On the Beurling-LASSO with Compressive Measurements

Clarice Poon
University of Cambridge

Joint work with:

Nicolas Keriven and **Gabriel Peyré**
École Normale Supérieure

April 24, 2018

A Dual Certificates Analysis of Compressive Off-the-Grid Recovery (arXiv:1802.08464)

Outline

The Recovery Problem

For $\mathcal{X} \subseteq \mathbb{R}^d$ or $\mathbb{T}^d \stackrel{\text{def.}}{=} \mathbb{R}^d \setminus \mathbb{Z}^d$, recover $\mu_0 \in \mathcal{M}(\mathcal{X})$ from m randomized linear observations $y \in \mathbb{C}^m$, where

$$y_k = \langle \varphi_{\omega_k}, \mu_0 \rangle + \varepsilon_k, \quad \text{where} \quad \langle \varphi, \mu \rangle \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mu(x) \in \mathbb{C}.$$

- $(\varepsilon_k)_{k=1}^m \in \mathbb{C}^m$ accounts of noise or modelling errors,
- $(\omega_1, \ldots, \omega_m)$ are i.i.d. according to some probability distribution $\Lambda(\omega)$ on $\omega \in \Omega$, and for $\omega \in \Omega$, $\varphi_\omega : \mathcal{X} \to \mathbb{C}$ is a continuous function.

Typically, the measure of interest is of the form $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ with $a_j \in \mathbb{R}$, where $a\delta_x$ denotes the Dirac at $x \in \mathcal{X}$ with amplitude $a \in \mathbb{R}$ (also called a 'spike').

The Recovery Problem

For $\mathcal{X} \subseteq \mathbb{R}^d$ or $\mathbb{T}^d \stackrel{\text{def.}}{=} \mathbb{R}^d \setminus \mathbb{Z}^d$, recover $\mu_0 \in \mathcal{M}(\mathcal{X})$ from m randomized linear observations $y \in \mathbb{C}^m$, where

$$y_k = \langle \varphi_{\omega_k}, \mu_0 \rangle + \varepsilon_k, \quad \text{where} \quad \langle \varphi, \mu \rangle \stackrel{\text{def.}}{=} \int_{\mathcal{X}} \varphi(x) \mathrm{d}\mu(x) \in \mathbb{C}.$$

- $(\varepsilon_k)_{k=1}^m \in \mathbb{C}^m$ accounts of noise or modelling errors,
- $(\omega_1, \ldots, \omega_m)$ are i.i.d. according to some probability distribution $\Lambda(\omega)$ on $\omega \in \Omega$, and for $\omega \in \Omega$, $\varphi_\omega : \mathcal{X} \to \mathbb{C}$ is a continuous function.

Typically, the measure of interest is of the form $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ with $a_j \in \mathbb{R}$, where $a\delta_x$ denotes the Dirac at $x \in \mathcal{X}$ with amplitude $a \in \mathbb{R}$ (also called a 'spike').

This is the continuous analogue of the compressive sensing problem!

Example: In many imaging applications such as astronomy and electron microscopy, one aims to recover the positions of point sources and their associated amplitudes, when observing a selection of its Fourier coefficients, i.e.

$$\mathcal{X} = [0,1]^d, \quad \varphi_{\omega_k}(x) = e^{2\pi i \omega_k^T x} \quad \text{and} \quad \omega_k \stackrel{iid}{\sim} \mathrm{Unif}(\{-f_c, \ldots, f_c\}).$$

First introduced by **Tang et al (2013)** for the 1D case as the *compressive sensing off the grid* problem.

Regression

Task: Given m training samples $(\omega_k, y_k)_{k=1}^m$, construct a function to predict the values $y_k \in \mathbb{R}$ from the features $\omega_k \in \Omega$ using a continuous dictionary of functions $\omega \mapsto \varphi_\omega(x)$ parametrized by $x \in \mathcal{X}$, i.e.

$$\text{find } \mu \text{ such that } y_k \approx \int_{\mathcal{X}} \varphi_{\omega_k}(x) \mathrm{d}\mu(x).$$

Example: **Bach (2017)** formulated the training of a neural network with a single hidden layer made of an infinite number of neurons as a sparse spikes problem:

$$\Omega = \mathcal{X} = \mathbb{R}^p, \qquad \varphi_\omega(x) = \max(\langle x, \omega \rangle, 0).$$

If $\mu = \sum_j a_j \delta_{x_j}$, then

$$\int_{\mathcal{X}} \varphi_\omega(x) \mathrm{d}\mu(x) = \sum_j a_j \max(\langle x_j, \omega \rangle, 0)$$

The positions of μ represent the weights of the neural network.

Sketching of a mixture of probability densities

Task: Fit a mixture of probability densities

$$\xi(t) = \sum_i a_i \xi_{x_i}(t)$$

on some domain $\mathcal{T}$, $(\xi_x)_{x \in \mathcal{X}}$ is a family of template distributions and $a_i \geqslant 0$, $\sum_i a_i = 1$.

Sketching of a mixture of probability densities

Task: Fit a mixture of probability densities

$$\xi(t) = \sum_i a_i \xi_{x_i}(t) = \int_{\mathcal{X}} \xi_x(t) \mathrm{d}\mu_0(x), \quad \text{where} \quad \mu_0 = \sum_j a_i \delta_{x_i}$$

on some domain $\mathcal{T}$, $(\xi_x)_{x\in\mathcal{X}}$ is a family of template distributions and $a_i \geqslant 0$, $\sum_i a_i = 1$.

Sketching of a mixture of probability densities

Task: Fit a mixture of probability densities

$$\xi(t) = \sum_i a_i \xi_{x_i}(t) = \int_{\mathcal{X}} \xi_x(t) \mathrm{d}\mu_0(x), \quad \text{where} \quad \mu_0 = \sum_j a_i \delta_{x_i}$$

on some domain $\mathcal{T}$, $(\xi_x)_{x \in \mathcal{X}}$ is a family of template distributions and $a_i \geqslant 0$, $\sum_i a_i = 1$.

Example: **Gribonval et al (2017)**: $\xi_x(t) \sim \sigma^{-1} e^{-(t-\tau)^2/(2\sigma^2)}$ with parameter space as mean and standard deviation $x = (\tau, \sigma) \in \mathcal{X} = \mathbb{R} \times \mathbb{R}^+$. In practice:

- No direct access to ξ, but n i.i.d. samples $(t_1, \ldots, t_n) \in \mathcal{T}^n$ drawn from ξ.
- You do not record this (possibly huge) set of data, but compute "online" a small set $y \in \mathbb{C}^m$ of m sketches against sketching functions $\theta_\omega(t)$:

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

So, we are back to solving our problem with $\varphi_\omega(x) \stackrel{\text{def.}}{=} \int_{\mathcal{T}} \theta_\omega(t) \xi_x(t) \mathrm{d}t$.

Popular choice of sketching function over $\mathcal{T} = \mathbb{R}^d$: $\theta_\omega(t) = e^{i\langle \omega, t \rangle}$.
NB: $\varphi_\cdot(x)$ is the characteristic function of ξ_x and often has closed form expression.

The Beurling LASSO

Solve

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \frac{1}{2m} \sum_{k=1}^{m} \left|\langle \varphi_{\omega_k}, \mu\rangle - y_k\right|^2 + \lambda \left|\mu\right|(\mathcal{X})$$

where

$$\left|\mu\right|(\mathcal{X}) \stackrel{\text{def.}}{=} \sup\left\{\langle f, \mu\rangle \;;\; f \in \mathcal{C}(\mathcal{X}),\; \|f\|_\infty \leqslant 1\right\}$$

Note that if $\mu = \sum_j a_j \delta_{x_j}$, then $\left|\mu\right|(\mathcal{X}) = \|a\|_1 = \sum_i |a_i|$.

The Beurling LASSO

Solve

$$\min_{\mu\in\mathcal{M}(\mathcal{X})} \frac{1}{2m}\sum_{k=1}^{m} \left|\langle \varphi_{\omega_k}, \mu\rangle - y_k\right|^2 + \lambda \left|\mu\right|(\mathcal{X})$$

where

$$\left|\mu\right|(\mathcal{X}) \overset{\text{def.}}{=} \sup\left\{\langle f, \mu\rangle \ ; \ f\in\mathcal{C}(\mathcal{X}),\ \|f\|_\infty \leqslant 1\right\}$$

Note that if $\mu = \sum_j a_j \delta_{x_j}$, then $\left|\mu\right|(\mathcal{X}) = \|a\|_1 = \sum_i |a_i|$.

- There are other (non-variational) approaches, notably Prony-type methods (1795): MUSIC (Schmidt, 1986), ESPRIT (Roy, 1987), Finite Rate of Innovation (Vetterli, 2002)... However, these are generally restricted to Fourier (or related) measurements, and the extension to the multivariate setting is nontrivial.

The Beurling LASSO

Solve

$$\min_{\mu \in \mathcal{M}(\mathcal{X})} \frac{1}{2m} \sum_{k=1}^{m} \left|\langle \varphi_{\omega_k}, \mu \rangle - y_k\right|^2 + \lambda \left|\mu\right|(\mathcal{X})$$

where

$$\left|\mu\right|(\mathcal{X}) \stackrel{\text{def.}}{=} \sup \left\{ \langle f, \mu \rangle \; ; \; f \in \mathcal{C}(\mathcal{X}), \; \|f\|_\infty \leqslant 1 \right\}$$

Note that if $\mu = \sum_j a_j \delta_{x_j}$, then $\left|\mu\right|(\mathcal{X}) = \|a\|_1 = \sum_i |a_i|$.

- There are other (non-variational) approaches, notably Prony-type methods (1795): MUSIC (Schmidt, 1986), ESPRIT (Roy, 1987), Finite Rate of Innovation (Vetterli, 2002)... However, these are generally restricted to Fourier (or related) measurements, and the extension to the multivariate setting is nontrivial.
- There are efficient algorithms for solving this infinite dimensional minimization problem, such as SDP approaches (Candès and Fernandez-Granda, 2014), and Frank-Wolfe approaches (Bredies and Pikkarainen 2013, Boyd et al 2017).

Background on the BLASSO

- The BLASSO problem was initially proposed in **De Castro & Gamboa (2012)** and **Bredies & Pikkarainen (2013)**.
- The first sharp analysis was by **Candès & Fernandez-Granda (2014)**. In the case of Fourier measurements on $\mathbb{T}^d$ $(d = 1, 2)$, they showed that, one can recover μ_0 uniquely (with no noise) if $\Delta \stackrel{\text{def.}}{=} \min_{i \neq j} |x_i - x_j| \geqslant \frac{C}{f_c}$, given $\left\{\hat{\mu}_0(k) \; ; \; |k|_\infty \leqslant f_c\right\}$.

Background on the BLASSO

- The BLASSO problem was initially proposed in **De Castro & Gamboa (2012)** and **Bredies & Pikkarainen (2013)**.
- The first sharp analysis was by **Candès & Fernandez-Granda (2014)**. In the case of Fourier measurements on $\mathbb{T}^d$ $(d = 1, 2)$, they showed that, one can recover μ_0 uniquely (with no noise) if $\Delta \overset{\text{def.}}{=} \min_{i \neq j} |x_i - x_j| \geqslant \frac{C}{f_c}$, given $\left\{\hat{\mu}_0(k) \; ; \; |k|_\infty \leqslant f_c\right\}$.
  - For positive spikes, no separation condition required, just $f_c \geqslant 2s$: **De Castro & Gamboa (2012)**, **Schiebinger et al (2015)**, **Denoyelle et at (2017)**.

Background on the BLASSO

- The BLASSO problem was initially proposed in **De Castro & Gamboa (2012)** and **Bredies & Pikkarainen (2013)**.
- The first sharp analysis was by **Candès & Fernandez-Granda (2014)**. In the case of Fourier measurements on $\mathbb{T}^d$ ($d = 1, 2$), they showed that, one can recover μ_0 uniquely (with no noise) if $\Delta \stackrel{\text{def.}}{=} \min_{i \neq j} |x_i - x_j| \geqslant \frac{C}{f_c}$, given $\left\{\hat{\mu}_0(k) \; ; \; |k|_\infty \leqslant f_c\right\}$.
  - For positive spikes, no separation condition required, just $f_c \geqslant 2s$: **De Castro & Gamboa (2012)**, **Schiebinger et al (2015)**, **Denoyelle et at (2017)**.
- Robustness to noise under this separability condition: **Candès & Fernandez-Granda (2014)**, **Azäis et al (2015)**, **Duval and Peyré (2015)**. More on this later.

Background on the BLASSO

- The BLASSO problem was initially proposed in **De Castro & Gamboa (2012)** and **Bredies & Pikkarainen (2013)**.
- The first sharp analysis was by **Candès & Fernandez-Granda (2014)**. In the case of Fourier measurements on $\mathbb{T}^d$ ($d = 1, 2$), they showed that, one can recover μ_0 uniquely (with no noise) if $\Delta \stackrel{\text{def.}}{=} \min_{i \neq j} |x_i - x_j| \geqslant \frac{C}{f_c}$, given $\left\{\hat{\mu}_0(k) \; ; \; |k|_\infty \leqslant f_c\right\}$.
  - For positive spikes, no separation condition required, just $f_c \geqslant 2s$: **De Castro & Gamboa (2012)**, **Schiebinger et al (2015)**, **Denoyelle et at (2017)**.
- Robustness to noise under this separability condition: **Candès & Fernandez-Granda (2014)**, **Azäis et al (2015)**, **Duval and Peyré (2015)**. More on this later.
- **Tang et al (2013)** extended the work of Candès & Fernandez-Granda to randomized compressed measurements:
  - Assume that $\Delta \geqslant \frac{C_d}{f_c}$ and $\{\text{sign}(a_j)\}$ are drawn uniformly at random from the unit circle.
  - Then, with high probability, μ_0 can be recovered exactly from m Fourier frequencies, drawn uniformly at random from $\{-f_c, \ldots, f_c\}$ with

$$m = \mathcal{O}(s \log(f_c) \log(s)).$$

This talk

Assume μ_0 is approximately sparse, i.e. $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j} + \tilde{\mu}$ with $|\tilde{\mu}|(\mathcal{X}) \ll |\mu|(\mathcal{X})$.

Measurement operator is $\Phi : \mu \in \mathcal{M}(\mathcal{X}) \mapsto \frac{1}{\sqrt{m}} (\langle \varphi_{\omega_k}, \mu \rangle)_{k=1}^m \in \mathbb{C}^m$ where $\omega_k \overset{iid}{\sim} \Lambda$.

Given $y = \Phi\mu_0 + \varepsilon$, consider

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

Goal

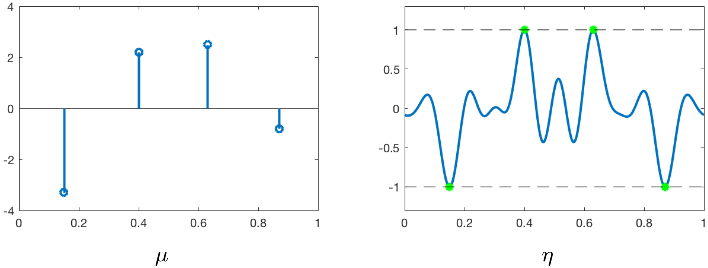
We will remove the random signs assumption, and present a general theorem with stability guarantees. The solution of $(\mathcal{P}_\lambda(y))$ is stable with respect to inexact sparsity and inexact measurements provided that $m = \mathcal{O}(s \times C_d \times \log \text{ factors})$. The implicit constant depends only on Φ.

Outline

Dual certificates

Let $X \stackrel{\text{def.}}{=} \{x_j\}_{j=1}^s$. A discrete measure $\mu_0 = \sum_i a_i \delta_{x_i}$ is a solution of $(\mathcal{P}_0(\Phi\mu_0))$ if and only if

$$\mathcal{D}(X) = \left\{\eta \in \mathrm{Im}(\Phi^*)\ ;\ \|\eta\|_{L^\infty} \leqslant 1,\ \eta(x_i) = \mathrm{sign}(a_i),\ \forall i\right\} \neq \emptyset.$$



- μ_0 is the unique solution if there exists $\eta \in \mathcal{D}(X)$ such that $|\eta(x)| < 1$ for all $x \notin X$.
- **Burger & Osher, 2004**: Existence of $\eta = \Phi^* p \in \mathcal{D}(X)$

  $\Longrightarrow$ choosing $\lambda \sim \|\varepsilon\| \stackrel{\text{def.}}{=} \delta$, solutions μ of $\mathcal{P}_\lambda(y)$ are stable wrt the Bregman "distance" :

$$\mathrm{d}(\mu, \mu_0) \stackrel{\text{def.}}{=} |\mu|(\mathcal{X}) - |\mu_0|(\mathcal{X}) - \langle \eta,\ \mu - \mu_0 \rangle = \mathcal{O}(\delta + \delta \|p\|)$$

$$\Longrightarrow\ \left||\mu|(\mathcal{X}) - |\mu_0|(\mathcal{X})\right| = \mathcal{O}(\delta + \delta\|p\|).$$

A stability result

For more precise control, look at how $|\eta|$ decays from 1 near the points $X \stackrel{\text{def.}}{=} \{x_j\}_{j=1}^s$.

Theorem (P., Keriven, Peyré '18, Variant of results by Candès & Fernandez-Granda '14 and Azäis et al. '13)

Suppose that there exists $C_0, C_2 > 0$, *neighbourhoods* $\mathcal{X}_j^{near}$ *around each point* x_j, *with* $\mathcal{X} = \mathcal{X}^{far} \cup \bigcup_{j=1}^s \mathcal{X}_j^{near}$, *and* $\eta \in \mathrm{Im}(\Phi^*)$ *such that*

- $\forall i,\ \eta(x_i) = \mathrm{sign}(a_i)$,
- $|\eta(x)| \leqslant 1 - C_0$ *for all* $x \in \mathcal{X}^{far}$,
- $\forall i,\ \forall x \in \mathcal{X}_i^{near},\ |\eta(x)| \leqslant 1 - C_2 \|x - x_i\|^2$.

Then, for $\lambda \sim \delta$, *any minimizer* $\hat{\mu}$ *of* $\mathcal{P}_\lambda(y)$ *satisfies*

$$C_0 \left|\hat{\mu} - \mu_0\right| (\mathcal{X}^{far}) + C_2 \sum_{j=1}^{s} \int_{\mathcal{X}_i^{near}} \|x - x_i\|^2 \, \mathrm{d} \left|\hat{\mu} - \mu_0\right|(x) \lesssim \delta(1 + \|p\|) + |\mu_0| (X^c).$$

A stability result

For more precise control, look at how $|\eta|$ decays from 1 near the points $X \overset{\text{def.}}{=} \{x_j\}_{j=1}^s$.

Theorem (P., Keriven, Peyré '18, Variant of results by Candès & Fernandez-Granda '14 and Azäis et al. '13)

Suppose that there exists $C_0, C_2 > 0$, *neighbourhoods* $\mathcal{X}_j^{near}$ *around each point* x_j, *with* $\mathcal{X} = \mathcal{X}^{far} \cup \bigcup_{j=1}^s \mathcal{X}_j^{near}$, *and* $\eta \in \mathrm{Im}(\Phi^*)$ *such that*

- $\forall i,\ \eta(x_i) = \mathrm{sign}(a_i)$,
- $|\eta(x)| \leqslant 1 - C_0$ *for all* $x \in \mathcal{X}^{far}$,
- $\forall i,\ \forall x \in \mathcal{X}_i^{near},\ |\eta(x)| \leqslant 1 - C_2 \|x - x_i\|^2$.

Then, for $\lambda \sim \delta$, *any minimizer* $\hat{\mu}$ *of* $\mathcal{P}_\lambda(y)$ *satisfies*

$$C_0 \left|\hat{\mu} - \mu_0\right| (\mathcal{X}^{far}) + C_2 \sum_{j=1}^s \int_{\mathcal{X}_i^{near}} \|x - x_i\|^2 \,\mathrm{d}\left|\hat{\mu} - \mu_0\right|(x) \lesssim \delta(1 + \|p\|) + |\mu_0|(X^c).$$

- If $\hat{\mu} = \sum_i \hat{a}_i \delta_{\hat{x}_i}$ and $Y_j = \{\hat{x}_i\}_i \cap \mathcal{X}_j^{near}$, then

$$C_2 \sum_{j=1}^s \sum_{\hat{x}_\ell \in Y_j} \|x_i - \hat{x}_\ell\|^2 \left|\hat{a}_\ell\right|^2 = \mathcal{O}(\delta + |\mu_0|(T^c)).$$

A stability result

For more precise control, look at how $|\eta|$ decays from 1 near the points $X \stackrel{\text{def.}}{=} \{x_j\}_{j=1}^s$.

Theorem (P., Keriven, Peyré '18, Variant of results by Candès & Fernandez-Granda '14 and Azäis et al. '13)

Suppose that there exists $C_0, C_2 > 0$, neighbourhoods $\mathcal{X}_j^{near}$ around each point x_j, with $\mathcal{X} = \mathcal{X}^{far} \cup \bigcup_{j=1}^s \mathcal{X}_j^{near}$, and $\eta \in \mathrm{Im}(\Phi^*)$ such that

- $\forall i,\ \eta(x_i) = \mathrm{sign}(a_i)$,
- $|\eta(x)| \leqslant 1 - C_0$ *for all* $x \in \mathcal{X}^{far}$,
- $\forall i, \forall x \in \mathcal{X}_i^{near}$, $|\eta(x)| \leqslant 1 - C_2 \|x - x_i\|^2$.

Then, for $\lambda \sim \delta$, any minimizer $\hat{\mu}$ of $\mathcal{P}_\lambda(y)$ satisfies

$$C_0 \left|\hat{\mu} - \mu_0\right|(\mathcal{X}^{far}) + C_2 \sum_{j=1}^s \int_{\mathcal{X}_i^{near}} \|x - x_i\|^2 \,\mathrm{d}\left|\hat{\mu} - \mu_0\right|(x) \lesssim \delta(1 + \|p\|) + |\mu_0|(X^c).$$

- Another interpretation: if $\hat{\mu}, \mu_0 \in \mathcal{M}_+(\mathcal{X})$, then $\mathcal{T}_2^2(\hat{\mu}, P_N(\hat{\mu})) = \mathcal{O}(\delta + |\mu_0|(T^c))$, where $P_N(\mu) = \sum_{j=1}^s \tilde{a}_j \delta_{x_j}$ with $\tilde{a}_j = |\mu|(\mathcal{X}_j^{near})$, and $\mathcal{T}_2$ is the W_2 partial optimal transport distance. (c.f. Eftekhari et al 2018).

A stability result

For more precise control, look at how $|\eta|$ decays from 1 near the points $X \stackrel{\text{def.}}{=} \{x_j\}_{j=1}^s$.

Theorem (P., Keriven, Peyré '18, Variant of results by Candès & Fernandez-Granda '14 and Azäis et al. '13)

Suppose that there exists $C_0, C_2 > 0$, neighbourhoods $\mathcal{X}_j^{near}$ around each point x_j, with $\mathcal{X} = \mathcal{X}^{far} \cup \bigcup_{j=1}^s \mathcal{X}_j^{near}$, and $\eta \in \mathrm{Im}(\Phi^*)$ such that

- $\forall i,\ \eta(x_i) = \mathrm{sign}(a_i)$,
- $|\eta(x)| \leqslant 1 - C_0$ *for all* $x \in \mathcal{X}^{far}$,
- $\forall i,\ \forall x \in \mathcal{X}_i^{near},\ |\eta(x)| \leqslant 1 - C_2 \|x - x_i\|^2$.

Then, for $\lambda \sim \delta$, any minimizer $\hat{\mu}$ of $\mathcal{P}_\lambda(y)$ satisfies

$$C_0 \left|\hat{\mu} - \mu_0\right| (\mathcal{X}^{far}) + C_2 \sum_{j=1}^s \int_{\mathcal{X}_i^{near}} \|x - x_i\|^2 \,\mathrm{d}\left|\hat{\mu} - \mu_0\right|(x) \lesssim \delta(1 + \|p\|) + |\mu_0| (X^c).$$

- Another interpretation: if $\hat{\mu}, \mu_0 \in \mathcal{M}_+(\mathcal{X})$, then $\mathcal{T}_2^2(\hat{\mu}, P_N(\hat{\mu})) = \mathcal{O}(\delta + |\mu_0| (T^c))$, where $P_N(\mu) = \sum_{j=1}^s \tilde{a}_j \delta_{x_j}$ with $\tilde{a}_j = |\mu| (\mathcal{X}_j^{near})$, and $\mathcal{T}_2$ is the W_2 partial optimal transport distance. (c.f. Eftekhari et al 2018).
- Under a slightly stronger condition, we have for each j, $\left|\int_{\mathcal{X}_j^{near}} \mathrm{d}(\hat{\mu} - \mu_0)(x)\right| = \mathcal{O}(\delta + |\mu_0| (X^c))$ which implies control on $\mathcal{T}_2^2(\hat{\mu}, \mu_0)$.

Support stability in the small noise regime

Duval & Peyré, 2015: One particular element of $\mathcal{D}(X)$ governs support stability.

The minimal norm certificate

$$\eta_0 = \Phi^* p_0, \qquad p_0 = \operatorname{argmin}\left\{\|p\|_2 \ ; \ \Phi^* p \in \mathcal{D}(X)\right\}.$$

If η_0 is nondegenerate, that is

$$\forall x \notin X, \ |\eta_0(x)| < 1 \quad \text{and} \quad \forall i, \ \operatorname{sign}(a_i)\nabla^2\eta_0(x_i) \prec 0.$$

- $\exists c_0$ such that for $\|\varepsilon\| \leqslant c_0\lambda$ and $\lambda \leqslant c_0$, $\mathcal{P}_\lambda(y)$ has a unique solution which consists of precisely s spikes and $(\hat{a}, \hat{x}) \to (a, x)$ smoothly as $\|\varepsilon\| \to 0$ with
$$\|x - \hat{x}\| + \|a - \hat{a}\| = \mathcal{O}(\|\varepsilon\|).$$
- In certain cases, if η_0 saturates at values other than X, then there is no support stability. So, for support stability, η_0 is the certificate to study.

Precertificates

We need to find $\eta = \Phi^* p$ such that $\eta(x_i) = \mathrm{sign}(a_i)$ for all i and $\|\eta\|_\infty \leqslant 1$. This is hard.

Vanishing derivatives precertificate

Consider instead: $\eta_V = \Phi^* p_V$ with

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

Precertificates

We need to find $\eta = \Phi^* p$ such that $\eta(x_i) = \mathrm{sign}(a_i)$ for all i and $\|\eta\|_\infty \leqslant 1$. This is hard.

Vanishing derivatives precertificate

Consider instead: $\eta_V = \Phi^* p_V$ with

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

The constraint consists of $(d+1)s$ equations and in fact,

$$\eta_V(x) = \sum_{i=1}^{N} \alpha_i K(x_i, x) + \sum_{i=1}^{N} \beta_i \partial_1 K(x_i, x), \qquad \begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix} M_0, & M_1 \\ M_1^T & M_2 \end{pmatrix}^{-1} \begin{pmatrix}\mathrm{sign}(a)\\ 0_N\end{pmatrix}$$

with correlation kernel $K(x, x') = \langle \varphi(x), \varphi(x') \rangle$,

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

Precertificates

We need to find $\eta = \Phi^* p$ such that $\eta(x_i) = \mathrm{sign}(a_i)$ for all i and $\|\eta\|_\infty \leqslant 1$. This is hard.

Vanishing derivatives precertificate

Consider instead: $\eta_V = \Phi^* p_V$ with

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

The constraint consists of $(d+1)s$ equations and in fact,

$$\eta_V(x) = \sum_{i=1}^N \alpha_i K(x_i, x) + \sum_{i=1}^N \beta_i \partial_1 K(x_i, x), \qquad \begin{pmatrix}\alpha\\ \beta\end{pmatrix} = \begin{pmatrix} M_0, & M_1 \\ M_1^T & M_2 \end{pmatrix}^{-1} \begin{pmatrix}\mathrm{sign}(a)\\ 0_N\end{pmatrix}$$

with correlation kernel $K(x, x') = \langle \varphi(x), \varphi(x')\rangle$,

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

- η_V is called the vanishing derivatives precertificate by Duval & Peyré (2015), coincides with the minimal norm certificate if $\|\eta_V\|_\infty \leqslant 1$ and is necessarily a valid certificate if there is support stability.

Precertificates

We need to find $\eta = \Phi^* p$ such that $\eta(x_i) = \mathrm{sign}(a_i)$ for all i and $\|\eta\|_\infty \leqslant 1$. This is hard.

Vanishing derivatives precertificate

Consider instead: $\eta_V = \Phi^* p_V$ with

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

The constraint consists of $(d+1)s$ equations and in fact,

$$\eta_V(x) = \sum_{i=1}^N \alpha_i K(x_i, x) + \sum_{i=1}^N \beta_i \partial_1 K(x_i, x), \qquad \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} M_0, & M_1 \\ M_1^T & M_2 \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{sign}(a) \\ 0_N \end{pmatrix}$$

with correlation kernel $K(x, x') = \langle \varphi(x), \varphi(x') \rangle$,

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

- η_V is called the vanishing derivatives precertificate by Duval & Peyré (2015), coincides with the minimal norm certificate if $\|\eta_V\|_\infty \leqslant 1$ and is necessarily a valid certificate if there is support stability.
- Typical strategy: compute some η_V based on a correlation kernel K, then check that it is nondegenerate.

Examples

Consider

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

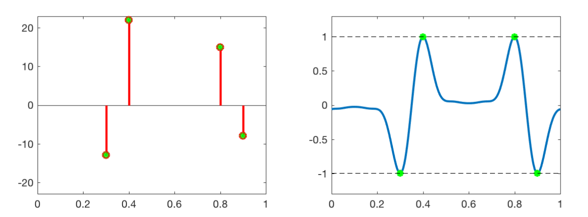
Solve

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

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

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

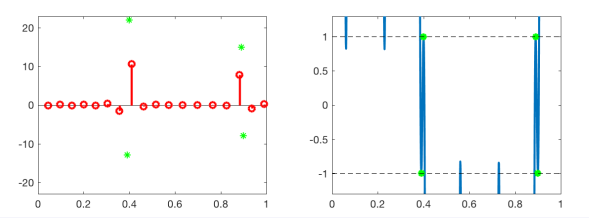
Solve

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

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 $\qquad\qquad$ η_V

Examples

Consider

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

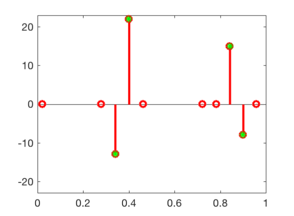
Solve

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

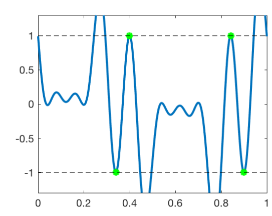
where $y = \Phi\mu_0 + \varepsilon$.

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

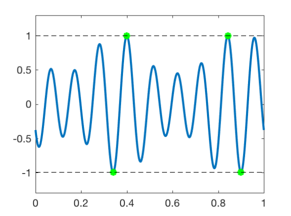
η_V is degenerate, but there exists a nondegenerate certificate:



Reconstruction



η_V



$\Phi^* p$

Examples

Consider

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

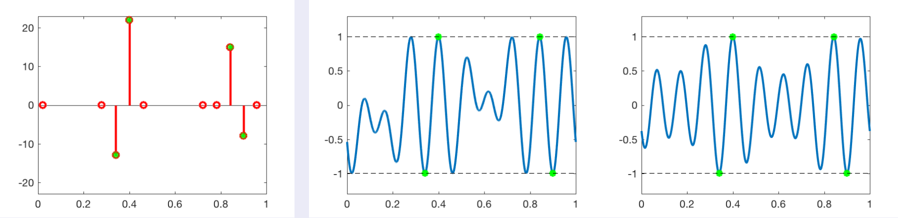
Solve

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

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 $\qquad$ η_0 $\qquad$ $\Phi^* p$

Methodology

What conditions should we impose on K and φ_ω to guarantee the existence of a nondegenerate dual certificate (and hence stable recovery) from $O(s \times \log\text{factors})$ randomized measurements?

- First observe that

$$K(x,x') = \frac{1}{m}\sum_{k=1}^{m} \mathrm{Re}(\overline{\varphi}_{\omega_k}(x)\varphi_{\omega_k}(x')) \overset{m\to\infty}{\longrightarrow} \overline{K}(x,x') \overset{\text{def.}}{=} \int \mathrm{Re}(\overline{\varphi}_\omega(x)\varphi_\omega(x'))\mathrm{d}\Lambda(\omega).$$

- What are the conditions on $\overline{K}$ such that $\overline{\eta}_V$ is nondegenerate in this limit case? This ensure stability of the limit problem

$$\min_{\mu} |\mu|(\mathcal{X}) + \frac{1}{2\lambda}\left\|\overline{\Phi}\mu - y\right\|^2_{L^2(\Lambda)}$$

  where $\overline{\Phi}\mu = \int \varphi(x)\mathrm{d}\mu(x)$ and $\varphi(x) = \omega \mapsto \varphi_\omega(x)$.

- How many measurements m are required such that η_V constructed from K is nondegenerate?

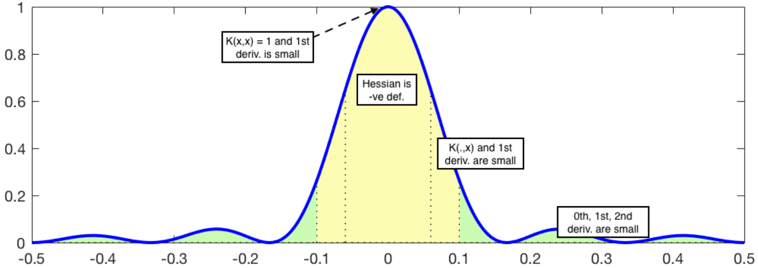
Outline

Admissible kernels

Recall:

$$\overline{\eta}_V(x) = \sum_{i=1}^{s} \alpha_i \overline{K}(x_i, x) + \sum_{i=1}^{s} \beta_i \partial_1 \overline{K}(x_i, x).$$

$$\overline{K}(x, \cdot)$$

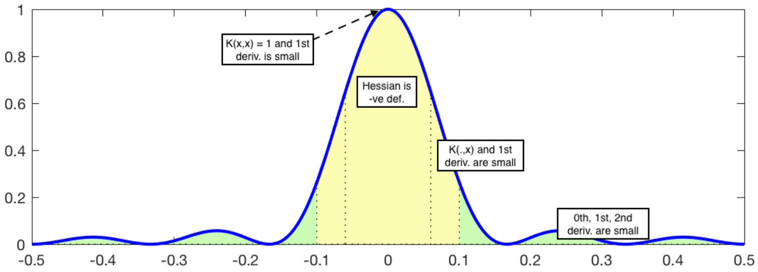


Admissible kernels

Recall:

$$\overline{\eta}_V(x) = \sum_{i=1}^{s} \alpha_i \overline{K}(x_i, x) + \sum_{i=1}^{s} \beta_i \partial_1 \overline{K}(x_i, x).$$

$$\overline{K}(x, \cdot)$$

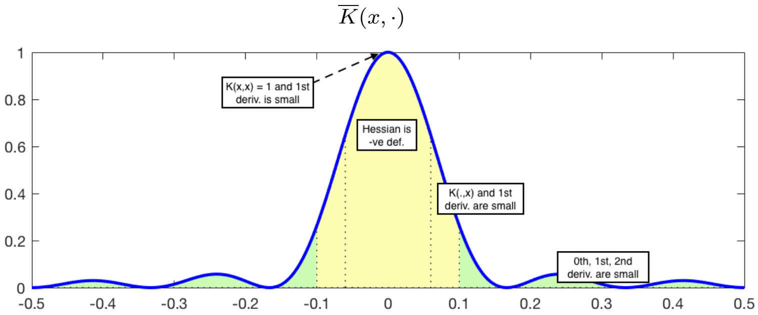


Ultimately, we want to interpolate a sign pattern using the kernel $\overline{K}$. Intuitively, locally around x_i, $\overline{K}(x_i, \cdot)$ should be sufficiently 'peaky' and when $|x_i - x'|$ is large, $\overline{K}(x_i, \cdot)$ should be sufficiently small.

Admissible kernels

Recall:

$$\overline{\eta}_V(x) = \sum_{i=1}^{s} \alpha_i \overline{K}(x_i, x) + \sum_{i=1}^{s} \beta_i \partial_1 \overline{K}(x_i, x).$$



Basically, if

- $\overline{K}(x_i, x)$ has a sufficient peak (negative definite Hessian) when $\|x - x_i\|_* \leqslant \varepsilon_{near}$,
- and the 0th, 1st, 2nd derivatives of $\overline{K}(x_i, \cdot)$ are sufficiently small wrt $1/s_{\max}$ for $\|x_i - x\|_* > \Delta$,

then η_V constructed for $s \leqslant s_{\max}$ spikes, $\{x_i\}_{i=1}^{s}$, with $\min_{k \neq j} \|x_k - x_j\|_* \geqslant \Delta$ will be nondegenerate, with control on its Hessian inside $B(x_i, \varepsilon_{near})$.

Admissible kernels

	Order 0	Order 1	Order 2	(Order 3)
$x = x'$	$\overline{K} = 1$	$\left\lvert\partial_{1,i}\overline{K}\right\rvert \leqslant \ell_1$	$\left\lvert\partial_{1,i}\partial_{2,j}\overline{K}\right\rvert \leqslant \ell_2,\ i \neq j,$ $\left\lvert\partial_{1,i}\partial_{2,i}\overline{K}\right\rvert \geqslant v$	n/a
$\left\lVert x - x'\right\rVert_* \leqslant \varepsilon_{\text{near}}$	n/a	n/a	$\operatorname{eig}\left(\nabla_2^2\overline{K}\right) \leqslant -\lambda_1$	$\left\lVert\nabla_1\nabla_2^2\overline{K}\right\rVert \leqslant \ell_3$
$\left\lVert x - x'\right\rVert_* \geqslant \varepsilon_{\text{near}}$	$\left\lvert\overline{K}\right\rvert \leqslant \ell_0$	$\left\lVert\nabla_1\overline{K}\right\rVert \leqslant \ell_1$	n/a	n/a
$\left\lVert x - x'\right\rVert_* \geqslant \frac{\Delta}{2}$	$\left\lvert\overline{K}\right\rvert \leqslant \frac{g_0}{s_{\max}}$	$\left\lVert\nabla_1\overline{K}\right\rVert \leqslant \frac{g_1}{s_{\max}}$	$\left\lVert\nabla_2^2\overline{K}\right\rVert \leqslant \frac{g_2}{s_{\max}}$	$\left\lVert\nabla_1\nabla_2^2\overline{K}\right\rVert \leqslant \frac{g_3}{s_{\max}}$
$\left\lVert x - x'\right\rVert_* \geqslant \Delta$	$\left\lvert\overline{K}\right\rvert \leqslant \frac{g_0}{s_{\max}}$	$\left\lvert\partial_{1,i}\overline{K}\right\rvert \leqslant \frac{g_1}{s_{\max}}$	$\left\lVert\partial_{1,i}\nabla_2\overline{K}\right\rVert_1 \leqslant \frac{g_2}{s_{\max}}$	n/a

Theorem (P., Keriven, Peyré (2018))

Let $u \overset{\text{def.}}{=} (\ell_1 + g_1)/\sqrt{v}$*. Assume that for some* $\delta, \delta' < 1$,

$$v^{-1}(d\ell_2 + g_2) \leqslant \delta, \quad g_0 + \frac{du^2}{1-\delta} \leqslant \delta'$$

$$\bar{C}_0 \overset{\text{def.}}{=} 1 - \left(\frac{\ell_0+g_0}{1-\delta'} + \frac{u\sqrt{d}}{\sqrt{v}(1-\delta)(1-\delta')} \cdot (c_1 + e_1)\right) > 0$$

$$\bar{C}_2 \overset{\text{def.}}{=} \left(1 - \frac{\delta'}{1-\delta'}\right) \cdot \lambda_1 - \frac{g_2}{1-\delta'} - \frac{u\sqrt{d}}{\sqrt{v}(1-\delta)(1-\delta')} \cdot (\ell_3 + g_3) > 0$$

then for $s \leqslant s_{\max}$, $a_1, \ldots, a_s \in \mathbb{R}$, $x_1, \ldots, x_s \in \mathcal{X}$ *s.t.* $\min_{j \neq k} \lVert x_j - x_k\rVert_* \geqslant \Delta$,

$$\operatorname{sign}(a_i)\nabla^2\bar{\eta}_V(x) \prec -\bar{C}_2\mathrm{Id},\ \forall \lVert x - x_i\rVert_* \leqslant \varepsilon_{\text{near}} \quad \textit{and} \quad |\bar{\eta}_V(x)| < 1 - \bar{C}_0,\ \forall x \notin \bigcup_i B(x_i, \varepsilon_{\text{near}}).$$

This extends the main result of Candès and Fernandez-Granda (2014) to general kernels by making explicit all the quantities which had to be bounded in their proof.

Examples

The **Fejér kernel** on $\mathbb{T}^d$:

$$\overline{K}(x,x') = \prod_{i=1}^{d} \kappa(x_i - x_i') \quad \text{where} \quad \kappa(t) = \left(\frac{\sin((f_c/2+1)\pi t)}{(f_c/2+1)\sin(\pi t)} \right)^4 = \sum_{\ell=-f_c}^{f_c} g(\ell) e^{i2\pi \ell t}.$$

where $g(\ell) \geqslant 0$ is such that $\sum_\ell g(\ell) = 1$. This corresponds to discrete Fourier sampling with

- $\Omega = \{-f_c, \ldots, f_c\}^d$, $\varphi_\omega(x) = e^{i2\pi\omega^T x}$ and $\Lambda(\omega) = \prod_{i=1}^d g(\omega_i)$.

To ensure that $\overline{K}$ is an admissible kernel, take

$$\varepsilon_{near} = \frac{0.1}{\sqrt{d} f_c} \quad \text{and} \quad \Delta = \frac{5\sqrt{d}\sqrt[4]{s_{\max}}}{f_c} \quad \text{and} \quad \|\cdot\|_* = \|\cdot\|_\infty$$

with f_c sufficiently large. Then, $\overline{\eta}_V$ is nondegenerate with $\bar{C}_0 \geqslant \frac{0.0056}{d}$ and $\bar{C}_2 \geqslant \frac{0.0318 f_c^2}{2}$.

Examples

The **Fejér kernel** on $\mathbb{T}^d$:

$$\overline{K}(x,x') = \prod_{i=1}^{d} \kappa(x_i - x_i') \quad \text{where} \quad \kappa(t) = \left(\frac{\sin((f_c/2+1)\pi t)}{(f_c/2+1)\sin(\pi t)}\right)^4 = \sum_{\ell=-f_c}^{f_c} g(\ell) e^{i2\pi \ell t}.$$

where $g(\ell) \geqslant 0$ is such that $\sum_\ell g(\ell) = 1$. This corresponds to discrete Fourier sampling with

- $\Omega = \{-f_c, \ldots, f_c\}^d$, $\varphi_\omega(x) = e^{i2\pi\omega^T x}$ and $\Lambda(\omega) = \prod_{i=1}^d g(\omega_i)$.

To ensure that $\overline{K}$ is an admissible kernel, take

$$\varepsilon_{near} = \frac{0.1}{\sqrt{d} f_c} \quad \text{and} \quad \Delta = \frac{5\sqrt{d}\sqrt[4]{s_{\max}}}{f_c} \quad \text{and} \quad \|\cdot\|_* = \|\cdot\|_\infty$$

with f_c sufficiently large. Then, $\overline{\eta}_V$ is nondegenerate with $\bar{C}_0 \geqslant \frac{0.0056}{d}$ and $\bar{C}_2 \geqslant \frac{0.0318 f_c^2}{2}$.

Remark

But, in the result of Candès and Fernandez-Granda, Δ does not depend on s!?

The separation distance Δ is the distance at which one 'minimizes the inteference between the spikes'. For $d=1$, if x_i's are Δ-separated, there are at most $N_d = 2$ elements of $\{x_j\}_{j\neq i}$ such that $\|x_j - x_i\| = \Delta$. As d increases, N_d grows exponentially in d.

Examples

The **Gaussian kernel** on $\mathbb{R}^d$:

$$\overline{K}(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right).$$

This corresponds to continuous Fourier sampling with

$$\Omega = \mathbb{R}^d \quad \text{and} \quad \varphi_\omega(x) = e^{i\omega^T x} \quad \text{and} \quad \Lambda = \mathcal{N}(0, \sigma^{-2}).$$

To ensure that $\overline{K}$ is an admissible kernel, Take

$$\varepsilon_{near} = \sigma/\sqrt{2} \quad \text{and} \quad \Delta = \sigma\sqrt{10\log(s_{\max}) + 4\log(d) + 24} \quad \text{and} \quad \|\cdot\|_* = \|\cdot\|_2 \,.$$

Then $\bar{C}_0 \geqslant 0.1712$ and $\bar{C}_2 \geqslant \frac{0.08}{2\sigma^2}$.

Outline

The sampled problem

Assume that $\overline{\eta}_V$ associated to $\overline{K}$ is nondegenerate. How many samples m do we need such that there exists (with high probability) a nondegenerate dual certificate in $\mathrm{Im}(\Phi^*)$?

In addition to the nondegeneracy parameters of $\overline{\eta}_V$ and properties of $\overline{K}$, the amount of subsampling is dependent on the Lipschitz constants of φ_ω and its derivatives.

Lipschitz constants

In particular, assume that $\varphi_\omega \in \mathcal{C}^2(\mathcal{X})$ with uniformly bounded derivatives and Lipschitz second derivative:

$$\sup_{\omega \in \Omega} \sup_{x \in \mathcal{X}} \|\nabla^r \varphi_\omega(x)\| \leqslant L_r, \qquad r \in \{0,1,2\}$$

and

$$\sup_{\omega \in \Omega} \left\|\nabla^2 \varphi_\omega(x) - \nabla^2 \varphi_\omega(x')\right\| \leqslant L_3 \left\|x - x'\right\|.$$

Define $L_{0,1} \overset{\text{def.}}{=} \sqrt{L_0^2 + L_1^2/v}$, where v s.t. $\left|\partial_{1,i}\partial_{2,i}\overline{K}\right| \geqslant v$.

Main result

Theorem (P., Keriven, Peyré (2018))

Let $\rho > 0$. Assume that the number of measurements m satisfies

$$m \gtrsim s \cdot \left(\mathbb{L}_1^2 \cdot \log(sd) \log\left(\frac{sd}{\rho}\right) + \sum_{j \in \{0,2\}} \mathbb{L}_j^2 \cdot \left(\log\left(\frac{(sN_j)^d}{\rho}\right) + \log\left(\frac{1}{\rho}\right) \frac{\log((sN_j)^d)}{\log(sd)} \right) \right)$$

where

$$\mathbb{L}_0^2 \overset{\text{def.}}{=} \left(L_0^2 \left(\frac{d}{\bar{C}_0} + \frac{1}{B_0^2} \right) + L_0 L_{01} \left(\frac{\sqrt{d}}{\bar{C}_0}, \frac{1}{B_0} \right) \cdot \log\left(\frac{L_0}{\bar{C}_0} + \frac{L_2}{\bar{C}_2} \right) \right.$$

$$\mathbb{L}_1^2 \overset{\text{def.}}{=} d^2 L_{01}^2 \left(\frac{B_0^2}{\bar{C}_0^2} + \frac{B_2^2}{\bar{C}_2^2} \right)$$

$$\mathbb{L}_2^2 \overset{\text{def.}}{=} \left(L_2^2 \left(\frac{d}{\bar{C}_2^2} + \frac{1}{B_2^2} \right) + L_2 L_{01} \left(\frac{\sqrt{d}}{\bar{C}_2} + \frac{1}{B_2} \right) \right) \cdot \log\left(\frac{L_0}{\bar{C}_0} + \frac{L_2}{\bar{C}_2} \right)$$

where B_0, B_2 depend only on $\overline{K}$, and

$$N_0 \overset{\text{def.}}{=} \frac{d L_{01} L_1 B_{\mathcal{X}}}{\bar{C}_0}, \qquad N_2 \overset{\text{def.}}{=} \frac{d L_{01} L_3 \varepsilon_{near}}{\bar{C}_2}$$

Then, with probability at least $1 - \rho$, there exists a dual certificate $\eta \in \mathrm{Im}(\Phi^*)$ which is nondegenerate with parameters $C_2 = \frac{\bar{C}_2}{4}$ and $C_0 = \frac{\bar{C}_0}{4}$.

Examples

The **Fejér kernel** on $\mathbb{T}^d$:

$$\overline{K}(x, x') = \prod_{i=1}^{d} \kappa(x_i - x_i') \quad \text{where} \quad \kappa(t) = \left(\frac{\sin((f_c/2+1)\pi t)}{(f_c/2+1)\sin(\pi t)} \right)^4 = \sum_{\ell=-f_c}^{f_c} g(\ell) e^{i2\pi \ell t}.$$

where $g(\ell) \geqslant 0$ is such that $\sum_\ell w_\ell = 1$. This corresponds to discrete Fourier sampling with

$$\Omega = \{-f_c, \ldots, f_c\}^d, \qquad \varphi_\omega(x) = e^{i2\pi\omega^T x}, \qquad \Lambda(\omega) = \prod_{i=1}^{d} g(\omega_i).$$

To ensure that $\overline{K}$ is an admissible kernel, take

$$\varepsilon_{near} = \frac{0.1}{\sqrt{d} f_c} \quad \text{and} \quad \Delta = \frac{5\sqrt{d}\,\sqrt[4]{s_{\max}}}{f_c} \quad \text{and} \quad \|\cdot\|_* = \|\cdot\|_\infty$$

with f_c sufficiently large. Then, $\bar{C}_0 \geqslant \frac{0.0056}{d}$ and $\bar{C}_2 \geqslant \frac{0.0318 f_c^2}{2}$.

Number of measurements

Under the assumption that the underlying positions are separated by Δ, we have a nondegenerate certificate when

$$m \gtrsim s \cdot \left(d^6 \log(sd) \log\left(\frac{sd}{\rho}\right) + d^4 \log\left(\frac{1}{\rho}\right) \log(df_c) \right).$$

Examples

The **Gaussian kernel** on $\mathbb{R}^d$:

$$\overline{K}(x,x') = \exp\left(-\frac{\|x-x'\|^2}{2\sigma^2}\right).$$

Corresponds to continuous Fourier sampling: $\varphi_\omega(x) = e^{i\omega^T x}$, $\omega \in \mathbb{R}$ and $\Lambda = \mathcal{N}(0,\sigma^{-2})$.

Admissible with $\bar{C}_0 \geqslant 0.1712$ and $\bar{C}_2 \geqslant \frac{0.08}{2\sigma^2}$ when

$$\varepsilon_{near} = \sigma/\sqrt{2} \quad \text{and} \quad \Delta = \sigma\sqrt{10\log(s_{\max}) + 4\log(d) + 24} \quad \text{and} \quad \|\cdot\|_* = \|\cdot\|_2\,.$$

Number of measurements

Problem: Since ω is unbounded, the derivatives of φ_ω are not uniformly bounded.
Fix: Weight each φ_ω by $f(\omega)$ and modify Λ to keep the same kernel $\overline{K}$. (This simply slightly modifies the way we sample). E.g.

- Let $f(\omega) = \frac{1}{2}\sqrt{\sum_{\ell=0}^{3} \frac{\|w\|^{2\ell}}{\gamma_{2\ell}}}$ where $\gamma_\ell = \mathbb{E}_{\omega\sim\mathcal{N}(0,\sigma^{-2})}\|\omega\|^\ell$.
- $\varphi_\omega(x) \overset{\text{def.}}{=} \frac{e^{i\omega^T x}}{f(\omega)}$ and $\Lambda = f(\cdot)^2\mathcal{N}(0,\sigma^{-2})$.

Number of measurements:

$$m \gtrsim s\cdot\left(d^3\log(sd)\log\left(\frac{sd}{\rho}\right) + d^4\log\left(\frac{1}{\rho}\right)\log(sd) + d^2\log\left(\frac{B_{\mathcal{X}}}{\sigma}\right)\right)$$

Mixture model learning

Assume that data points $t_1, \ldots, t_n$ are drawn according to a mixture of Gaussians

$$\sum_{i=1}^{s} a_i \mathcal{N}(x_i, \sigma_0 \mathrm{Id}),$$

where σ_0 is given and the means $\{x_i\}_{i=1}^s$ are unknown. Choose any $\sigma_* > 0$ and draw $\omega_1, \ldots, \omega_m$ iid from $\mathcal{N}(0, \sigma_*^2 \mathrm{Id})$. Let $M_* \stackrel{\text{def.}}{=} (1 + 2\sigma_0^2 \sigma_*^2)^{d/4}$.

1. Sketch the data: for $k = 1, \ldots, m$, compute $y_k = \frac{M_*}{n} \sum_{j=1}^n e^{i\omega_k^T t_j}$.
2. Solve $(\mathcal{P}_\lambda(y))$ with $\varphi_\omega(x) = M_* \exp\left(i\omega^T x - \frac{1}{2}\|\omega\|^2\right)$.

Mixture model learning

Assume that data points $t_1, \ldots, t_n$ are drawn according to a mixture of Gaussians

$$\sum_{i=1}^{s} a_i \mathcal{N}(x_i, \sigma_0 \mathrm{Id}),$$

where σ_0 is given and the means $\{x_i\}_{i=1}^{s}$ are unknown. Choose any $\sigma_* > 0$ and draw $\omega_1, \ldots, \omega_m$ iid from $\mathcal{N}(0, \sigma_*^2 \mathrm{Id})$. Let $M_* \stackrel{\text{def.}}{=} (1 + 2\sigma_0^2 \sigma_*^2)^{d/4}$.

1. Sketch the data: for $k = 1, \ldots, m$, compute $y_k = \frac{M_*}{n} \sum_{j=1}^{n} e^{i\omega_k^T t_j}$.
2. Solve $(\mathcal{P}_\lambda(y))$ with $\varphi_\omega(x) = M_* \exp\left(i\omega^T x - \frac{1}{2}\|\omega\|^2\right)$.

Then, with probability ρ,

$$\delta \stackrel{\text{def.}}{=} \|\Phi\mu_0 - y\|_2 \leqslant \frac{M_*}{\sqrt{n}} \left(1 + \sqrt{2\log(2\rho^{-1})}\right).$$

Mixture model learning

Assume that data points $t_1, \dots, t_n$ are drawn according to a mixture of Gaussians

$$\sum_{i=1}^{s} a_i \mathcal{N}(x_i, \sigma_0 \mathrm{Id}),$$

where σ_0 is given and the means $\{x_i\}_{i=1}^s$ are unknown. Choose any $\sigma_* > 0$ and draw $\omega_1, \dots, \omega_m$ iid from $\mathcal{N}(0, \sigma_*^2 \mathrm{Id})$. Let $M_* \stackrel{\text{def.}}{=} (1 + 2\sigma_0^2 \sigma_*^2)^{d/4}$.

1. Sketch the data: for $k = 1, \dots, m$, compute $y_k = \frac{M_*}{n} \sum_{j=1}^n e^{i\omega_k^T t_j}$.
2. Solve $(\mathcal{P}_\lambda(y))$ with $\varphi_\omega(x) = M_* \exp\left(i\omega^T x - \frac{1}{2} \|\omega\|^2\right)$.

Then, with probability ρ,

$$\delta \stackrel{\text{def.}}{=} \|\Phi\mu_0 - y\|_2 \leqslant \frac{M_*}{\sqrt{n}} \left(1 + \sqrt{2 \log(2\rho^{-1})}\right).$$

If $x_i's$ are Δ-separated with $\Delta \sim (\sigma_0 + \sigma_*^{-1})\sqrt{\log(sd)}$ and letting $\sigma^2 = \sigma_0^2 + \sigma_*^{-2}$,

Mixture model learning

Assume that data points $t_1, \dots, t_n$ are drawn according to a mixture of Gaussians

$$\sum_{i=1}^{s} a_i \mathcal{N}(x_i, \sigma_0 \mathrm{Id}),$$

where σ_0 is given and the means $\{x_i\}_{i=1}^s$ are unknown. Choose any $\sigma_* > 0$ and draw $\omega_1, \dots, \omega_m$ iid from $\mathcal{N}(0, \sigma_*^2 \mathrm{Id})$. Let $M_* \stackrel{\text{def.}}{=} (1 + 2\sigma_0^2 \sigma_*^2)^{d/4}$.

1. Sketch the data: for $k = 1, \dots, m$, compute $y_k = \frac{M_*}{n} \sum_{j=1}^n e^{i\omega_k^T t_j}$.
2. Solve $(\mathcal{P}_\lambda(y))$ with $\varphi_\omega(x) = M_* \exp\left(i\omega^T x - \frac{1}{2} \|\omega\|^2\right)$.

Then, with probability ρ,

$$\delta \stackrel{\text{def.}}{=} \|\Phi\mu_0 - y\|_2 \leqslant \frac{M_*}{\sqrt{n}} \left(1 + \sqrt{2\log(2\rho^{-1})}\right).$$

If $x_i' s$ are Δ-separated with $\Delta \sim (\sigma_0 + \sigma_*^{-1})\sqrt{\log(sd)}$ and letting $\sigma^2 = \sigma_0^2 + \sigma_*^{-2}$, then w.p. $\geqslant 1 - \rho$, there exists a nondegenerate dual certificate with

$$C_0 = \mathcal{O}(1) \quad \text{and} \quad C_2 = \mathcal{O}(\sigma^{-2}) \quad \text{and} \quad \varepsilon_{near} = \mathcal{O}(\sigma)$$

provided that

$$m \gtrsim sd^2 M_*^2 \left(\sigma^2 \log(sd) \log\left(\frac{sd}{\rho}\right) + \log\left(\frac{1}{\rho}\right) (\log(MB_{\mathcal{X}}) + \sigma^4 \log(sdM_*))\right)$$

Mixture model learning

Assume that data points $t_1, \dots, t_n$ are drawn according to a mixture of Gaussians

$$\sum_{i=1}^{s} a_i \mathcal{N}(x_i, \sigma_0 \mathrm{Id}),$$

where σ_0 is given and the means $\{x_i\}_{i=1}^s$ are unknown. Choose any $\sigma_* > 0$ and draw $\omega_1, \dots, \omega_m$ iid from $\mathcal{N}(0, \sigma_*^2 \mathrm{Id})$. Let $M_* \stackrel{\text{def.}}{=} (1 + 2\sigma_0^2 \sigma_*^2)^{d/4}$.

1. Sketch the data: for $k = 1, \dots, m$, compute $y_k = \frac{M_*}{n} \sum_{j=1}^n e^{i \omega_k^T t_j}$.
2. Solve $(\mathcal{P}_\lambda(y))$ with $\varphi_\omega(x) = M_* \exp\left(i\omega^T x - \frac{1}{2} \|\omega\|^2 \right)$.

Then, with probability ρ,

$$\delta \stackrel{\text{def.}}{=} \|\Phi \mu_0 - y\|_2 \leqslant \frac{M_*}{\sqrt{n}} \left(1 + \sqrt{2 \log(2\rho^{-1})} \right).$$

If $x_i' s$ are Δ-separated with $\Delta \sim (\sigma_0 + \sigma_*^{-1}) \sqrt{\log(sd)}$ and letting $\sigma^2 = \sigma_0^2 + \sigma_*^{-2}$, then w.p. $\geqslant 1 - \rho$, there exists a nondegenerate dual certificate with

$$C_0 = \mathcal{O}(1) \quad \text{and} \quad C_2 = \mathcal{O}(\sigma^{-2}) \quad \text{and} \quad \varepsilon_{near} = \mathcal{O}(\sigma)$$

provided that

$$m \gtrsim s d^2 M_*^2 \left(\sigma^2 \log(sd) \log\left(\frac{sd}{\rho}\right) + \log\left(\frac{1}{\rho}\right) \left(\log(M B_{\mathcal{X}}) + \sigma^4 \log(sd M_*) \right) \right)$$

On the dependence on d: There is a tradeoff between the number of measurements and the separation distance Δ in the choice of σ_*.

- $\sigma_* \gtrsim \sigma_0^{-1}$ implies both m and n are $\mathcal{O}(e^d)$, but $\Delta = \mathcal{O}(\sigma_0 \sqrt{\log(sd)})$.
- $\sigma_*^2 = 1/(\sigma_0^2 d)$ implies that $m = \mathrm{poly}(d)$ and $n = \mathcal{O}(d)$, but $\Delta = \mathcal{O}(\sigma_0 \sqrt{d \log(sd)})$.

Remarks on the main theorem

- The certificate is constructed via an infinite dimensional version of the **golfing scheme** (initially introduced by **Gross (2009)** for the problem of matrix completion).
  - Although it is a dual certificate with some of the nice properties of the limiting minimal norm certificate, it is not the minimal norm certificate. So, stability and uniqueness guarantees, but no support stability.

Remarks on the main theorem

- The certificate is constructed via an infinite dimensional version of the **golfing scheme** (initially introduced by **Gross (2009)** for the problem of matrix completion).
  - Although it is a dual certificate with some of the nice properties of the limiting minimal norm certificate, it is not the minimal norm certificate. So, stability and uniqueness guarantees, but no support stability.
- Following the work of Tang et al, by bounding the variations of the limiting minimal norm certificate $\overline{\eta}_0$, we can show that the minimal norm certificate η_0 for Φ is nondegenerate if

$$m = \mathcal{O}\left(s \cdot \mathrm{poly}(d) \cdot \log\left(\frac{sd}{\rho}\right) \log\left(\frac{(sN)^d}{\rho}\right)\right).$$

  - This requires a random signs assumption.
  - Without the random signs assumption, we can guarantee nondegeneracy of η_0 with

$$m = \mathcal{O}\left(s^2 \cdot \mathrm{poly}(d) \cdot \log\left(\frac{(sN)^d}{\rho}\right)\right)$$

Outline

Ideas from compressed sensing

Let $\Phi \in \mathbb{C}^{m\times N}$. Consider the finite dimensional CS problem

$$\min_x \|x\|_1 + \frac{1}{2\lambda}\|\Phi x - y\|_2^2$$

If $y = \Phi x^0 + \varepsilon$, To guarantee a solution which is stable to noisy measurements, we need to find $\mathrm{Im}(\Phi^*)$ such

$$v_j = \mathrm{sign}(x_j^0), \quad \forall j \in \mathrm{Supp}(x^0) \quad \text{and} \quad |v_j| < 1, \quad \forall j \notin \mathrm{Supp}(x^0).$$

Ideas from compressed sensing

Let $\Phi \in \mathbb{C}^{m \times N}$. Consider the finite dimensional CS problem

$$\min_x \|x\|_1 + \frac{1}{2\lambda} \|\Phi x - y\|_2^2$$

If $y = \Phi x^0 + \varepsilon$, To guarantee a solution which is stable to noisy measurements, we need to find $\mathrm{Im}(\Phi^*)$ such

$$v_j = \mathrm{sign}(x_j^0), \quad \forall j \in \mathrm{Supp}(x^0) \quad \text{and} \quad |v_j| < 1, \quad \forall j \notin \mathrm{Supp}(x^0).$$

It is sufficient to construct an approximate certificate:

Theorem (Gross (2011); Candès and Plan (2011))

Let T index the largest s entries of $\left|x^0\right|$. Suppose that there exists $v = \Phi^* p$ such that

$$\left\|v_T - \mathrm{sign}(x_T^0)\right\|_2 \leqslant \frac{1}{4} \quad \textit{and} \quad \|v_{T^c}\|_\infty \leqslant \frac{1}{4}$$

and

$$\left\|(\Phi_T^* \Phi_T)^{-1}\right\|_{2\to 2} \leqslant 2 \quad \textit{and} \quad \max_{i \in T^c} \left\|\Phi_T^* \Phi_{\{i\}}\right\|_2 \leqslant 1,$$

then one can guarantee that $\left\|\hat{x} - x^0\right\|_2 \lesssim (1 + \|p\|_2) \|\varepsilon\|_2 + \sigma_1(x^0)_s$ provided that $\lambda \sim \delta$.

Alternative proof: $\exists$ inexact certificate $\implies$ $\exists$ dual certificate

Theorem (Gross (2011); Candès and Plan (2011))

Let T index the largest s entries of $|x^0|$. Suppose that there exists $v = \Phi^* p$ such that

$$\left\|v_T - \operatorname{sign}(x^0_T)\right\|_2 \leqslant \frac{1}{4} \quad \text{and} \quad \|v_{T^c}\|_\infty \leqslant \frac{1}{4}$$

and

$$\left\|(\Phi_T^* \Phi_T)^{-1}\right\|_{2\to 2} \leqslant 2 \quad \text{and} \quad \max_{i \in T^c} \left\|\Phi_T^* \Phi_{\{i\}}\right\|_2 \leqslant 1,$$

then one can guarantee that $\left\|\hat{x} - x^0\right\|_2 \lesssim (1 + \|p\|_2) \|\varepsilon\|_2 + \sigma_1(x^0)_s$ provided that $\lambda \sim \delta$.

Proof:

Observe that $u \overset{\text{def.}}{=} v + \tilde{v}$ where

$$\tilde{v} \overset{\text{def.}}{=} \Phi^* \Phi_T (\Phi_T^* \Phi_T)^{-1} e \quad \text{and} \quad e = \operatorname{sign}(x^0_T) - v_T,$$

satisfies $u_T = \operatorname{sign}(x^0_T)$ and

$$\begin{aligned} \|u_{T^c}\|_\infty &\leqslant \|v_{T^c}\|_\infty + \left\|\Phi_{T^c}^* \Phi_T (\Phi_T^* \Phi_T)^{-1} e\right\|_\infty \\ &\leqslant \frac{1}{4} + \|\Phi_{T^c}^* \Phi_T\|_{2\to\infty} \left\|(\Phi_T^* \Phi_T)^{-1}\right\|_{2\to 2} \|e\|_2 \leqslant \frac{3}{4}. \end{aligned}$$

Outline of the proof of our main theorem

Define $\Psi : \eta \in \mathcal{C}^1(\mathcal{X}) \mapsto \left(\eta(x_1), \ldots, \eta(x_s), \nabla\eta(x_1)^T, \ldots, \nabla\eta(x_s)^T\right)^T \in \mathbb{C}^{s(d+1)}$.

Outline of the proof of our main theorem

Define $\Psi : \eta \in \mathcal{C}^1(\mathcal{X}) \mapsto \big(\eta(x_1), \dots, \eta(x_s), \nabla\eta(x_1)^T, \dots, \nabla\eta(x_s)^T\big)^T \in \mathbb{C}^{s(d+1)}$.

1. Construct (using a golfing scheme) $\tilde{\eta}$ such that
   - $\left\| \Psi\tilde{\eta} - \binom{\mathrm{sign}(a)}{0_{sd}} \right\|_{D,2} \leqslant c\min\{\bar{C}_0, \bar{C}_2\}$
   - For all $x \in \mathcal{X}^{far}_{grid}$, $|\tilde{\eta}(x)| \leqslant 1 - \frac{3\bar{C}_0}{8}$
   - For all $x \in \mathcal{X}^{near}_{grid}$, $\mathrm{sign}(a_j)\nabla^2\tilde{\eta}(x) \prec -\frac{3\bar{C}_2}{8}\mathrm{Id}$.

Outline of the proof of our main theorem

Define $\Psi : \eta \in \mathcal{C}^1(\mathcal{X}) \mapsto \left(\eta(x_1), \ldots, \eta(x_s), \nabla\eta(x_1)^T, \ldots, \nabla\eta(x_s)^T\right)^T \in \mathbb{C}^{s(d+1)}$.

1. Construct (using a golfing scheme) $\tilde{\eta}$ such that
   - $\left\| \Psi\tilde{\eta} - \binom{\mathrm{sign}(a)}{0_{sd}} \right\|_{D,2} \leqslant c \min\{\bar{C}_0, \bar{C}_2\}$
   - For all $x \in \mathcal{X}_{grid}^{far}$, $|\tilde{\eta}(x)| \leqslant 1 - \frac{3\bar{C}_0}{8}$
   - For all $x \in \mathcal{X}_{grid}^{near}$, $\mathrm{sign}(a_j)\nabla^2\tilde{\eta}(x) \prec -\frac{3\bar{C}_2}{8}\mathrm{Id}$.
2. Fix this function to obtain a certificate which is nondegenerate when evaluated on the finite grid.
   - $\Psi\eta = \binom{\mathrm{sign}(a)}{0_{sd}}$
   - For all $x \in \mathcal{X}_{grid}^{far}$, $|\eta(x)| \leqslant 1 - \frac{\bar{C}_0}{4}$
   - For all $x \in \mathcal{X}_{grid}^{near}(j)$, $\mathrm{sign}(a_j)\nabla^2\eta(x) \prec -\frac{\bar{C}_2}{4}\mathrm{Id}$.

Outline of the proof of our main theorem

Define $\Psi : \eta \in \mathcal{C}^1(\mathcal{X}) \mapsto \left(\eta(x_1), \ldots, \eta(x_s), \nabla\eta(x_1)^T, \ldots, \nabla\eta(x_s)^T\right)^T \in \mathbb{C}^{s(d+1)}$.

1. Construct (using a golfing scheme) $\tilde{\eta}$ such that
   - $\left\| \Psi\tilde{\eta} - \binom{\mathrm{sign}(a)}{0_{sd}} \right\|_{D,2} \leqslant c \min\{\bar{C}_0, \bar{C}_2\}$
   - For all $x \in \mathcal{X}_{grid}^{far}$, $|\tilde{\eta}(x)| \leqslant 1 - \frac{3\bar{C}_0}{8}$
   - For all $x \in \mathcal{X}_{grid}^{near}$, $\mathrm{sign}(a_j)\nabla^2\tilde{\eta}(x) \prec -\frac{3\bar{C}_2}{8}\mathrm{Id}$.
2. Fix this function to obtain a certificate which is nondegenerate when evaluated on the finite grid.
   - $\Psi\eta = \binom{\mathrm{sign}(a)}{0_{sd}}$
   - For all $x \in \mathcal{X}_{grid}^{far}$, $|\eta(x)| \leqslant 1 - \frac{\bar{C}_0}{4}$
   - For all $x \in \mathcal{X}_{grid}^{near}(j)$, $\mathrm{sign}(a_j)\nabla^2\eta(x) \prec -\frac{\bar{C}_2}{4}\mathrm{Id}$.
3. Extend the nondegeneracy properties to the entire domain via covering arguments.

Summary

We have

- discussed the role of dual certificates in analysing the BLASSO.
- presented conditions under which there exists nondegenerate dual certificate when the number of samples is linear with sparsity up to log factors. This gives sharp estimates in the cases of Fourier sampling in low dimensions.

Ongoing work...

- Further analysis of our conditions for a comprehensive understanding of more complex sampling operators, e.g Gaussian mixtures with varying standard deviation.
- A full understanding of the impact of dimension is still missing.

Summary

We have

- discussed the role of dual certificates in analysing the BLASSO.
- presented conditions under which there exists nondegenerate dual certificate when the number of samples is linear with sparsity up to log factors. This gives sharp estimates in the cases of Fourier sampling in low dimensions.

Ongoing work...

- Further analysis of our conditions for a comprehensive understanding of more complex sampling operators, e.g Gaussian mixtures with varying standard deviation.
- A full understanding of the impact of dimension is still missing.

Thanks for your attention.

A Dual Certificates Analysis of Compressive Off-the-Grid Recovery (arXiv:1802.08464)