangle = |\mu|\,(\mathcal{X}) - \langle \eta,\, \mu \rangle$

(ii) $\langle \eta,\, \mu \rangle \leqslant \sum_i \int_{B_\varepsilon(x_i)} |\eta(x)| \,\mathrm{d}\,|\mu|\,(x) + \|\eta\|_{L^\infty(\mathcal{X}^{far})}\,|\mu|\,(\mathcal{X}^{far})$.

(iii) Plugging in the assumptions on η into (ii) yields

$$\begin{aligned}\langle \eta,\, \mu \rangle &\leqslant \sum_i |\mu| \left(\bigcup_i B_\varepsilon(x_i) \right) - c_2 \int_{B_\varepsilon(x_i)} |x - x_i|^2 \,\mathrm{d}\,|\mu|\,(x) + (1 - c_0)\,|\mu| \left(\mathcal{X}^{far} \right) \\ &= |\mu|\,(\mathcal{X}) - c_2 \sum_i \int_{B_\varepsilon(x_i)} |x - x_i|^2 \,\mathrm{d}\,|\mu|\,(x) - c_0\,|\mu| \left(\mathcal{X}^{far} \right)\end{aligned}$$

(iv) Combining (i) and (iii) yields the required conclusion. □

Outline

1. The sparse spikes problem
2. The BLASSO and dual certificates
3. Minimal norm certificate and support stability
4. Analysis of the minimal norm certificate
5. Recovery statements
6. Numerical algorithms

The minimal norm certificate

Checking the existence of a dual certificate which saturates only at X guarantees uniqueness of solutions to $\mathcal{P}_0(y)$ and to some extent, stability. However, for *support* stability, we need to consider the certificate of minimal norm [Duval & Peyré '15].

Minimal norm certificate

Given any μ^* solution to $(\mathcal{P}_0(y))$, define

$$p_0 \stackrel{\text{def.}}{=} \min \{\|p\| \ ; \ p \in (\mathcal{D}_0(y))\}$$

If p_0 exists, then we call it the *minimal norm certificate*

A key property is that it is the limit of the (unique) dual solutions of $(\mathcal{D}_\lambda(y))$ as $\lambda \to 0$.

Limit of p_λ

Lemma (Duval & Peyré '15)

Let p_λ be the solution to $(\mathcal{D}_\lambda(y))$. If p_0 exists, then $\|p_\lambda - p_0\| \to 0$ and $\eta_\lambda^{(k)} \to \eta_0^{(k)}$ uniformly for all k.

Proof.

Step 1, extract a weakly convergent subsequence: Since p_λ is a solution to $\mathcal{D}_\lambda(y)$, we have

$$\langle y, p_\lambda\rangle - \frac{\lambda}{2}\|p_\lambda\|^2 \geqslant \langle y, p_0\rangle - \frac{\lambda}{2}\|p_0\|^2, \tag{4.1}$$

and p_0 being a solution to $\mathcal{D}_0(y)$ implies that

$$\langle y, p_0\rangle \geqslant \langle y, p_\lambda\rangle.$$

Therefore, $\|p_0\| \geqslant \|p_\lambda\|$, and given $\lambda_n \to 0$, we may extract a subsequence such that $p_{\lambda_{n_k}}$ weakly converges to p_* for some $p_* \in \mathcal{H}$ (recall that the closed unit ball of a Hilbert space is weakly sequentially compact).

□

Limit of p_λ

Lemma (Duval & Peyré '15)

Let p_λ be the solution to $(\mathcal{D}_\lambda(y))$. If p_0 exists, then $\|p_\lambda - p_0\| \to 0$ and $\eta_\lambda^{(k)} \to \eta_0^{(k)}$ uniformly for all k.

Proof.

Step 1, extract a weakly convergent subsequence: Since p_λ is a solution to $\mathcal{D}_\lambda(y)$, we have

$$\langle y, p_\lambda \rangle - \frac{\lambda}{2} \|p_\lambda\|^2 \geqslant \langle y, p_0 \rangle - \frac{\lambda}{2} \|p_0\|^2 , \tag{4.1}$$

and p_0 being a solution to $\mathcal{D}_0(y)$ implies that

$$\langle y, p_0 \rangle \geqslant \langle y, p_\lambda \rangle.$$

Therefore, $\|p_0\| \geqslant \|p_\lambda\|$, and given $\lambda_n \to 0$, we may extract a subsequence such that $p_{\lambda_{n_k}}$ weakly converges to p_* for some $p_* \in \mathcal{H}$ (recall that the closed unit ball of a Hilbert space is weakly sequentially compact).

Step 2, the weak limit solves $(\mathcal{D}_0(y))$**:** Taking the limit of $\lambda \to 0$ in (3.1) yields $\langle y, p_* \rangle \geqslant \langle y, p_0 \rangle$.

Note that $\Phi^* p_{\lambda_{n_k}}$ converges weakly to $\Phi^* p$ in $\mathcal{C}(\mathcal{X})$, and so,

$$\|\Phi^* p\|_\infty \leqslant \liminf_k \left\|\Phi^* p_{\lambda_{n_k}}\right\|_\infty = 1.$$

Therefore, p_* solves $\mathcal{D}_0(y)$.

□

Limit of p_λ

Lemma (Duval & Peyré '15)

Let p_λ be the solution to $(\mathcal{D}_\lambda(y))$. If p_0 exists, then $\|p_\lambda - p_0\| \to 0$ and $\eta_\lambda^{(k)} \to \eta_0^{(k)}$ uniformly for all k.

Proof.

Step 3, the weak limit is the minimal norm solution: p_* is the solution of minimal norm since

$$\|p_*\| \leqslant \liminf_k \left\|p_{\lambda_{n_k}}\right\| \leqslant \|p_0\|,$$

and hence, $p_* = p_0$, $\left\|p_{\lambda_{n_k}}\right\| \to \|p_0\|$ and $p_{\lambda_{n_k}} \to p_0$ strongly in $\mathcal{H}$.

□

Limit of p_λ

Lemma (Duval & Peyré '15)

Let p_λ be the solution to $(\mathcal{D}_\lambda(y))$. If p_0 exists, then $\|p_\lambda - p_0\| \to 0$ and $\eta_\lambda^{(k)} \to \eta_0^{(k)}$ uniformly for all k.

Proof.

Step 3, the weak limit is the minimal norm solution: p_* is the solution of minimal norm since

$$\|p_*\| \leqslant \liminf_k \left\|p_{\lambda_{n_k}}\right\| \leqslant \|p_0\|,$$

and hence, $p_* = p_0$, $\left\|p_{\lambda_{n_k}}\right\| \to \|p_0\|$ and $p_{\lambda_{n_k}} \to p_0$ strongly in $\mathcal{H}$.

Step 4, full convergence: This implies $\lim_{\lambda\to 0} \|p_\lambda - p_0\| = 0$, since otherwise, we can extract a subsequence p_{λ_k} such that $\left\|p_{\lambda_k} - p_0\right\| > \varepsilon$ and by the above argument, extract a further subsequence which converges strongly to p_0. Finally, for the convergence of $\eta_\lambda^{(k)}$, note that

$$\left|\eta_\lambda^{(k)}(x) - \eta_0^{(k)}(x)\right| \leqslant \left\|\varphi^{(k)}\right\|_\infty \|p_\lambda - p_0\| \to 0, \qquad \lambda \to 0.$$

□

Exact support stability under small noise

Suppose that $\mu_0 = \sum_{i=1}^{s} a_j \delta_{x_j}$.

Theorem (Duval & Peyré '15)

Suppose that η_0 is nondegenerate, there exists r, λ_0, c_0 such that for all $\lambda \leqslant \lambda_0$ and $\|w\| \leqslant c_0\lambda$, any solution $\mu_{\lambda,w}$ of $(\mathcal{P}_\lambda(y))$ with $y = \Phi\mu_0 + w$ has support contained in $\bigcup_{i=1}^{s} B_r(x_i)$. Moreover, if μ_0 is identifiable, then $\mu_{\lambda,w}$ consist of exactly s spikes.

Proof.

- Note that since the solution to $\mathcal{D}_\lambda(y)$ is the projection onto a closed convex set, we have

$$\left\|p_{\lambda,0} - p_{\lambda,w}\right\| \leqslant \frac{\|w\|}{\lambda}.$$

- Suppose that $\eta_0''(x) \neq 0$ in $x \in B_r(x_j)$, $j = 1, \ldots, s$, and $|\eta_0(x)| < 1$ for $x \notin \cup_j B_r(x_j)$. Then, for all $\varepsilon > 0$, for all λ and $\|w\|/\lambda$ sufficiently small, $\left|\eta_0^{(k)} - \eta_{\lambda,w}^{(k)}\right| < \varepsilon$ for $k \in \{0, 2\}$.

- Therefore, $\eta_{\lambda,w}$ is such that $\left|\eta_{\lambda,w}^{(2)}(x)\right| \neq 0$ in $B_r(x_j)$ for each j and $\left|\eta_{\lambda,w}(x)\right| < 1$ for $x \notin \cup_j B_r(x_j)$. So, there exists at most 1 point in $B_\varepsilon(x_j)$ for which $\left|\eta_{\lambda,w}\right| = 1$.
- But if $\mathcal{P}_0$ has a unique solution μ_0, then we know that $\mu_{\lambda,w}$ converges in the weak-* topology as $\lambda, \|w\| \to 0$. Therefore $\mu_{\lambda,w}(B_r(x_j)) \to \mu_0(B_r(x_j)) \neq 0$ and hence, for λ, w sufficiently small, $\mu_{\lambda,w}$ has exactly one spike in $B_r(x_j)$.

□

Exact support stability under small noise

In fact, the following (stronger) result holds:

Theorem (Duval & Peyré '15)

Suppose that η_0 is nondegenerate and μ_0 is identifiable, then there exists λ_*, c_* such that for all $\lambda \leqslant \lambda_*$ and $\|w\| \leqslant c_* \lambda$, $\mathcal{P}_\lambda(y)$ has a unique solution which consists of precisely s spikes. Writing $v = (\lambda, w)$, we have $\mu_v = \sum_{i=1}^s a_i^v \delta_{x_i^v}$. the mapping $v \mapsto (a^v, X^v)$ is $\mathcal{C}^1$ and

$$\|a^v - a_0\| + \|X^v - X_0\| \leqslant C(\lambda + \|w\|).$$

Summary

- The extremal points of solutions to the dual problem inform on the support of the primal solutions.
- Existence of a nondegenerate dual certificate guarantees exact recovery in the noiseless setting, and support clustering stability in the noisy setting.
- For support stability, we look to a special solution of $\mathcal{D}_0(y)$, the one of minimal norm.
- the MNC is the limit of p_λ and so, it informs on the support of μ_λ for λ small.

Outline

1 The sparse spikes problem

2 The BLASSO and dual certificates

3 Minimal norm certificate and support stability

4 Analysis of the minimal norm certificate

5 Recovery statements

6 Numerical algorithms

Vanishing derivatives precertificate

We need to find $\eta = \Phi^* p$ such that $\eta(x_i) = \text{sign}(a_i)$ for all i and $\|\eta\|_\infty \leqslant 1$. What is a good (closed form) candidate?

Minimal norm certificate

Consider instead the vanishing derivatives precertificate, defined as $\eta_V = \Phi^* p_V$ with

$$p_V = \text{argmin}\left\{\|p\| \ ; \forall i,\ (\Phi^* p)(x_i) = \text{sign}(a_i) \quad \text{and} \quad \|\Phi^* p\|_\infty \leqslant 1\right\}.$$

Vanishing derivatives precertificate

We need to find $\eta = \Phi^* p$ such that $\eta(x_i) = \mathrm{sign}(a_i)$ for all i and $\|\eta\|_\infty \leqslant 1$. What is a good (closed form) candidate?

Vanishing derivatives precertificate

Consider instead the vanishing derivatives precertificate, defined as $\eta_V = \Phi^* p_V$ with

$$p_V = \operatorname{argmin}\left\{\|p\| \;;\; \forall i,\ (\Phi^* p)(x_i) = \mathrm{sign}(a_i) \quad \text{and} \quad \nabla(\Phi^* p)(x_i) = 0\right\}.$$

Closed form expression: The constraint consists of $(d+1)s$ equations. Writing

$$\Gamma_X : \mathbb{R}^{2s} \to \mathcal{H}, \quad \binom{a}{b} \mapsto \sum_j a_j \varphi(x_j) + b_j \varphi'(x_j)$$

the constraints can be written as $\Gamma_X^* p = \binom{\mathrm{sign}(a)}{0}$ and so, $p_V = \Gamma_X^{*,\dagger} \binom{\mathrm{sign}(a)}{0}$.

Vanishing derivatives precertificate

We need to find $\eta = \Phi^* p$ such that $\eta(x_i) = \text{sign}(a_i)$ for all i and $\|\eta\|_\infty \leqslant 1$. What is a good (closed form) candidate?

Vanishing derivatives precertificate

Consider instead the vanishing derivatives precertificate, defined as $\eta_V = \Phi^* p_V$ with

$$p_V = \operatorname{argmin}\left\{\|p\| \;;\; \forall i,\ (\Phi^* p)(x_i) = \text{sign}(a_i) \quad \text{and} \quad \nabla(\Phi^* p)(x_i) = 0\right\}.$$

Closed form expression: The constraint consists of $(d+1)s$ equations. Writing

$$\Gamma_X : \mathbb{R}^{2s} \to \mathcal{H}, \quad \begin{pmatrix} a \\ b \end{pmatrix} \mapsto \sum_j a_j \varphi(x_j) + b_j \varphi'(x_j)$$

the constraints can be written as $\Gamma_X^* p = \binom{\text{sign}(a)}{0}$ and so, $p_V = \Gamma_X^{*,\dagger} \binom{\text{sign}(a)}{0}$.

Kernel expression: Writing the covariance kernel $K(x, x') \stackrel{\text{def.}}{=} \langle \varphi(x), \varphi(x') \rangle$, we have

$$\eta_V(x) = \sum_{i=1}^N \alpha_i K(x_i, x) + \sum_{i=1}^N \beta_i \partial_1 K(x_i, x), \qquad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = D_{K,X}^{-1} \begin{pmatrix} \text{sign}(a) \\ 0_N \end{pmatrix}$$

with correlation kernel $K(x, x') = \langle \varphi(x), \varphi(x') \rangle$, $D_{K,X} \stackrel{\text{def.}}{=} \begin{pmatrix} M_0, & M_1 \\ M_1^* & M_2 \end{pmatrix}$,

$$\text{where} \quad M_0 = (K(x_i, x_j))_{i,j}, \quad M_1 = (\partial_1 K(x_i, x_j))_{i,j}, \quad M_2 = (\partial_1 \partial_2 K(x_i, x_j))_{i,j}.$$

Necessity of η_V

η_V coincides with the minimal norm certificate if $\|\eta_V\|_\infty \leqslant 1$ and is necessarily a valid certificate if there is support stability.

Given $X = \{x_j\}_{j=1}^s$, define $\Gamma : \mathbb{R}^{2s} \to \mathcal{H}$ by $\Gamma_X \binom{a}{b} = \sum_j a_j \varphi(x_j) + b_j \varphi'(x_j)$.

Lemma

Let $X_0 = \{x_{0,i}\}_{i=1}^s$ and Suppose that $\mu_0 = \sum_{i=1}^s a_{0,i} \delta_{x_{0,i}}$ and Γ_{X_0} is full rank. Suppose that there exists a $\mathcal{C}^1$ path $g : [0, \lambda_0) \to \mathbb{R}^s \times \mathcal{X}^s$, $\lambda \mapsto (a_\lambda, X_\lambda)$ such that $\mu_\lambda \overset{\text{def.}}{=} \sum_{i=1}^s a_{\lambda,i} x_{\lambda,i}$ solves $(\mathcal{P}_\lambda(y))$ with $y = \Phi\mu_0$. Then $\eta_V = \eta_0$.

Typical strategy: compute some η_V based on a correlation kernel K, then check that it is nondegenerate.

Proof of necessity of η_V (bc $\lim_{\lambda\to 0} p_\lambda = p_V$)

Given $\lambda \in [0, \lambda_0)$, let $(a, X) = g(\lambda)$. For all λ sufficiently small, we have $\text{sign}(a) = \text{sign}(a_0)$ by continuity of g, and recall that $p_\lambda = \frac{1}{\lambda}\left(\Phi_{X_0} a_0 - \Phi_X a\right)$. Therefore,

$$\Gamma_X^* \left(\Phi_X a - \Phi_{X_0} a_0\right) + \lambda \begin{pmatrix} \text{sign}(a_0) \\ 0 \end{pmatrix} = 0.$$

Proof of necessity of η_V (bc $\lim_{\lambda\to 0} p_\lambda = p_V$)

Given $\lambda \in [0, \lambda_0)$, let $(a, X) = g(\lambda)$. For all λ sufficiently small, we have $\text{sign}(a) = \text{sign}(a_0)$ by continuity of g, and recall that $p_\lambda = \frac{1}{\lambda}\left(\Phi_{X_0} a_0 - \Phi_X a\right)$. Therefore,

$$\Gamma_X^* \left(\Phi_X a - \Phi_{X_0} a_0\right) + \lambda \binom{\text{sign}(a_0)}{0} = 0.$$

Applying $\Gamma_X (\Gamma_X^* \Gamma_X)^\dagger$ to both sides gives

$$\Gamma_X \binom{a}{0} - \Gamma_X \Gamma_X^\dagger \Gamma_{X_0} \binom{a_0}{0} + \lambda \Gamma_X^{*,\dagger} \binom{\text{sign}(a_0)}{0_s} = 0.$$

Let Π_X be the projection onto $\text{Im}(\Gamma_X)^\perp$. Then, $\Pi_X = (\text{Id} - \Gamma_X \Gamma_X^\dagger)$, so

$$\underbrace{\frac{1}{\lambda}\left(-\Phi_X a + \Phi_{X_0} a_0\right)}_{p_\lambda} - \frac{1}{\lambda} \Pi_X \Phi_{X_0} a_0 = \underbrace{\Gamma_X^{*,\dagger} \binom{\text{sign}(a_0)}{0_s}}_{\text{cvg. } \Gamma_{X_0}^{*,\dagger} \binom{\text{sign}(a_0)}{0_s} = p_V}.$$

Proof of necessity of η_V (bc $\lim_{\lambda\to 0} p_\lambda = p_V$)

Given $\lambda \in [0, \lambda_0)$, let $(a, X) = g(\lambda)$. For all λ sufficiently small, we have $\mathrm{sign}(a) = \mathrm{sign}(a_0)$ by continuity of g, and recall that $p_\lambda = \frac{1}{\lambda}\left(\Phi_{X_0} a_0 - \Phi_X a\right)$. Therefore,

$$\Gamma_X^* \left(\Phi_X a - \Phi_{X_0} a_0\right) + \lambda \begin{pmatrix} \mathrm{sign}(a_0) \\ 0 \end{pmatrix} = 0.$$

Applying $\Gamma_X (\Gamma_X^* \Gamma_X)^\dagger$ to both sides gives

$$\Gamma_X \begin{pmatrix} a \\ 0 \end{pmatrix} - \Gamma_X \Gamma_X^\dagger \Gamma_{X_0} \begin{pmatrix} a_0 \\ 0 \end{pmatrix} + \lambda \Gamma_X^{*,\dagger} \begin{pmatrix} \mathrm{sign}(a_0) \\ 0_s \end{pmatrix} = 0.$$

Let Π_X be the projection onto $\mathrm{Im}(\Gamma_X)^\perp$. Then, $\Pi_X = (\mathrm{Id} - \Gamma_X \Gamma_X^\dagger)$, so

$$\underbrace{\frac{1}{\lambda}\left(-\Phi_X a + \Phi_{X_0} a_0\right)}_{p_\lambda} - \frac{1}{\lambda} \Pi_X \Phi_{X_0} a_0 = \underbrace{\Gamma_X^{*,\dagger} \begin{pmatrix} \mathrm{sign}(a_0) \\ 0_s \end{pmatrix}}_{\text{cvg. } \Gamma_{X_0}^{*,\dagger} \begin{pmatrix} \mathrm{sign}(a_0) \\ 0_s \end{pmatrix} = p_V}.$$

Since Π_X is a projection and $\Phi_{X_0} a_0 = \sum_j a_{0,j} \varphi(x_{0,j})$ is

$$\sum_j a_{0,j} \left(\varphi(x_j) + \varphi'(x_j)(x_j - x_{0,j}) + (x_j - x_{0,j})^2 \int_0^1 \varphi''(t(x_j - x_{0,j}))\mathrm{d}t \right),$$

we have

$$\frac{1}{\lambda} \left\| \Pi_X \Phi_{X_0} a_0 \right\| \leqslant \|a_0\|_\infty \left\|\varphi''\right\|_\infty \frac{1}{\lambda} \|X - X_0\|^2 \leqslant \|a_0\|_\infty \left\|\varphi''\right\|_\infty \frac{1}{\lambda} \|g(\lambda) - g(0)\|^2 \lesssim \lambda$$

since g is differentiable. Therefore, $\lim_{\lambda\to 0} p_\lambda = p_V$ and hence, $p_V = p_0$.

Examples

Consider

$$\varphi_k = \left(1 - \frac{|k|}{f_c + 1}\right) e^{2\pi i k \cdot} \quad \text{and} \quad \Phi\mu = (\langle \varphi_k, \mu \rangle)_{k=-f_c,\dots,f_c}.$$

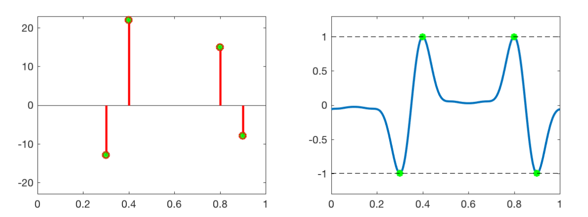
Solve

$$\min_{\mu} |\mu|(\mathbb{T}) + \frac{1}{2\lambda} \|\Phi\mu - y\|_2^2$$

where $y = \Phi\mu_0 + \varepsilon$.

- μ_0 consists of 4 spikes.
- Let $f_c = 10$, $\lambda = 10^{-3}$ and $\|\varepsilon\| = 10^{-4}\|y\|$.

η_V is nondegenerate



Reconstruction η_V

Examples

Consider

$$\varphi_k = \left(1 - \frac{|k|}{f_c + 1}\right) e^{2\pi i k \cdot} \quad \text{and} \quad \Phi\mu = (\langle \varphi_k,\, \mu\rangle)_{k=-f_c,\ldots,f_c}\,.$$

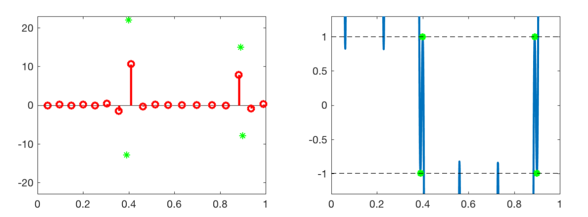
Solve

$$\min_{\mu} |\mu|\,(\mathbb{T}) + \frac{1}{2\lambda}\,\|\Phi\mu - y\|_2^2$$

where $y = \Phi\mu_0 + \varepsilon$.

- μ_0 consists of 4 spikes.
- Let $f_c = 10$, $\lambda = 10^{-3}$ and $\|\varepsilon\| = 10^{-4}\,\|y\|$.

η_V is degenerate



Reconstruction η_V

Examples

Consider

$$\varphi_k = \left(1 - \frac{|k|}{f_c + 1}\right) e^{2\pi i k \cdot} \quad \text{and} \quad \Phi\mu = (\langle \varphi_k, \mu \rangle)_{k=-f_c,\ldots,f_c}.$$

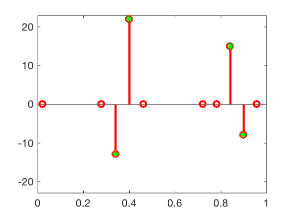
Solve

$$\min_{\mu} |\mu|(\mathbb{T}) + \frac{1}{2\lambda} \|\Phi\mu - y\|_2^2$$

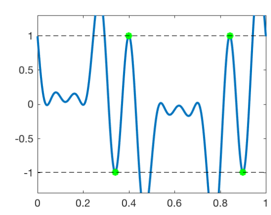
where $y = \Phi\mu_0 + \varepsilon$.

- μ_0 consists of 4 spikes.
- Let $f_c = 10$, $\lambda = 10^{-3}$ and $\|\varepsilon\| = 10^{-4}\|y\|$.

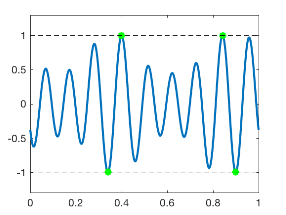
η_V is degenerate, but there exists a nondegenerate certificate:



Reconstruction



η_V



$\Phi^* p$

Examples

Consider

$$\varphi_k = \left(1 - \frac{|k|}{f_c + 1}\right) e^{2\pi i k \cdot} \quad \text{and} \quad \Phi\mu = \left(\langle \varphi_k, \mu \rangle\right)_{k=-f_c,\dots,f_c}.$$

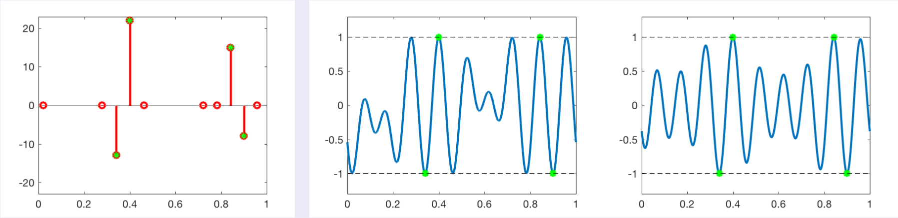
Solve

$$\min_{\mu} |\mu|(\mathbb{T}) + \frac{1}{2\lambda} \|\Phi\mu - y\|_2^2$$

where $y = \Phi\mu_0 + \varepsilon$.

- μ_0 consists of 4 spikes.
- Let $f_c = 10$, $\lambda = 10^{-3}$ and $\|\varepsilon\| = 10^{-4}\|y\|$.

η_V is degenerate, but there exists a nondegenerate certificate:



Reconstruction $\qquad$ η_0 $\qquad$ $\Phi^* p$

Key takeaway point

The vanishing derivatives certificate has a closed form expression and leads to an understanding of the recovery properties of the BLASSO.

- If $\|\eta_V\|_\infty > 1$, then no support stability is possible (arbitrarily small noise can lead to the appearance of spurious spikes).
- η_V nondegenerate implies support stability in the small noise regime, and unique recovery in the noiseless regime.
- η_V nondegenerate implies clustering stability in the large noise regime.

Key takeaway point

The vanishing derivatives certificate has a closed form expression and leads to an understanding of the recovery properties of the BLASSO.

- If $\|\eta_V\|_\infty > 1$, then no support stability is possible (arbitrarily small noise can lead to the appearance of spurious spikes).
- η_V nondegenerate implies support stability in the small noise regime, and unique recovery in the noiseless regime.
- η_V nondegenerate implies clustering stability in the large noise regime.

Next: precise recovery statements obtained via the analysis of vanishing derivatives certificates.

Outline

1 The sparse spikes problem

2 The BLASSO and dual certificates

3 Minimal norm certificate and support stability

4 Analysis of the minimal norm certificate

5 Recovery statements

6 Numerical algorithms

Sampling the Fourier transform

One of the seminal papers on the BLASSO is by Candès and Fernandez-Granda, *Towards a Mathematical Theory of Superresolution* published in CPAM, 2014.

Setting: We want to recover $\mu_{a,X} = \sum_j a_j \delta_{x_j}$ for $x_j \in \mathbb{T}$, from samples of its Fourier transform:

$$\Phi\mu \stackrel{\text{def.}}{=} \left\{ \langle e^{-i2\pi k \cdot}, \mu \rangle \; ; \; k \in \mathbb{Z}, |k| \leqslant f_c \right\}.$$

The minimum separation condition is defined as

$$\Delta(X) \stackrel{\text{def.}}{=} \min_{i \neq j} |x_i - x_j| \,.$$

Theorem (Candès & Fernandez-Granda '14)

Suppose that $\Delta(X) \geqslant \frac{C}{f_c}$. *Then,* $\mu_{a,X}$ *is the unique solution to* $(\mathcal{P}_0(y))$ *with* $y = \Phi\mu_{a,X}$.

- Here, $C > 0$ is a universal constant, $C \stackrel{\text{def.}}{=} 2$ in the original paper of Candès and Fernandez-Granda, with improvement to $C = 1.26$ by Fernandez-Granda in 2016.
- Since the proof constructs a nondegenerate dual certificate, "clustering stability" is also guaranteed in the noisy regime. Stability bounds on $\left\| \varphi_{\text{high}} \star (\hat{\mu} - \mu_{a,X}) \right\|_{L^1}$ are also possible.

Comment 1: the separation condition is necessary

In order to recover spikes of arbitrary signs, the minimum separation condition is necessary.

Comment 1: the separation condition is necessary

In order to recover spikes of arbitrary signs, the minimum separation condition is necessary.

To recover $\mu_{a,X}$, it is necessary that there exists a trigonometric function η which interpolates $\text{sign}(a)$ at points X.

Comment 1: the separation condition is necessary

In order to recover spikes of arbitrary signs, the minimum separation condition is necessary.

To recover $\mu_{a,X}$, it is necessary that there exists a trigonometric function η which interpolates $\mathrm{sign}(a)$ at points X.

Suppose that $|x_j - x_i| = \Delta$, $\mathrm{sign}(a_j) = 1$, $\mathrm{sign}(a_i) = -1$. Then, for some $x \in [x_i, x_j]$,

$$\eta(x_i) - \eta(x_j) = \eta'(x)(x_i - x_j)$$

Therefore,

$$\left|\eta'(x)\right| \geqslant \left|\frac{\eta(x_i) - \eta(x_j)}{(x_i - x_j)}\right| = \frac{2}{\Delta}.$$

The classical Bernstein's inequality asserts that for every trigonometric polynomial of degree at most f, $|q'(x)| \leqslant f \left\|q\right\|_\infty$. In our case, η is a trigonometric polynomial of degree $2f_c$. Therefore, we must have $\Delta \geqslant 1/f_c$.

Comment 1: the separation condition is necessary

In order to recover spikes of arbitrary signs, the minimum separation condition is necessary.

To recover $\mu_{a,X}$, it is necessary that there exists a trigonometric function η which interpolates $\text{sign}(a)$ at points X.

Suppose that $|x_j - x_i| = \Delta$, $\text{sign}(a_j) = 1$, $\text{sign}(a_i) = -1$. Then, for some $x \in [x_i, x_j]$,

$$\eta(x_i) - \eta(x_j) = \eta'(x)(x_i - x_j)$$

Therefore,

$$|\eta'(x)| \geqslant \left| \frac{\eta(x_i) - \eta(x_j)}{(x_i - x_j)} \right| = \frac{2}{\Delta}.$$

The classical Bernstein's inequality asserts that for every trigonometric polynomial of degree at most f, $|q'(x)| \leqslant f \|q\|_\infty$. In our case, η is a trigonometric polynomial of degree $2f_c$. Therefore, we must have $\Delta \geqslant 1/f_c$.

Remark

- Note that if the spikes are all positive, then it can be shown that the BLASSO does not require any separation [De Castro et al '12]
- For the arbitrary signs case, the separation condition is fundamental only for the BLASSO, it is known that other methods, such as Prony type methods do not require any separation.

Comment 2: analysis via the Jackson kernel

$$\eta_V(x) = \sum_{i=1}^{s} \alpha_i K(x_i, x) + \sum_{i=1}^{N} \beta_i \partial_1 K(x_i, x) \quad \text{where} \quad \binom{\alpha}{\beta} = D_{K,X} \binom{\text{sign}(a)}{0_S}$$

Comment 2: analysis via the Jackson kernel

$$\eta_V(x) = \sum_{i=1}^{s} \alpha_i K(x_i, x) + \sum_{i=1}^{N} \beta_i \partial_1 K(x_i, x) \quad \text{where} \quad \binom{\alpha}{\beta} = D_{K,X} \binom{\text{sign}(a)}{0_S}$$

Since $\varphi(x) = \left(e^{-i2\pi kx}\right)_{|k| \leqslant f_c}$, we have

$$K(x,y) = \sum_{|k| \leqslant f_c} e^{i2\pi k(x-y)} = \kappa(x-y)$$

where $\kappa(t) = \frac{\sin((2f_c+1)\pi t)}{(2f_c+1)\sin(\pi t)}$ is the Dirichlet kernel.

Comment 2: analysis via the Jackson kernel

$$\eta_V(x) = \sum_{i=1}^{s} \alpha_i K_{\mathrm{CF}}(x_i, x) + \sum_{i=1}^{N} \beta_i \partial_1 K_{\mathrm{CF}}(x_i, x) \quad \text{where} \quad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = D_{K_{\mathrm{CF}},X} \begin{pmatrix} \mathrm{sign}(a) \\ 0_S \end{pmatrix}$$

Since $\varphi(x) = \left(e^{-i2\pi kx}\right)_{|k|\leqslant f_c}$, we have

$$K(x,y) = \sum_{|k|\leqslant f_c} e^{i2\pi k(x-y)} = \kappa(x-y)$$

where $\kappa(t) = \frac{\sin((2f_c+1)\pi t)}{(2f_c+1)\sin(\pi t)}$ is the Dirichlet kernel.

The κ has slow decay $1/(1+f_c\,|t|)$, so it was proposed to replace κ by κ_{CF} (4th power of Dirichlet kernel):

$$K_{\mathrm{CF}}(x,x') = \kappa_{\mathrm{CF}}(x-x') = \left(\frac{\sin\left(\pi t \left(\frac{f_c}{2}+1\right)\right)}{\left(\frac{f_c}{2}+1\right)\sin(\pi t)} \right)^4 .$$

Comment 2: analysis via the Jackson kernel

$$\eta_V(x) = \sum_{i=1}^{s} \alpha_i K_{\mathrm{CF}}(x_i, x) + \sum_{i=1}^{N} \beta_i \partial_1 K_{\mathrm{CF}}(x_i, x) \quad \text{where} \quad \binom{\alpha}{\beta} = D_{K_{\mathrm{CF}},X} \binom{\mathrm{sign}(a)}{0_S}$$

Since $\varphi(x) = \left(e^{-i2\pi kx}\right)_{|k| \leqslant f_c}$, we have

$$K(x,y) = \sum_{|k| \leqslant f_c} e^{i2\pi k(x-y)} = \kappa(x-y)$$

where $\kappa(t) = \frac{\sin((2f_c+1)\pi t)}{(2f_c+1)\sin(\pi t)}$ is the Dirichlet kernel.

The κ has slow decay $1/(1+f_c\,|t|)$, so it was proposed to replace κ by κ_{CF} (4th power of Dirichlet kernel):

$$K_{\mathrm{CF}}(x,x') = \kappa_{\mathrm{CF}}(x-x') = \left(\frac{\sin\left(\pi t \left(\frac{f_c}{2}+1\right)\right)}{\left(\frac{f_c}{2}+1\right)\sin(\pi t)} \right)^4 .$$

- Under K_{CF}, η_V is still a trigonometric polynomial with frequencies $|k| \leqslant f_c$. So nondegeneracy of η_F still guarantees exact and stable clustering recovery.

Comment 2: analysis via the Jackson kernel

$$\eta_V(x) = \sum_{i=1}^{s} \alpha_i K_{\mathrm{CF}}(x_i, x) + \sum_{i=1}^{N} \beta_i \partial_1 K_{\mathrm{CF}}(x_i, x) \quad \text{where} \quad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = D_{K_{\mathrm{CF}}, X} \begin{pmatrix} \mathrm{sign}(a) \\ 0_S \end{pmatrix}$$

Since $\varphi(x) = \left(e^{-i2\pi kx}\right)_{|k| \leqslant f_c}$, we have

$$K(x, y) = \sum_{|k| \leqslant f_c} e^{i2\pi k(x-y)} = \kappa(x - y)$$

where $\kappa(t) = \frac{\sin((2f_c+1)\pi t)}{(2f_c+1)\sin(\pi t)}$ is the Dirichlet kernel.

The κ has slow decay $1/(1 + f_c |t|)$, so it was proposed to replace κ by κ_{CF} (4th power of Dirichlet kernel):

$$K_{\mathrm{CF}}(x, x') = \kappa_{\mathrm{CF}}(x - x') = \left(\frac{\sin\left(\pi t \left(\frac{f_c}{2} + 1\right)\right)}{\left(\frac{f_c}{2} + 1\right) \sin(\pi t)} \right)^4 .$$

- Under K_{CF}, η_V is still a trigonometric polynomial with frequencies $|k| \leqslant f_c$. So nondegeneracy of η_F still guarantees exact and stable clustering recovery.
- $K_{\mathrm{CF}}(x, x') = \langle \tilde{\varphi}(x), \tilde{\varphi}(x') \rangle$ with $\varphi_F(x) = \left(\sqrt{g_k} e^{-i2\pi kx}\right)_{|k| \leqslant f_c}$, for some appropriate weight g. So, the result of C-FG guarantees support stability for weighted Fourier sampling.

Extension: Subsampling

Observe $\left\{\langle e^{-i2\pi k\cdot}, \mu_0\rangle \; ; \; k \in \Omega\right\}$ where $\Omega \subset \left\{k \in \mathbb{Z} \; ; \; |k| \leqslant \frac{f_c}{2}\right\}$ drawn uniformly at random.

Theorem (Tang et al '13)

Let $\mu_0 = \sum_j a_j \delta_{x_j}$ *with* $\min_{i\neq j} |x_i - x_j| \geqslant C/f_c$*. Suppose that* $\operatorname{sign}(a)$ *is a Steinhaus sequence and*

$$|\Omega| \gtrsim \max\left(s \log(s/\delta) \log(f_c/\delta), \log(f_c/\delta)^2\right).$$

Then, w.p. at least $1-\delta$*,* μ_0 *can be exactly recovered from* $\mathcal{P}_0(y)$*.*

Convolution

Let Φ be a convolution operator $\Phi : \mathcal{M}(\mathcal{X};\mathbb{R}) \to L^2(\mathbb{R})$ with $\varphi(x) = t \mapsto \psi(t-x) \in L^2(\mathbb{R})$:

$$\Phi\mu = t \mapsto \int \psi(t-x)\mathrm{d}\mu(x).$$

Then,

$$K(x,x') \overset{\text{def.}}{=} \kappa(x-x'), \quad \text{where} \quad \kappa \overset{\text{def.}}{=} \psi \star \psi.$$

Convolution

Let Φ be a convolution operator $\Phi : \mathcal{M}(\mathcal{X};\mathbb{R}) \to L^2(\mathbb{R})$ with $\varphi(x) = t \mapsto \psi(t-x) \in L^2(\mathbb{R})$:

$$\Phi\mu = t \mapsto \int \psi(t-x)\mathrm{d}\mu(x).$$

Then,

$$K(x,x') \stackrel{\text{def.}}{=} \kappa(x-x'), \quad \text{where} \quad \kappa \stackrel{\text{def.}}{=} \psi \star \psi.$$

Let $\sigma \stackrel{\text{def.}}{=} 1/\sqrt{|\kappa''(0)|}$, and assume that for $p > \frac{1}{2}$, $r, b > 0$, we have

- Sufficient decay: for $k = 0, 1, 2, 3$, $\sigma^k \left|\kappa^{(k)}(t)\right| \leqslant \frac{A_k}{(1+Ct^2)^p}$.
- Sufficient peak: $\sigma^2\kappa''(t) < -b$, $\forall\, |t| < \sigma r$.

Convolution

Let Φ be a convolution operator $\Phi : \mathcal{M}(\mathcal{X};\mathbb{R}) \to L^2(\mathbb{R})$ with $\varphi(x) = t \mapsto \psi(t-x) \in L^2(\mathbb{R})$:

$$\Phi\mu = t \mapsto \int \psi(t-x)\mathrm{d}\mu(x).$$

Then,

$$K(x,x') \overset{\text{def.}}{=} \kappa(x-x'), \quad \text{where} \quad \kappa \overset{\text{def.}}{=} \psi \star \psi.$$

Let $\sigma \overset{\text{def.}}{=} 1/\sqrt{|\kappa''(0)|}$, and assume that for $p > \frac{1}{2}$, $r, b > 0$, we have

- Sufficient decay: for $k = 0, 1, 2, 3$, $\sigma^k \left|\kappa^{(k)}(t)\right| \leqslant \frac{A_k}{(1+Ct^2)^p}$.
- Sufficient peak: $\sigma^2 \kappa''(t) < -b$, $\forall\, |t| < \sigma r$.

Theorem ([Bendory et al '15])

Let $|x_i - x_j| > \Delta$ for all $i \neq j$, with $\Delta \overset{\text{def.}}{=} \frac{D}{\sqrt{C}}$. Then, η_V is nondegenerate. Here $D > 0$ is a constant which depends only on A_k, b, r and p.

Convolution

Let Φ be a convolution operator $\Phi : \mathcal{M}(\mathcal{X};\mathbb{R}) \to L^2(\mathbb{R})$ with $\varphi(x) = t \mapsto \psi(t-x) \in L^2(\mathbb{R})$:

$$\Phi\mu = t \mapsto \int \psi(t-x)\mathrm{d}\mu(x).$$

Then,

$$K(x,x') \overset{\text{def.}}{=} \kappa(x-x'), \quad \text{where} \quad \kappa \overset{\text{def.}}{=} \psi \star \psi.$$

Let $\sigma \overset{\text{def.}}{=} 1/\sqrt{|\kappa''(0)|}$, and assume that for $p > \frac{1}{2}$, $r, b > 0$, we have

- Sufficient decay: for $k = 0, 1, 2, 3$, $\sigma^k \left|\kappa^{(k)}(t)\right| \leqslant \frac{A_k}{(1+Ct^2)^p}$.
- Sufficient peak: $\sigma^2\kappa''(t) < -b$, $\forall\, |t| < \sigma r$.

Theorem ([Bendory et al '15])

Let $|x_i - x_j| > \Delta$ for all $i \neq j$, with $\Delta \overset{\text{def.}}{=} \frac{D}{\sqrt{C}}$. Then, η_V is nondegenerate. Here $D > 0$ is a constant which depends only on A_k, b, r and p.

- Gaussian kernel: $\psi(t) = \frac{1}{\sqrt[4]{\pi}\sqrt{\sigma}} \exp\left(-t^2\sigma^{-2}/2\right)$, then $\kappa(t) = \exp\left(-t^2\sigma^{-2}/4\right)$.
- Cauchy kernel: $\psi(t) = \frac{2}{\sqrt{\sigma\pi}(4t^2\sigma^{-2}+1)}$, then $\kappa(t) = \frac{1}{(t^2\sigma^{-2}+1)}$.

We have a scaling factor σ, but b, r, A_k and p can be chosen to to be constants independent of σ and $C \sim |\kappa''(0)| \sim \sigma^{-2}$. Therefore, η_V is nondegenerate provided that $\Delta \gtrsim \sigma$.

Summary

On conditions for recovery:

- The extremal points of solutions to the dual problem inform on the support of the primal solutions.
- Existence of a nondegenerate dual certificate guarantees exact recovery in the noiseless setting, and support clustering stability in the noisy setting.
- For support stability, we look to a special solution of $\mathcal{D}_0(y)$, the one of minimal norm $\eta_0 = \Phi^* p_0$.
- the MNC is the limit of p_λ and so, it informs on the support of μ_λ for λ small.

Analysis of dual certificates:

- To analyse the MNC, we typically look at the vanishing derivatives precertificate which has a closed form expression.
- $\eta_V = \eta_0$ when $\|\eta_V\|_\infty \leqslant 1$. In fact, we must have $\|\eta_V\|_\infty \leqslant 1$ if we expect support stability.
- To guarantee exact recovery of spikes of arbitrary signs, it is necessary that that the underlying positions satisfy a minimum separation condition.
  - Case of sampling Fourier coefficients from $[-f_c, f_c]$, need $\min_{i\neq j} |x_i - x_j| \gtrsim \frac{1}{f_c}$.
  - Case of Gaussian deconvolution with scaling σ need $\min_{i\neq j} |x_i - x_j| \gtrsim \sigma$.

Outline

1. The sparse spikes problem
2. The BLASSO and dual certificates
3. Minimal norm certificate and support stability
4. Analysis of the minimal norm certificate
5. Recovery statements
6. Numerical algorithms

Numerical algorithms for the BLASSO

$(\mathcal{P}_\lambda(y))$ is an optimisation problem over the set of measures. One straightforward way of solving $(\mathcal{P}_\lambda(y))$ is to simply discretize over a fine grid $X \stackrel{\text{def.}}{=} (x_j)_{j=1}^N \subset \mathcal{X}$, that is, solve

$$\min_{a\in\mathbb{R}^N} \lambda \|a\|_1 + \frac{1}{2} \|\Phi_X a - y\|^2$$

where $\Phi_X : \mathbb{R}^N \to \mathcal{H}$ is defined by $\Phi_X a = \sum_{j=1}^N a_j \varphi(x_j)$. This is then simply the LASSO and when $\mathcal{H}$ is a finite dimensional space, this can be solved by a wide range of first order methods, such as projected gradient descent.

$$a^{n+1} = \text{Prox}_{\gamma\lambda\|\cdot\|_1}(a^n - \gamma\Phi_X^*(\Phi_X a - y))$$

where $\text{Prox}_{\gamma\lambda\|\cdot\|_1} = \text{argmin}_z \frac{1}{2}\|z-x\|_2^2 + \lambda\gamma\|x\|_1$.

Numerical algorithms for the BLASSO

$(\mathcal{P}_\lambda(y))$ is an optimisation problem over the set of measures. One straightforward way of solving $(\mathcal{P}_\lambda(y))$ is to simply discretize over a fine grid $X \stackrel{\text{def.}}{=} (x_j)_{j=1}^N \subset \mathcal{X}$, that is, solve

$$\min_{a\in\mathbb{R}^N} \lambda \|a\|_1 + \frac{1}{2}\|\Phi_X a - y\|^2$$

where $\Phi_X : \mathbb{R}^N \to \mathcal{H}$ is defined by $\Phi_X a = \sum_{j=1}^N a_j \varphi(x_j)$. This is then simply the LASSO and when $\mathcal{H}$ is a finite dimensional space, this can be solved by a wide range of first order methods, such as projected gradient descent.

$$a^{n+1} = \mathrm{Prox}_{\gamma\lambda\|\cdot\|_1}(a^n - \gamma\Phi_X^*(\Phi_X a - y))$$

where $\mathrm{Prox}_{\gamma\lambda\|\cdot\|_1} = \mathrm{argmin}_z \frac{1}{2}\|z-x\|_2^2 + \lambda\gamma\|x\|_1$.

Other approach which are better aligned to the infinite dimensional nature of $(\mathcal{P}_\lambda(y))$ include SDP approaches/Lasserre hierarchies (for certain measurements, e.g. Fourier) or the Frank-Wolfe/conditional gradient algorithm.

SDP approach of Candès and Fernandez-Granda

Let us consider the case where we observe Fourier coefficients up to some cut-off $f_c \in \mathbb{N}$. Let $n = 2f_c + 1$. The dual to $\mathcal{P}_\lambda(y)$ is a finite dimensional problem:

$$\max_{c \in \mathbb{C}^n} \mathrm{Re}\langle y,\, c\rangle - \frac{\lambda}{2} \|c\|^2 \text{ subject to } \|\mathcal{F}_n^* c\|_\infty \leqslant 1$$

where

$$\mathcal{F}_n^* c(t) = \sum_{|k| \leqslant f_c} c_k e^{i2\pi k t}.$$

SDP approach of Candès and Fernandez-Granda

Let us consider the case where we observe Fourier coefficients up to some cut-off $f_c \in \mathbb{N}$. Let $n = 2f_c + 1$. The dual to $\mathcal{P}_\lambda(y)$ is a finite dimensional problem:

$$\max_{c \in \mathbb{C}^n} \mathrm{Re}\langle y,\, c\rangle - \frac{\lambda}{2} \|c\|^2 \;\text{ subject to }\; \|\mathcal{F}_n^* c\|_\infty \leqslant 1$$

where

$$\mathcal{F}_n^* c(t) = \sum_{|k| \leqslant f_c} c_k e^{i2\pi kt}.$$

Theorem (Dumitrescu)

A causal trigonometric polynomial $p(t) \stackrel{\text{def.}}{=} \sum_{k=0}^{n-1} c_k e^{i2\pi kt}$ *with* $c \in \mathbb{C}^n$ *is bounded by 1 in magnitude iff there exists* $Q \in \mathbb{C}^{n\times n}$ *Hermitian s.t.*

$$0 \preceq \begin{pmatrix} Q & c \\ c^* & 1 \end{pmatrix} \quad \text{and} \quad \sum_{i=1}^{n-j} Q_{i,i+j} = \delta_{0,j}, \quad j = 1, \ldots, n-1, \tag{7.1}$$

where $\delta_{0,j} = 1$ *if* $j = 0$ *and* 0 *otherwise.*

SDP approach of Candès and Fernandez-Granda

Let us consider the case where we observe Fourier coefficients up to some cut-off $f_c \in \mathbb{N}$. Let $n = 2f_c + 1$. The dual to $\mathcal{P}_\lambda(y)$ is a finite dimensional problem:

$$\max_{c \in \mathbb{C}^n} \mathrm{Re}\langle y,\, c\rangle - \frac{\lambda}{2}\,\|c\|^2 \;\text{ subject to }\; \|\mathcal{F}_n^* c\|_\infty \leqslant 1$$

where

$$\mathcal{F}_n^* c(t) = \sum_{|k| \leqslant f_c} c_k e^{i2\pi kt}.$$

Theorem (Dumitrescu)

A causal trigonometric polynomial $p(t) \overset{\text{def.}}{=} \sum_{k=0}^{n-1} c_k e^{i2\pi kt}$ *with* $c \in \mathbb{C}^n$ *is bounded by 1 in magnitude iff there exists* $Q \in \mathbb{C}^{n\times n}$ *Hermitian s.t.*

$$0 \preceq \begin{pmatrix} Q & c \\ c^* & 1 \end{pmatrix} \quad \text{and} \quad \sum_{i=1}^{n-j} Q_{i,i+j} = \delta_{0,j}, \quad j = 1, \ldots, n-1, \tag{7.1}$$

where $\delta_{0,j} = 1$ *if* $j = 0$ *and* 0 *otherwise.*

One direction is easy to see: since $\langle z, \begin{pmatrix} Q & c \\ c^* & 1 \end{pmatrix} z\rangle \geqslant 0$, choose $z = (x^\top, \langle x,\, c\rangle)^\top$. Then, $x^* Q x - |\langle x,\, c\rangle| \geqslant 0$. Choosing $x = \left(e^{2\pi ikt}\right)_{k=0}^n$, we have $|p(t)| \leqslant x^* Q x$. The constraint on the diagonals of Q implies that $x^* Q x = 1$.

SDP approach of Candès and Fernandez-Granda

Note that $e^{i2\pi f_c t}(\mathcal{F}_n^* c)(t)$ is a causal trigonometric polynomial. This observation allows $(\mathcal{D}_\lambda(y))$ to be formulated as a SDP problem, as the dual problem becomes
Step I:

$$\max_{c \in \mathbb{C}^n, Q \in \mathbb{C}^{n \times n}} \mathrm{Re}\langle y, c\rangle - \frac{\lambda}{2} \|c\|^2 \text{ subject to (6.1)}$$

This is a finite dimensional semidefinite program.

SDP approach of Candès and Fernandez-Granda

Note that $e^{i2\pi f_c t}(\mathcal{F}_n^* c)(t)$ is a causal trigonometric polynomial. This observation allows $(\mathcal{D}_\lambda(y))$ to be formulated as a SDP problem, as the dual problem becomes
Step I:

$$\max_{c\in\mathbb{C}^n, Q\in\mathbb{C}^{n\times n}} \mathrm{Re}\langle y, c\rangle - \frac{\lambda}{2}\|c\|^2 \text{ subject to (6.1)}$$

This is a finite dimensional semidefinite program.

To find the solution to the primal problem, note that $\mathcal{F}_n^* c$ achieves its extremal points on the support of μ.

SDP approach of Candès and Fernandez-Granda

Note that $e^{i2\pi f_c t}(\mathcal{F}_n^* c)(t)$ is a causal trigonometric polynomial. This observation allows $(\mathcal{D}_\lambda(y))$ to be formulated as a SDP problem, as the dual problem becomes
Step I:

$$\max_{c\in\mathbb{C}^n, Q\in\mathbb{C}^{n\times n}} \mathrm{Re}\langle y,\, c\rangle - \frac{\lambda}{2}\left\|c\right\|^2 \text{ subject to (6.1)}$$

This is a finite dimensional semidefinite program.

To find the solution to the primal problem, note that $\mathcal{F}_n^* c$ achieves its extremal points on the support of μ.

To locate these extremal points:

- $$p_{2n-2}(e^{i2\pi t}) = 1 - \left|(\mathcal{F}_n^* c)(t)\right|^2 = 1 - \sum_{|k|\leqslant 2f_c} u_k e^{i2\pi kt} \quad \text{where} \quad u_k = \sum_j c_j \bar{c}_{j-k}.$$
- $z^{2f_c} p_{2n-2}(z)$ is a polynomial of degree $2n-2 = 4f_c$ and has the same roots as p_{2n-2} (ignoring $z=0$).
- $p_{2n-2}(e^{i2\pi t})$ has at most $2n-2$ roots.
- $p_{2n-2}(e^{i2\pi t})$ is real-valued and nonnegative, so it cannot have single roots on the unit circle. i.e. either $p_{2n-2}(e^{i2\pi t}) = 0$ or there are at most $n-1$ roots on the unit circle.

SDP approach of Candès and Fernandez-Granda

Step I:

$$\max_{c,Q} \mathrm{Re}\langle y,\, c\rangle - \frac{\lambda}{2}\,\|c\|^2 \text{ subject to}$$

$$0 \preceq \begin{pmatrix} Q & c \\ c^* & 1 \end{pmatrix} \quad \text{and} \quad \sum_{i=1}^{n-j} Q_{i,i+j} = \delta_{0,j}, \quad j = 1, \ldots, n-1,$$

Step II: Find the support $\hat{X}$ of μ by locating the roots of p_{2n-2} on the unit circle (eigenvalues of its companion matrix).
Step III: After finding the support $\hat{X}$, solve $\sum_{t \in \hat{X}} e^{-i2\pi kt} a_t = y_k$ to recover the amplitudes a (we have at most $n-1$ unknowns and n observed values in y).

Check this out later: `http://nbviewer.jupyter.org/github/gpeyre/numerical-tours/blob/master/matlab/sparsity_8_sparsespikes_measures.ipynb`

The multivariate setting

For the multivariate case when $d > 1$, one needs to make use of a so-called Lasserre Hierarchy. Consider the semidefinite relaxation of order m with $m \geqslant n = 2f_c + 1$:

$$\max_{c \in \mathbb{C}^{n^d}, Q \in \mathbb{C}^{n^d \times n^d}} \operatorname{Re}\langle y, c \rangle$$

$$\text{subject to} \begin{cases} \text{(i)} \;\; 0 \preceq \begin{bmatrix} Q & \tilde{c} \\ \tilde{c}^* & 1 \end{bmatrix} \\ \qquad \text{where} \quad \tilde{c}_k = \begin{cases} c_k & k \in [-f_c, f_c]^d \\ 0 & \text{otherwise} \end{cases} \\ \text{(ii)} \;\; \operatorname{Trace} \Theta_k Q = \delta_{0,k}, \qquad k \in (-m, m)^d \cap \mathbb{Z}, \end{cases} \qquad (\hat{\mathcal{D}}_{\lambda,m}(y))$$

where $\Theta_k \stackrel{\text{def.}}{=} \theta_{k_d} \otimes \cdots \otimes \theta_{k_1}$ with $\otimes$ denoting the Kronecker product and θ_{k_j} denoting the $m \times m$ Toeplitz matrix with ones on its k_j^{th} diagonal and zeros elsewhere.

The multivariate setting

For the multivariate case when $d > 1$, one needs to make use of a so-called Lasserre Hierarchy. Consider the semidefinite relaxation of order m with $m \geqslant n = 2f_c + 1$:

$$\max_{c \in \mathbb{C}^{n^d}, Q \in \mathbb{C}^{n^d \times n^d}} \operatorname{Re}\langle y, c\rangle$$

$$\text{subject to } \begin{cases} \text{(i)} \;\; 0 \preceq \begin{bmatrix} Q & \tilde{c} \\ \tilde{c}^* & 1 \end{bmatrix} \\ \qquad \text{where} \quad \tilde{c}_k = \begin{cases} c_k & k \in [-f_c, f_c]^d \\ 0 & \text{otherwise} \end{cases} \\ \text{(ii)} \;\; \operatorname{Trace}\Theta_k Q = \delta_{0,k}, \qquad k \in (-m, m)^d \cap \mathbb{Z}, \end{cases} \tag{$\hat{\mathcal{D}}_{\lambda,m}(y)$}$$

where $\Theta_k \stackrel{\text{def.}}{=} \theta_{k_d} \otimes \cdots \otimes \theta_{k_1}$ with $\otimes$ denoting the Kronecker product and θ_{k_j} denoting the $m \times m$ Toeplitz matrix with ones on its k_j^{th} diagonal and zeros elsewhere.

- It is known that $(\hat{\mathcal{D}}_{\lambda,m}(y))$ converges to $\mathcal{D}_\lambda(y)$ as $m \to +\infty$. If we have finite convergence, then the hierarchy is said to collapse.

The multivariate setting

For the multivariate case when $d > 1$, one needs to make use of a so-called Lasserre Hierarchy. Consider the semidefinite relaxation of order m with $m \geqslant n = 2f_c + 1$:

$$\max_{c \in \mathbb{C}^{n^d}, Q \in \mathbb{C}^{n^d \times n^d}} \operatorname{Re}\langle y, c\rangle$$

$$\text{subject to} \begin{cases} \text{(i)} \quad 0 \preceq \begin{bmatrix} Q & \tilde{c} \\ \tilde{c}^* & 1 \end{bmatrix} \\ \qquad \text{where} \quad \tilde{c}_k = \begin{cases} c_k & k \in [-f_c, f_c]^d \\ 0 & \text{otherwise} \end{cases} \\ \text{(ii)} \quad \text{Trace}\Theta_k Q = \delta_{0,k}, \qquad k \in (-m, m)^d \cap \mathbb{Z}, \end{cases} \qquad (\hat{\mathcal{D}}_{\lambda,m}(y))$$

where $\Theta_k \stackrel{\text{def.}}{=} \theta_{k_d} \otimes \cdots \otimes \theta_{k_1}$ with $\otimes$ denoting the Kronecker product and θ_{k_j} denoting the $m \times m$ Toeplitz matrix with ones on its k_j^{th} diagonal and zeros elsewhere.

- It is known that $(\hat{\mathcal{D}}_{\lambda,m}(y))$ converges to $\mathcal{D}_\lambda(y)$ as $m \to +\infty$. If we have finite convergence, then the hierarchy is said to collapse.
- In general, it is not know if we have finite convergence. However, as discussed above, in $d = 1$, this relaxation is tight in the sense that $(\hat{\mathcal{D}}_{\lambda,m}(y))$ is equivalent to $\mathcal{D}_\lambda(y)$ for any $m \geqslant n$. For $d = 2$, it is known that we have finite convergence for some $m \geqslant n$ (although in practice, it sufficies to take $m \geqslant n^2$.)

The multivariate setting

For the multivariate case when $d > 1$, one needs to make use of a so-called Lasserre Hierarchy. Consider the semidefinite relaxation of order m with $m \geqslant n = 2f_c + 1$:

$$\max_{c \in \mathbb{C}^{n^d}, Q \in \mathbb{C}^{n^d \times n^d}} \operatorname{Re}\langle y, c\rangle$$

$$\text{subject to} \begin{cases} \text{(i)} \;\; 0 \preceq \begin{bmatrix} Q & \tilde{c} \\ \tilde{c}^* & 1 \end{bmatrix} \\ \qquad \text{where} \quad \tilde{c}_k = \begin{cases} c_k & k \in [-f_c, f_c]^d \\ 0 & \text{otherwise} \end{cases} \\ \text{(ii)} \;\; \operatorname{Trace}\Theta_k Q = \delta_{0,k}, \qquad k \in (-m, m)^d \cap \mathbb{Z}, \end{cases} \qquad (\hat{\mathcal{D}}_{\lambda,m}(y))$$

where $\Theta_k \stackrel{\text{def.}}{=} \theta_{k_d} \otimes \cdots \otimes \theta_{k_1}$ with $\otimes$ denoting the Kronecker product and θ_{k_j} denoting the $m \times m$ Toeplitz matrix with ones on its k_j^{th} diagonal and zeros elsewhere.

- It is known that $(\hat{\mathcal{D}}_{\lambda,m}(y))$ converges to $\mathcal{D}_\lambda(y)$ as $m \to +\infty$. If we have finite convergence, then the hierarchy is said to collapse.
- In general, it is not know if we have finite convergence. However, as discussed above, in $d = 1$, this relaxation is tight in the sense that $(\hat{\mathcal{D}}_{\lambda,m}(y))$ is equivalent to $\mathcal{D}_\lambda(y)$ for any $m \geqslant n$. For $d = 2$, it is known that we have finite convergence for some $m \geqslant n$ (although in practice, it sufficies to take $m \geqslant n^2$.)
- To detect collapse of the hierarchy, it suffices to recover a measure $\mu_{\lambda,m}$ whose positions are the roots of $\Phi^* c$ which lie on the complex unit circle and amplitudes are found by solving the linear system of Step III above. If $\Phi^* c$ is a dual certificate to $\mu_{\lambda,m}$, then $\mu_{\lambda,m}$ is a solution to $(\mathcal{P}_\lambda(y))$.

Frank Wolfe

Frank-Wolfe algorithm aims to solve

$$\min_{m \in C} f(m) \tag{7.2}$$

where C is a weakly compact convex set of a Banach space, and f is a differentiable convex function.

Algorithm 1 Frank-Wolfe

for $k = 0, \ldots, n$ **do**
  $s^k \in \operatorname{argmin}_{s \in C} f(m^k) + \mathrm{d}f(m^k)(s - m^k)$
  if $\mathrm{d}f(m^k)(s^k - m^k) = 0$ **then** m^k is a solution. Stop.
  else
    $\gamma^k \leftarrow \frac{2}{k+2}$ or $\gamma^k \in \operatorname{argmin}_{\gamma \in [0,1]} f(m^k + \gamma(s^k - m^k))$
    $m^{k+1} \leftarrow m^k + \gamma^k(s^k - m^k)$
  end if
end for

Some comments on the Frank Wolfe algorithm

- The key advantage of this algorithm is that it is better suited to optimisation over Banach spaces as it does not rely on any underlying Hilbertian structure (for example, the proximal gradient decent algorithm involves a proximal term which is often in terms of the Euclidean distance), and only uses directional derivatives of f.

Some comments on the Frank Wolfe algorithm

- The key advantage of this algorithm is that it is better suited to optimisation over Banach spaces as it does not rely on any underlying Hilbertian structure (for example, the proximal gradient decent algorithm involves a proximal term which is often in terms of the Euclidean distance), and only uses directional derivatives of f.
- Note that given a differentiable convex function,

  $$f(x) \geqslant f(y) + \mathrm{d}f(y)(x - y)$$

  so the stopping criterion does ensure that m^k is a global minimiser, since minimality of s^k in step 2 implies that for all $s \in C$,

  $$f(s) \geqslant f(m^k) + \mathrm{d}f(m^k)(s - m^k) \geqslant f(m^k) + \mathrm{d}f(m^k)(s^k - m^k) = f(m^k).$$

Some comments on the Frank Wolfe algorithm

- The key advantage of this algorithm is that it is better suited to optimisation over Banach spaces as it does not rely on any underlying Hilbertian structure (for example, the proximal gradient decent algorithm involves a proximal term which is often in terms of the Euclidean distance), and only uses directional derivatives of f.
- Note that given a differentiable convex function,

$$f(x) \geqslant f(y) + \mathrm{d}f(y)(x-y)$$

so the stopping criterion does ensure that m^k is a global minimiser, since minimality of s^k in step 2 implies that for all $s \in C$,

$$f(s) \geqslant f(m^k) + \mathrm{d}f(m^k)(s-m^k) \geqslant f(m^k) + \mathrm{d}f(m^k)(s^k-m^k) = f(m^k).$$

- We remark that in line 6, we can replace m^{k+1} by any element of $\tilde{m} \in C$ such that $f(\tilde{m}) \leqslant f(m^{k+1})$ without adversely affecting the convergence properties of this algorithm.

Application of FW to our problem

In our setting, we are interested in recovering m as a measure, and $C \subseteq \mathcal{M}(\mathcal{X})$. In our case, we are interested in applying Frank-Wolfe to

$$f_\lambda(\mu) \stackrel{\text{def.}}{=} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu| (\mathcal{X}).$$

Application of FW to our problem

In our setting, we are interested in recovering m as a measure, and $C \subseteq \mathcal{M}(\mathcal{X})$. In our case, we are interested in applying Frank-Wolfe to

$$f_\lambda(\mu) \overset{\text{def.}}{=} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu| (\mathcal{X}).$$

Immediate problems:

1. f_λ is not differentiable
2. $\mathcal{M}(\mathcal{X})$ is unbounded.

Application of FW to our problem

In our setting, we are interested in recovering m as a measure, and $C \subseteq \mathcal{M}(\mathcal{X})$. In our case, we are interested in applying Frank-Wolfe to

$$f_\lambda(\mu) \stackrel{\text{def.}}{=} \frac{1}{2}\left\|\Phi\mu - y\right\|^2 + \lambda\left|\mu\right|(\mathcal{X}).$$

Immediate problems:

1. f_λ is not differentiable
2. $\mathcal{M}(\mathcal{X})$ is unbounded.

The following lemma allows us to rewrite minimisation of f_λ over $\mathcal{M}(\mathcal{X})$ into the form (6.2).

Lemma (Denoyelle et al '18)

μ_* is a minimiser of f_λ if and only if $(|\mu_*|(\mathcal{X}), \mu_*)$ minimises

$$\min_{(t,\mu)\in C} \tilde{f}_\lambda(\mu, t) \stackrel{\text{def.}}{=} \frac{1}{2}\left\|\Phi\mu - y\right\| + \lambda t$$

where $C \stackrel{\text{def.}}{=} \{(t,m) \in \mathbb{R}_+ \times \mathcal{M}(\mathcal{X})\,;\, |\mu|(\mathcal{X}) \leqslant t \leqslant M\}$ *and* $M \stackrel{\text{def.}}{=} \frac{\|y\|^2}{2\lambda}$.

Application of FW to our problem

In our setting, we are interested in recovering m as a measure, and $C \subseteq \mathcal{M}(\mathcal{X})$. In our case, we are interested in applying Frank-Wolfe to

$$f_\lambda(\mu) \stackrel{\text{def.}}{=} \frac{1}{2} \|\Phi\mu - y\|^2 + \lambda |\mu| (\mathcal{X}).$$

Immediate problems:

1. f_λ is not differentiable
2. $\mathcal{M}(\mathcal{X})$ is unbounded.

The following lemma allows us to rewrite minimisation of f_λ over $\mathcal{M}(\mathcal{X})$ into the form (6.2).

Lemma (Denoyelle et al '18)

μ_* is a minimiser of f_λ if and only if $(|\mu_*|(\mathcal{X}), \mu_*)$ minimises

$$\min_{(t,\mu)\in C} \tilde{f}_\lambda(\mu, t) \stackrel{\text{def.}}{=} \frac{1}{2} \|\Phi\mu - y\| + \lambda t$$

where $C \stackrel{\text{def.}}{=} \{(t, m) \in \mathbb{R}_+ \times \mathcal{M}(\mathcal{X}) \; ; \; |\mu| (\mathcal{X}) \leqslant t \leqslant M\}$ *and* $M \stackrel{\text{def.}}{=} \frac{\|y\|^2}{2\lambda}$.

Proof.

Note that if μ_* is a minimiser of f_λ, then $|\mu_*| (\mathcal{X}) \leqslant \frac{1}{\lambda} f_\lambda(\mu_*) \leqslant \frac{1}{\lambda} f_\lambda(0) \leqslant \frac{\|y\|}{2\lambda}$. Therefore, it suffices to minimise f_λ over all measure with $|\mu| (\mathcal{X}) \leqslant M$. It is then easy to check that μ_* minimises f_λ if and only if it minimises $\tilde{f}_\lambda$. □

Convergence of FW

Note that $\tilde{f}_\lambda$ is now differentiable over $\mathbb{R} \times \mathcal{M}(\mathcal{X})$ with $\mathrm{d}\tilde{f}_\lambda = (\lambda, \Phi^*(\Phi\mu - y))$, so

$$\mathrm{d}\tilde{f}_\lambda : (t', \mu') \mapsto \lambda t' + \int_{\mathcal{X}} \Phi^*(\Phi\mu - y)\mathrm{d}\mu'.$$

Moreover, even though C is not weakly compact, it is compact in the weak$*$ topology, and the convergence arguments for Algorithm 1 can be applied to conclude that

Lemma

Let (t^k, μ^k) be a sequence generated by Algorithm 1 applied to $\tilde{f}_\lambda$. Then, there exists $C > 0$ such that for any solution μ^* of $(\mathcal{P}_\lambda(y))$, we have

$$f_\lambda(\mu^k) - f_\lambda(\mu_*) \leqslant \frac{C}{k}.$$

Convergence of FW

As a corollary of this lemma, we have the following result, which shows under a nondegeneracy condition, μ^k increasingly clusters around the support of the solution μ^*.

Corollary

Suppose that $\mu_* \stackrel{\text{def.}}{=} \mu_{a,X} = \sum_i a_i \delta_{x_i}$ *is the unique solution to* $(\mathcal{P}_\lambda(y))$ *and* $\frac{1}{\lambda}\Phi^*(y - \Phi\mu_*)$ *is nondegenerate and satisfies the conditions of Theorem 2.2. Then,*

1. $\left|\mu^k\right| \left(\mathcal{X} \setminus \bigcup_i B_\varepsilon(x_i)\right) + \sum_{i=1}^s \int_{B_\varepsilon(x_i)} |x - x_i|^2 \, \mathrm{d} \left|\mu^k\right|(x) \lesssim \frac{1}{k}$.
2. *Suppose* Φ_X *is injective. Then,* $a_j^k \stackrel{\text{def.}}{=} \mu^k(B_\varepsilon(x_j))$ *satisfies* $\left\|a^k - a\right\|^2 \lesssim \frac{1}{k}$.

Convergence of FW

As a corollary of this lemma, we have the following result, which shows under a nondegeneracy condition, μ^k increasingly clusters around the support of the solution μ^*.

Corollary

Suppose that $\mu_* \stackrel{\text{def.}}{=} \mu_{a,X} = \sum_i a_i \delta_{x_i}$ *is the unique solution to* $(\mathcal{P}_\lambda(y))$ *and* $\frac{1}{\lambda}\Phi^*(y - \Phi\mu_*)$ *is nondegenerate and satisfies the conditions of Theorem 2.2. Then,*

1. $\left|\mu^k\right|\left(\mathcal{X} \setminus \bigcup_i B_\varepsilon(x_i)\right) + \sum_{i=1}^s \int_{B_\varepsilon(x_i)} |x - x_i|^2 \,\mathrm{d}\left|\mu^k\right|(x) \lesssim \frac{1}{k}$.
2. *Suppose* Φ_X *is injective. Then,* $a_j^k \stackrel{\text{def.}}{=} \mu^k(B_\varepsilon(x_j))$ *satisfies* $\left\|a^k - a\right\|^2 \lesssim \frac{1}{k}$.

Step 1, relate to Bregman distance

Let $r_k = f_\lambda(\mu^k) - f_\lambda(\mu_*)$. Let $F(\mu) \stackrel{\text{def.}}{=} \frac{1}{2\lambda}\left\|\Phi\mu - y\right\|^2$ and $J(\mu) \stackrel{\text{def.}}{=} |\mu|(\mathcal{X})$. Then, $f_\lambda = \lambda(J + F)$. By convexity of F,

$$\lambda^{-1} r_k \geqslant J(\mu^k) - J(\mu^*) + \langle F'(\mu^*),\, \mu^k - \mu^* \rangle.$$

Since $-F'(\mu^*) = \frac{1}{\lambda}\Phi^*(y - \Phi\mu_*) \in \partial J(\mu^*)$, and $-F'(\mu^*)$ is nondegenerate, by Theorem 2.2,

$$\lambda^{-1} r_k \geqslant c_0 \left|\mu^k\right|\left(\mathcal{X} \setminus \bigcup_i B_\varepsilon(x_i)\right) + c_2 \sum_{i=1}^s \int_{B_\varepsilon(x_i)} |x - x_i|^2 \,\mathrm{d}\left|\mu^k\right|(x).$$

Convergence of FW

Step 2, using injectivity of Φ_X

For the second claim, define

$$R(\nu) \overset{\text{def.}}{=} J(\nu) - J(\mu^*) + \langle F'(\mu^*), \nu - \mu^* \rangle \quad \text{and} \quad T(\nu) \overset{\text{def.}}{=} F(\nu) - F(\mu^*) - \langle F'(\mu^*), \nu - \mu^* \rangle.$$

- $R(\nu) \geqslant 0$ since $-F'(\mu^*) \in \partial J(\mu^*)$.
- $T(\nu) \geqslant 0$ by convexity of F.
- $\lambda^{-1} r_k = J(\mu^k) + T(\mu^k) \geqslant T(\mu^k)$.

Convergence of FW

Step 2, using injectivity of Φ_X

For the second claim, define

$$R(\nu) \overset{\text{def.}}{=} J(\nu) - J(\mu^*) + \langle F'(\mu^*), \nu - \mu^* \rangle \quad \text{and} \quad T(\nu) \overset{\text{def.}}{=} F(\nu) - F(\mu^*) - \langle F'(\mu^*), \nu - \mu^* \rangle.$$

- $R(\nu) \geqslant 0$ since $-F'(\mu^*) \in \partial J(\mu^*)$.
- $T(\nu) \geqslant 0$ by convexity of F.
- $\lambda^{-1} r_k = J(\mu^k) + T(\mu^k) \geqslant T(\mu^k)$.

Let $a_j^k = \mu^k(B_\varepsilon(x_j))$ and let $\hat{\mu}^k = \sum_j a_j^k \delta_{x_j}$. If Φ_X is injective with $\|\Phi_X a\|^2 \geqslant C \|a\|^2$, then

$$r_k \geqslant \lambda T(\mu^k) = \frac{\left\|\Phi(\mu^k - \mu^*)\right\|^2}{2} \geqslant \frac{3}{8} \left\|\Phi(\hat{\mu}^k - \mu^*)\right\|^2 - \frac{3}{2} \left\|\Phi(\hat{\mu}^k - \mu^k)\right\|^2$$
$$\geqslant \frac{3}{8} C \sum_k \left|a_j^k - a_j\right|^2 - \frac{3}{2} \left\|\Phi(\hat{\mu}^k - \mu^k)\right\|^2,$$

where we used $(a-b)^2/2 \geqslant 3a^2/8 - 3b^2/2$.

Step 3, bounding deviation of μ^k from its sparse projection

Finally, note that

$$\left\|\Phi(\hat{\mu}^k - \mu^k)\right\|^2 \leqslant \left\|\sum_j \int_{B_\varepsilon(x_j)} (\varphi(x) - \varphi(x_j))\mathrm{d}\mu^k(x) + \int_{\mathcal{X}^{far}} \varphi(x)\mathrm{d}\mu^k(x)\right\|^2$$

$$\leqslant 2\left(\sum_j \int_{B_\varepsilon(x_j)} \|\varphi'\|_\infty |x - x_j| \,\mathrm{d}\left|\mu^k\right|(x)\right)^2 + 2\left|\mu^k\right|(\mathcal{X}^{far})^2$$

$$\leqslant 2\left(\sum_j \|\varphi'\|_\infty \sqrt{\left|\mu^k\right|(B_\varepsilon(x_j)) \int_{B_\varepsilon(x_j)} |x - x_j|^2 \,\mathrm{d}\left|\mu^k\right|(x)}\right)^2 + 2\left|\mu^k\right|(\mathcal{X}^{far})^2$$

$$\leqslant 2\|\varphi'\|_\infty \left|\mu^k\right|(\mathcal{X}^{near})\left(\sum_j \int_{B_\varepsilon(x_j)} |x - x_j|^2 \,\mathrm{d}\left|\mu^k\right|(x)\right) + 2\left|\mu^k\right|(\mathcal{X}^{far})^2$$

$$\lesssim \lambda^{-1} c_2^{-1} r_k + \lambda^{-2} c_0^{-2} r_k^2.$$

Comments on lines 2 and 3 of Algorithm 1

- For step 2: Note that given $(t^k, \mu^k) \in C$, $s \mapsto \mathrm{d}\tilde{f}_\lambda(t^k, \mu^k)$ is a linear form, and since C is convex, it achieves its minimum at an extremal point of C. These extremal points are of the form $s = (M, \pm M\delta_x)$ with $x \in \mathcal{X}$.

Comments on lines 2 and 3 of Algorithm 1

- For step 2: Note that given $(t^k, \mu^k) \in C$, $s \mapsto \mathrm{d}\tilde{f}_\lambda(t^k, \mu^k)$ is a linear form, and since C is convex, it achieves its minimum at an extremal point of C. These extremal points are of the form $s = (M, \pm M\delta_x)$ with $x \in \mathcal{X}$. Therefore,

$$\begin{aligned} \mathrm{argmin}_{s\in C}\, \mathrm{d}\tilde{f}(t^k, m^k)(s) &= \mathrm{argmin}_{x\in\mathcal{X}}\, \pm M(\Phi^*(\Phi\mu^k - y))(x) + \lambda M \\ &= \mathrm{argmin}_{x\in\mathcal{X}}\, \pm\eta^k(x) + 1 \quad \text{where} \quad \eta^k \stackrel{\text{def.}}{=} \frac{1}{\lambda}\Phi^*(\Phi\mu^k - y) \\ &= \mathrm{argmax}_{x\in\mathcal{X}}\, \left|\eta^k(x)\right|. \end{aligned}$$

  Therefore, for each k, we introduce a new support point x_*^k, $s^k = (M, \sigma M\delta_{x_*^k})$ where $\left|\eta^k(x_*^k)\right| = \left\|\eta^k\right\|_\infty$.

Comments on lines 2 and 3 of Algorithm 1

- For step 2: Note that given $(t^k, \mu^k) \in C$, $s \mapsto \mathrm{d}\tilde{f}_\lambda(t^k, \mu^k)$ is a linear form, and since C is convex, it achieves its minimum at an extremal point of C. These extremal points are of the form $s = (M, \pm M\delta_x)$ with $x \in \mathcal{X}$. Therefore,

$$\begin{aligned}\mathrm{argmin}_{s\in C}\, \mathrm{d}\tilde{f}(t^k, m^k)(s) &= \mathrm{argmin}_{x\in\mathcal{X}}\, \pm M(\Phi^*(\Phi\mu^k - y))(x) + \lambda M \\ &= \mathrm{argmin}_{x\in\mathcal{X}}\, \pm\eta^k(x) + 1 \quad \text{where} \quad \eta^k \overset{\text{def.}}{=} \frac{1}{\lambda}\Phi^*(\Phi\mu^k - y) \\ &= \mathrm{argmax}_{x\in\mathcal{X}}\, \left|\eta^k(x)\right|.\end{aligned}$$

  Therefore, for each k, we introduce a new support point x_*^k, $s^k = (M, \sigma M\delta_{x_*^k})$ where $\left|\eta^k(x_*^k)\right| = \left\|\eta^k\right\|_\infty$.

- The halting condition of step 3 implies that μ^k is a minimiser of $(\mathcal{P}_\lambda(y))$ and hence, η^k is a dual certificate.

Comments on line 4 of Algorithm 1

If $\mu^k = \sum_{j=1}^k a_j^k \delta_{x_j^k}$, then the line search in step 4 is

$$\min_{\gamma} (1-\gamma) \left\| a^k \right\|_1 + \gamma M + \frac{1}{2} \left\| \Phi \mu_\gamma - y \right\|^2$$

where $\mu_\gamma = (1-\gamma) \sum_{j=1}^k a_j^k \delta_{x_j^k} + \gamma M \delta_{x_*^k}$.

- Note that since we can replace this step with any (t, μ) which improves the objective value, it seems sensible to simply perform in step 4

  $$\min_{a \in \mathbb{R}^{k+1}} \left\| a \right\|_1 + \frac{1}{2} \left\| \Phi_X a - y \right\|^2$$

  where $X = \{x_1^k, \ldots, x_k^k, x_*^k\}$. This is a finite dimensional nonsmooth convex optimisation problem and can be tackled using a variety of algorithms such as Forward Backward or FISTA.
- We can further improve the objective value by optimising over the positions as well [Bredies & Pikkarainnen '13, Boyd et al '17]
- More recently, [Denoyelle et al '18] proposed the sliding Frank-Wolfe algorithm, where step 4 is augmented to optimise over the positions and the amplitudes simultaneously. This minor modification in fact leads to *finite termination*.

Algorithm 2 Sliding Frank-Wolfe [Denoyelle et al '18]

Initialise with $m^0 = 0$.
for $k = 0, \dots, n$ **do**
    $\mu^k = \sum_{i=1}^{N^k} a_i^k \delta_{x_i^k}$, $a_i^k \in \mathbb{R}$, $x_i^k \in \mathcal{X}$ distinct, find $x_*^k \in \mathcal{X}$ s.t.

$$x_*^k \in \operatorname{argmin}_{x\in\mathcal{X}} \left|\eta^k(x)\right| \quad \text{where} \quad \eta^k \stackrel{\text{def.}}{=} \frac{1}{\lambda}\Phi^*(y - \Phi\mu^k).$$

    if then $\left|\eta^k(x_*^k)\right| \leqslant 1$
        μ^k is a solution. Stop.
    else
        $a_i^{k+\frac{1}{2}} \leftarrow \eta^k(x_*^k)$
        $m^{k+\frac{1}{2}} = \sum_{i=1}^{N^k} a_i^{k+\frac{1}{2}} \delta_{x_i^k} + a_i^{k+\frac{1}{2}} \delta_{x_*^k}$ s.t.

$$a^{k+\frac{1}{2}} \in \operatorname{argmin}_{a\in\mathbb{R}^{N^k+1}} \frac{1}{2}\left\|\Phi_{x^{k+\frac{1}{2}}} a - y\right\|^2 + \lambda \|a\|_1$$

where $x^{k+\frac{1}{2}} = (x_1^k, \cdots, x_{N^k}^k, x_*^k)$.
        $m^{k+1} = \sum_{i=1}^{N^k+1} a_i^{k+1} \delta_{x_i^{k+1}}$ s.t.

$$(a^{k+1}, x^{k+1}) \in \operatorname{argmin}_{(a,x)\in\mathbb{R}^{N^k}\times\mathcal{X}^{N^k+1}} \frac{1}{2}\|\Phi_x a - y\|^2 + \lambda \|a\|_1,$$

using a non-convex solver initialised with $(a^{k+\frac{1}{2}}, x^{k+\frac{1}{2}})$.
    end if
end for

Finite termination

Theorem (Denoyelle et al '18)

Let $\mu_{a,X} = \sum_i a_i \delta_{x_i}$ *be the unique solution to* $(\mathcal{P}_\lambda(y))$ *and suppose that* $\eta_\lambda = \frac{1}{\lambda}\Phi^*(y - \Phi\mu_{a,X})$ *is nondegenerate. Then, Algorithm 2 recovers* $\mu_{a,X}$ *after a finite number of steps.*

Sketch of proof.

Step 1, η^k converges to η_λ:

□

Finite termination

Theorem (Denoyelle et al '18)

Let $\mu_{a,X} = \sum_i a_i \delta_{x_i}$ *be the unique solution to* $(\mathcal{P}_\lambda(y))$ *and suppose that* $\eta_\lambda = \frac{1}{\lambda}\Phi^*(y - \Phi\mu_{a,X})$ *is nondegenerate. Then, Algorithm 2 recovers* $\mu_{a,X}$ *after a finite number of steps.*

Sketch of proof.

Step 1, η^k converges to η_λ:

- First note that μ^k converges to $\mu_{a,X}$ in the weak-$*$ topology.
- Since Φ is weak-$*$ to weak continuous, we have $p^k = \frac{1}{\lambda}(y - \Phi\mu^k)$ converges weakly to p_λ. Furthermore, p^k must be uniformly bounded in $\mathcal{H}$.
- This implies that the functions $\eta^k \stackrel{\text{def.}}{=} x \mapsto \langle \varphi(x), p^k \rangle$ are uniformly bounded and equicontinuous. So, by Arzela-Ascoli, we can extract a subsequence of η^k which converges to η_λ in L^∞ norm.

This is true also for the first and second derivatives of η^k.

□

Finite termination

Theorem (Denoyelle et al '18)

Let $\mu_{a,X} = \sum_i a_i \delta_{x_i}$ *be the unique solution to* $(\mathcal{P}_\lambda(y))$ *and suppose that* $\eta_\lambda = \frac{1}{\lambda}\Phi^*(y - \Phi\mu_{a,X})$ *is nondegenerate. Then, Algorithm 2 recovers* $\mu_{a,X}$ *after a finite number of steps.*

Sketch of proof.

Step 1, η^k converges to η_λ:
Step 2, η^k becomes a valid certificate in finite time: □

Finite termination

Theorem (Denoyelle et al '18)

Let $\mu_{a,X} = \sum_i a_i \delta_{x_i}$ *be the unique solution to* $(\mathcal{P}_\lambda(y))$ *and suppose that* $\eta_\lambda = \frac{1}{\lambda}\Phi^*(y - \Phi\mu_{a,X})$ *is nondegenerate. Then, Algorithm 2 recovers* $\mu_{a,X}$ *after a finite number of steps.*

Sketch of proof.

Step 1, η^k converges to η_λ:
Step 2, η^k becomes a valid certificate in finite time:

- Now, η_λ is nondegenerate implies that there exists a small neighbourhood around each x_i on which $\eta_\lambda'' \neq 0$. Therefore, there exists $\varepsilon > 0$ and $k_1 \in \mathbb{N}$ such that for all $k \geqslant k_1$, $(\eta^k)''(x) \neq 0$ for $x \in (x_i - \varepsilon, x_i + \varepsilon) \overset{\text{def.}}{=} I_{x_i,\varepsilon}$, and $|\eta^k(x)| < 1$ for all $x \notin \cup_i I_{x_i,\varepsilon}$. The optimality condition of step 8 is

$$0 \in \Phi_x^*(\Phi_x a - y) + \lambda\partial \|a\|_1 \quad \text{and} \quad \forall j,\ \langle(\Phi_x a - y), \varphi'(x_j)\rangle = 0.$$

  So, $\eta^k = -\frac{1}{\lambda}\Phi^*(\Phi_{x^k} a^k - y)$ satisfies $\eta^k(x_j^k) = \text{sign}(a_j^k)$ and $(\eta^k)'(x_j) = 0$. Hence, $|\eta^k(x)| < 1$ except at x^k.

□

Finite termination

Theorem (Denoyelle et al '18)

Let $\mu_{a,X} = \sum_i a_i \delta_{x_i}$ *be the unique solution to* $(\mathcal{P}_\lambda(y))$ *and suppose that* $\eta_\lambda = \frac{1}{\lambda}\Phi^*(y - \Phi\mu_{a,X})$ *is nondegenerate. Then, Algorithm 2 recovers* $\mu_{a,X}$ *after a finite number of steps.*

Sketch of proof.

Step 1, η^k converges to η_λ:
Step 2, η^k becomes a valid certificate in finite time: □

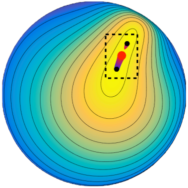
Remark

- Step 8 of Algorithm 2 requires solving a nonconvex optimisation problem, however, the proof utilises only the optimality condition of the optimisation problem and hence, finite convergence still holds even if we compute a stationary point.
- Under the nondegeneracy assumption, numerical observations suggest that we in fact have convergence in s iterations.

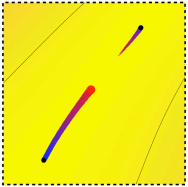
Example 1: nondegenerate case

Measurements: $y = \Phi m_0 + \lambda w$, where $w = \Phi \tilde{m}$, $\tilde{m} = \sum_{j=1}^{20} b_j \delta_{u_j}$, b is white noise with standard deviation 10^{-3}.

Let $\mathcal{X} = \left\{x \in \mathbb{R}^2 \; ; \; \|x\| \leqslant 1\right\}$. To model MEG/EEG, $\varphi(x) = u \mapsto \|x - u\|^{-2} \in \mathcal{H} = L^2(\partial\mathcal{X})$.



η_V and μ_λ

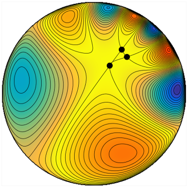


Zoom

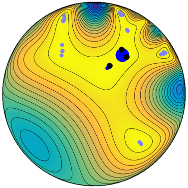
- Background image shows η_V
- Blue for $\lambda = 0$, Red for $\lambda = \lambda_{\max}$.

Example 2: Degenerate case

Measurements: $y = \Phi m_0 + \lambda w$, where $w = \Phi\tilde{m}$, $\tilde{m} = \sum_{j=1}^{20} b_j \delta_{u_j}$, b is white noise with standard deviation 10^{-3}.



η_V and μ_0



$\eta^{(\ell)}$ and $\mu^{(\ell)}$

- $\eta_{W,Z}$ is not a valid certificate implies support instability.
- Dot size proportional to amplitude of corresponding spikes.
- Light blue dots indicate the support of $m^{(\ell)}$ with very small amplitude.
- The additional spikes are required to force $\eta^{(l)} \leqslant 1$, this is not satisfied by $\eta_{W,Z}$.
- Numerically, no convergence in a finite number of iterations.

Summary

On conditions for recovery:

- The extremal points of solutions to the dual problem inform on the support of the primal solutions.
- Existence of a nondegenerate dual certificate guarantees exact recovery in the noiseless setting, and support clustering stability in the noisy setting.
- For support stability, we look to a special solution of $\mathcal{D}_0(y)$, the one of minimal norm $\eta_0 = \Phi^* p_0$. The MNC informs on the support of μ_λ for λ small.

Analysis of dual certificates:

- To analyse the MNC, we typically look at the vanishing derivatives precertificate which has a closed form expression.
- $\eta_V = \eta_0$ when $\|\eta_V\|_\infty \leqslant 1$. In fact, we must have $\|\eta_V\|_\infty \leqslant 1$ if we expect support stability.
- To guarantee exact recovery of spikes of arbitrary signs, it is necessary that that the underlying positions satisfy a minimum separation condition.

Numerical algorithms

- For Fourier type measurements, one can look to SDP type algorithms. However, convergence for dimensions higher than 2 are not guaranteed. Also computationally expensive.
- For more general measurements, we saw that the Frank-Wolfe algorithm can be applied.
- This is basically OMP where you add a new support point at each iteration, then locally improve over the recovered amplitudes and positions.
- **Simultaneously** optimising over the amplitudes and positions leads to substantial improvements!

http://nbviewer.jupyter.org/github/gpeyre/numerical-tours/blob/master/matlab/sparsity_8_sparsespikes_measures.ipynb