An introduction to sparse spikes recovery via the BLASSO

Clarice Poon
University of Bath
C.M.H.S.Poon@maths.cam.ac.uk

January 21, 2019

Abstract

This document describes the recovery of a train of sparse spikes using the Beurling-LASSO. This problem can be seen as a generalisation of the LASSO, which is commonly used for regression or sparse recovery purposes.

1 Regression/sparse recovery via the LASSO

We begin by a reminder of the LASSO.

1.1 Regression

Given training pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$ for $i = 1, \ldots, n$, find a predictor (function) $f : \mathbb{R}^p \to \mathbb{R}$ such that $f(x_i) \approx y_i$. Then, given $x \in \mathbb{R}^p$, we can compute a prediction $y = f(x)$.

On popular approach is via the Lasso:

$$\operatorname{argmin}_{f \in \mathcal{F}} \frac{1}{n} \sum_i \|f(x_i) - y_i\|^2 + \lambda \|f\| . \tag{1}$$

Linear model

$$f_\theta(x) = \omega^\top x + b = \langle \begin{pmatrix} \omega \\ b \end{pmatrix}, \begin{pmatrix} x \\ 1 \end{pmatrix} \rangle, \quad \text{where} \quad \theta = \{\omega \in \mathbb{R}^p,\ b \in \mathbb{R}\},$$

and $\|f\| = \|\omega\|_1$.

Two Layer neural network (with one hidden layer)

$$f_\theta(x) = \sum_{j=1}^N a_j \mathrm{RELU}(\omega_j^\top x), \quad \text{where} \quad \theta = \{a_j \in \mathbb{R}, \omega_j \in \mathbb{R}^p \ ; \ j = 1, \ldots, N\},$$

$\mathrm{RELU}(x) = \max(x, 0)$ and $\|f\| = \|a\|_1$. *In this case, the LASSO is a nonconvex problem.*

1.2 Sparse recovery

In signal processing/compressed sensing, the Lasso is often referred to as Basis pursuit denoising. $\mathcal{F} = \mathbb{R}^N$ and choose $\|f\| \stackrel{\text{def.}}{=} \|f\|_1$ and $f(x) \stackrel{\text{def.}}{=} x^\top f$. Then, given a vector $y = Xf_0 \in \mathbb{R}^m$ with $X\omega = \left(x_k^\top \omega\right)_{k=1}^m$ and for some $f_0 \in \mathbb{R}^N$, (1) becomes

$$\min_{f \in \mathbb{R}^N} \frac{1}{2\lambda} \|Xf - y\|_2^2 + \|f\|_1 .$$

An example of this setup is the problem of reconstructing a sparse sum of complex sinusoids from a small sampling of its time samples: Find the parameters $(a_j, t_j) \in \mathbb{C} \times [0,1]$ for $j = 1, \dots, s$ of $S(\xi) = \sum_{j=1}^s a_j e^{2\pi i \xi t_j}$ from finitely many samples $\xi \in \Lambda = \{\xi_j \; ; \; j = 1, \dots, m\}$. Often, we discretize the interval $[0,1]$ and assume that $t_j \in \left\{\frac{n}{N} \; ; \; n = 0, \dots, N-1\right\}$. Then, $x_k \overset{\text{def.}}{=} \left(e^{2\pi i \xi_k \frac{n}{N}}\right)_{n=0}^{N-1}$. Solving Lasso recovers a vector f, if this has support S, the recovered parameters are $a = (f_n)_{n \in S}$ and $t = (n/N)_{n \in S}$.

Can we avoid discretization of $[0,1]$ *and recover the true positions and amplitudes?*

In both cases, what do we choose N *to be? Choosing* N *too small leads to discretization errors and choosing* N *too large could lead to numerical instabilities.*

2 The sparse spikes problem

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

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

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

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

This is called the **total variation norm.** Moreover, $\mathcal{M}(\mathcal{X})$ is weak* compact, i.e. given any $\mu^n \in \mathcal{M}(\mathcal{X})$ with $|\mu^n|(\mathcal{X}) \leqslant B$, there exists a subsequence and $\mu^* \in \mathcal{M}(\mathcal{X})$ such that μ^{n_k} weak* converges to μ^*.

The sparse spikes problem is as follows:

Recover $\mu_0 \in \mathcal{M}(\mathcal{X})$, $\mathcal{X} \subseteq \mathbb{R}^d$, from observations, $y = \Phi\mu_0 + w$. Here, $w \in \mathcal{H}$ is the additive noise and $\Phi : \mathcal{M}(\mathcal{X}) \to \mathcal{H}$, $\Phi\mu = \int \varphi(x) \mathrm{d}\mu(x)$ with $\varphi \in \mathcal{C}(\mathcal{X}, \mathcal{H})$. Typically, the measure of interest is $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$ where $a_j \in \mathbb{R}$ and $a\delta_x$ denotes the Dirac at $x \in \mathcal{X}$ with amplitude $a \in \mathbb{R}$.

2.1 Examples

Sampling the Fourier transform (e.g. astronomy) 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. For example, if $\Omega = \left\{k \in \mathbb{Z}^d \; ; \; |k|_\infty \leqslant f_c\right\}$, then $m = (2f_c + 1)^d$ and $\Phi\mu = \left(\sum_j a_j e^{2\pi i \langle k, \, x_j \rangle}\right)_{|k| \leqslant f_c}$.

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

Sampling the Laplace transform (e.g. microscopy) Recover $\mu \in \mathcal{M}(\mathbb{R}^d_+)$ from $(\Phi\mu)(t) = \sum_j a_j \exp(-\langle x_j, \, t \rangle)$. Here, $\varphi(x) = t \mapsto \exp(-\langle x, \, t \rangle)$.

Regression Given m training samples $\{(\omega_k, y_k) \; ; \; k = 1, \dots, m\}$, construct a function to predict the values y_k from ω_k using a continuous dictionary of functions $\omega \mapsto \varphi_\omega(x)$, parametrised by $x \in \mathcal{X}$.

The training of a two layer neural network can be placed into this framework [Bac17].

- $\mathcal{X} = \mathbb{R}^d$.
- Let $\Omega \subset \mathbb{R}^d$, and (Ω, Λ) be a probability space (with probability distribution Λ). For $\omega \in \Omega$, let $\varphi_\omega(x) = \max(\langle x, \, \omega \rangle, 0)$.

Let ω_k be m iid points drawn from (Ω, Λ) and $\varphi(x) = (\varphi_{\omega_k}(x))_{k=1}^m$. Then, $\Phi : \mathcal{M}(\mathbb{R}^d) \to \mathbb{C}^m$ and given $\mu_0 = \sum_{j=1}^s a_j \delta_{x_j}$, then

$$(\Phi\mu_0)_k = \sum_{j=1}^s a_j \max(\langle x_j, \, \omega_k \rangle, 0).$$

So, we can interpret a_j as the weights in the output layer and x_j as the weights in the hidden layer.

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

$$\xi(t) = \sum_{j=1}^s a_j \xi_{x_j}(t) = \int_{\mathcal{X}} \xi_x(t) \mathrm{d}\mu_0(x) \tag{2}$$

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 $\xi_x = \mathcal{N}(m, \sigma)$. So, this is a sparse spikes problem with $\varphi(x) \stackrel{\text{def.}}{=} \xi_x$.

Sketching of density mixtures We can also consider an extension of the density estimation problem: Typically, there is no direct access to ξ (from equation (2)) but instead, we have access to 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 $\theta_\omega(t)$ [, keriven]:

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

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

3 The 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 in the noiseless case, consider

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

This is called the Beurling LASSO, and was initially proposed in [DCG12] and [BP13].

Note that,

(i) $\mu \mapsto |\mu|(\mathcal{X})$ is lower semicontinous with respect to weak* convergence and Φ is weak* to weak continuous, so it is straightforward to establish the existence of solutions to $(\mathcal{P}_\lambda(y))$ and $(\mathcal{P}_0(y))$ via the direct method of calculus.

(ii) Recall that $|\mu|(\mathcal{X}) = \|a\|_1$ for $\mu = \sum_j a_j\delta_{x_j}$. So, this is a generalisation of the LASSO.

Questions:

1. 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 $(\mathcal{P}_0(y))$?
2. If μ_0 can be recovered in the noiseless setting, can it be stably recovered via $(\mathcal{P}_\lambda(y))$?
3. 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 $|\mu_1 - \mu_2|(\mathcal{X}) = \sum_j |a_j| + |a'_j|$.
4. When do we have support stability? That is, we recover exactly s spikes and have control on error of the amplitudes and positions.
5. Numerical algorithms which respect the infinite dimensional structure?

4 Dual certificates and recovery

4.1 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

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

One can show that

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

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

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

FACT: μ is a minimiser of a convex functional F if and only if $0 \in \partial F(\mu)$.

In particular, μ is a solution of $(\mathcal{P}_\lambda(y))$ iff $0 \in \Phi^*(\Phi\mu - y) + \lambda\partial\left|\mu\right|(\mathcal{X})$. So, if $\mu = \sum_j a_j\delta_{x_j}$, then $\eta \overset{\text{def.}}{=} \frac{1}{\lambda}\Phi^*(y - \Phi\mu)$ satisfies $0 = -\eta + \partial\left|\mu\right|(\mathcal{X})$, $\eta(x_j) = \mathrm{sign}(a_j)$, and $\left\|\eta\right\|_\infty \leqslant 1$.

4.2 Dual problems

Relevant details on duality can be found in the appendix. The Fenchel dual problem to $(\mathcal{P}_\lambda(y))$ is

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

which is equivalent to

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

This is a projection onto a closed convex set and we have immediately existence and uniqueness of the dual solution.

The dual problem of $\mathcal{P}_0(y)$ is

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

Here, existence is not guaranteed, but is true when $\mathrm{Im}(\Phi^*)$ is finite dimensional.

We have strong duality. Primal solution to $(\mathcal{P}_\lambda(y))$ and dual solution p_λ satisfy

$$\Phi^* p_\lambda \in \partial\left|\mu_\lambda\right|(\mathcal{X}) \quad \text{and} \quad p_\lambda = -\frac{1}{\lambda}(\Phi\mu_\lambda - y) \tag{3}$$

Conversely, any pair p_λ and μ_λ which satisfy the relationship (3) must be primal and dual solutions of $(\mathcal{P}_\lambda(y))$ and $(\mathcal{D}_\lambda(y))$ respectively.

If there exists $p \in \mathcal{D}_0(y)$, then it is linked to any solution μ of $\mathcal{P}_0(y)$ by

$$\Phi^* p \in \partial\left|\mu\right|(\mathcal{X}). \tag{4}$$

Again, any pair μ and p which satisfy (4) must be primal and dual solutions of $(\mathcal{P}_0(y))$ and $(\mathcal{D}_0(y))$ respectively.

4.2.1 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 1. *Suppose that $\mu_0 = \sum_j a_j\delta_{x_j}$, $y = \Phi\mu_0$ and there exists $p \in \mathcal{D}_0(y)$ such that $\Phi^* p(x_j) = \mathrm{sign}(a_j)$ and $|\Phi^* p(x)| < 1$ for all $x \notin \{x_j\}_j$, and assume that Φ_X is injective. Then, μ_0 is the unique solution to $(\mathcal{P}_0(y))$.*

Proof. Suppose that μ is a solution of $(\mathcal{P}_0(y))$. We must have $\mathrm{Supp}(\mu) \subset X$. 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, $\mu = \mu_0$. □

4.2.2 Stability

We say that a certificate is nondegenerate wrt $\operatorname{sign}(a)$, X if $\eta(x_j) = \operatorname{sign}(a_j)$, $\eta(x) < 1$ for all $x \notin \{x_j\}_j$ and $\operatorname{sign}(a_j)\nabla^2\eta(x_j) \prec 0$. In the following, we show that more 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 the support $\{x_j\}_j$.

Theorem 2. *[CFG13, ADCG15] For $i = 1, \dots, s$, let $B_\varepsilon(x_i)$ be a neighbourhood around x_i of radius ε. Suppose that there exists $c_2, c_0 > 0$ and $\eta = \Phi^* p$ 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 $\hat\mu$ to $(\mathcal{P}_\lambda(y))$ with $\|w\| \leqslant \delta$ satisfies

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

Remark 1. Suppose that $\hat\mu = \sum_{j=1}^s \sum_k \hat a_{j,k} \delta_{\hat x_{j,k}} + \sum_j \hat b_k \delta_{\hat z_k}$ where $\hat x_{j,k} \in B_\varepsilon(x_j)$ and $\hat z_j \in \mathcal{X} \setminus \bigcup_j B_\varepsilon(x_j)$. Then, this theore implies that

$$c_0 \sum_k \left|\hat b_k\right| + c_2 \sum_j \sum_k |\hat x_{j,k} - x_j|^2 |\hat a_{j,k}| \lesssim \delta \|p\|$$

which suggest that the close $\hat x_{j,k}$ is to x_j, the smaller $|\hat a_{j,k}|$ should be.

Remark 2. If $\mathcal{H}$ is a finite dimensional space, such as $\mathbb{C}^m$ or $\mathbb{R}^m$, then there exists a solution μ which is a sparse measure of at most m diracs. However, μ_0 and μ need not be sparse in general. For example, consider the case of deconvolution with $\Phi : \mathcal{M}(\mathcal{X}) \to L^2(\mathbb{R})$ with $\varphi(x) = \psi(\cdot - x)$, where

$$\psi(t) = \begin{cases} 2 - 4x & x \in [0, \frac{1}{2}] \\ 2 + 4x & x \in [-\frac{1}{2}, 0] \end{cases}.$$

If $\frac{\mathrm{d}\mu_0}{\mathrm{d}x} \overset{\text{def.}}{=} \chi_{[-1,1]}$, then $\Phi\mu_0$ is symmetric about 0 with

$$(\Phi\mu_0)(t) = \begin{cases} 1 & t \in [0, \frac{1}{2}] \\ 1 - 2(t - 1/2)^2 & t \in [1/2, 1] \\ 2(3/2 - t)^2 & t \in (1, 3/2] \\ 0 & t > 3/2 \end{cases}$$

Note that for $p \overset{\text{def.}}{=} \chi_{[-3/2,3/2]}$, $(\psi \star p)(t) = 1$ for $t \in [-1, 1]$ and $|(\psi \star p)(t)| < 1$ for all $|t| > 1$. Therefore, $\Phi^* p \in \partial |\mu_0|(\mathcal{X})$ and hence, p solves $(\mathcal{D}_0(y))$ and μ_0 solves $(\mathcal{P}_0(y))$ with $y = \Phi\mu_0$.

Remark 3. This stability result bounds the measure on $\mathcal{X} \setminus \operatorname{Supp}(\mu_0)$. In the case of μ_0 being a nonsparse measure, let $\tilde{\mathcal{X}} \overset{\text{def.}}{=} \mathcal{X} \setminus \operatorname{Supp}(\mu_0)$. Then,

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

The proof of Theorem 2 makes use of the following two lemmas. The first provides a bound on the Bregman "distance" with respect to η.

Lemma 1. *[BO04] 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,*

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

Remark 4. Given a function $J : X \to \mathbb{R} \cup \{-\infty\}$, the Bregman distance for $x, x_0 \in X$ and $p \in \partial J(x_0)$ is defined as $d^p(x, x_0) = J(x) - J(x_0) - \langle p, x - x_0 \rangle$. By definition, $d^p(x, x_0) \geqslant 0$. For example, if $X = \mathbb{R}^n$ and $J(x) = \frac{1}{2}\|x\|_2^2$, then $\partial J(x_0) = \{x_0\}$ and $d^p(x, x_0) = \frac{1}{2}\|x - x_0\|_2^2$. However, it is not a true distance for general J as it is not symmetric.

Proof. Since μ is a minimizer,

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

So,

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

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

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

and by rearranging the above inequality,

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

□

Therefore, up to a constant, the choice of $\lambda \sim \delta/\left\|p\right\|$ yields a upper bound of $\delta\left\|p\right\|$ for the Bregman distance. The claim of Theorem 2 follows combining this result with the following lower bound for $d^\eta(\mu,\mu_0)$:

Lemma 2. *Under the assumptions of Theorem 2, we have*

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

Proof. We have

$$\begin{aligned}
&\left|\mu\right|(\mathcal{X}) - \left|\mu_0\right| - \langle\eta,\, \mu-\mu_0\rangle = \left|\mu\right|(\mathcal{X}) - \langle\eta,\, \mu\rangle \\
&\geqslant \left|\mu\right|(\mathcal{X}) - \sum_i \left|\mu\right|\left(\bigcup_i B_\varepsilon(x_i)\right) + c_2\sum_j\int_{B_\varepsilon(x_j)}\left|x-x_j\right|^2\mathrm{d}\left|\mu\right|(x) - (1-c_0)\left|\mu\right|\left(\mathcal{X}\setminus\bigcup_i B_\varepsilon(x_i)\right) \\
&= c_0\left|\mu\right|\left(\mathcal{X}\setminus\bigcup_i B_\varepsilon(x_i)\right) + c_2\sum_j\int_{B_\varepsilon(x_j)}\left\|x-x_j\right\|^2\mathrm{d}\left|\mu\right|(x)
\end{aligned}$$

□

4.3 The minimal norm certificate (MNC) and support stability

Checking the existence of a dual certificate which saturates only at X guarantees uniqueness of solutions to $\mathcal{P}_0(y)$ and clustering stability. However, for *support* stability, we need to consider the certificate of minimal norm [DP15]. For simplicity, we restrict to the case of $d = 1$, although all results can be easily extended to the higher dimensional case.

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

$$p_0 \overset{\text{def.}}{=} \min\left\{\left\|p\right\| \;;\; \Phi^* p \in \partial\left|\mu^*\right|(\mathcal{X})\right\}$$

If p_0 exists, then we call it the *minimal norm certificate*, and a key property is that it is the limit of the (unique) dual solutions of $(\mathcal{D}_\lambda(y))$ as $\lambda \to 0$.

Lemma 3. *[DP15] Let p_λ be the solution to $(\mathcal{D}_\lambda(y))$ If p_0 exists, then $\left\|p_\lambda - p_0\right\| \to 0$ and $\eta_\lambda^{(k)} \to \eta_0^{(k)}$ uniformly for all k.*

Proof. Since p_λ is a solution to $\mathcal{D}_\lambda(y)$, we have

$$\langle y,\, p_\lambda\rangle - \frac{\lambda}{2}\left\|p_\lambda\right\|^2 \geqslant \langle y,\, p_0\rangle - \frac{\lambda}{2}\left\|p_0\right\|^2, \tag{5}$$

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). Taking the limit of $\lambda \to 0$ in (5) yields $\langle y,\, p_*\rangle \geqslant \langle y,\, p_0\rangle$.

Note that $\Phi^* p_{\lambda_{n_k}}$ converges weakly to $\Phi^* p$, and so,

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

Therefore, p_* solves $\mathcal{D}_0(y)$. Finally, 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 $\mathcal{H}$. This implies $\lim_{\lambda\to 0} \|p_\lambda - p_0\| = 0$, since otherwise, we can extract a subsequence p_{λ_k} such that $\|p_{\lambda_k} - p_0\| > \varepsilon$ 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

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

□

Theorem 3. *[DP15] 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 $(\mathcal{P}_\lambda(y))$ has support contained in $\bigcup_{i=1}^s B_\varepsilon(x_i)$. Moreover, if μ_0 is identifiable, then $\mu_{\lambda,w}$ has exact support $\{x_i\}$.*

Proof. Suppose that η_0 is nondegenerate. Note that since the solution to $\mathcal{D}_\lambda(y)$ is the projection onto a closed convex set, we have

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

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 $|\eta_{\lambda,w}(x)| < 1$ for $x \notin \cup_j B_r(x_j)$. So, there exists at most 1 point in $B_\varepsilon(x_j)$ for which $|\eta_{\lambda,w}| = 1$.

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 $\mu_{\lambda,w}(\mathcal{X}_j) \to \mu_0(\mathcal{X}_j) \neq 0$ and hence, for λ, w sufficiently small, $\mu_{\lambda,w}$ has exactly one spike in $B_\varepsilon(x_j)$. □

In fact, if $\Gamma_X : \mathbb{R}^{2s} \to \mathcal{H}$ defined by

$$\Gamma_X \begin{pmatrix} a \\ b \end{pmatrix} = \sum_j a_j \varphi(x_j) + \sum_j b_j \varphi'(x_j).$$

is full rank, then the following (stronger) result holds:

Theorem 4. *[DP15] Suppose that Γ_X is full rank and η_0 is nondegenerate. 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*

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

The proof can be found in Appendix C.

4.4 Precertificates

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

Vanishing derivatives precertificate [DP15] Consider instead: $\eta_V = \Phi^* p_V$ with

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

The constraint consists of $(d+1)s$ equations and in fact, writing the covariance kernel $K(x, x') \overset{\text{def.}}{=} \langle \varphi(x), \varphi(x') \rangle$, we have

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

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

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

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:

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 4. *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 $\lambda_0 > 0$ such that there exists a continuous path $g : [0, \lambda_0) \to \mathbb{R}^s \times \mathcal{X}^s$, $\lambda \mapsto (a_\lambda, X_\lambda)$ such that for all $\lambda \in [0, \lambda_0)$, $\mu_\lambda \overset{\text{def.}}{=} \sum_{i=1}^s a_{\lambda,i} x_{\lambda,i}$ solves $(\mathcal{P}_\lambda(y))$ with $y = \Phi \mu_0$. Then, η_V exists and $\|\eta_V\|_\infty \leqslant 1$, so $\eta_V = \eta_0$.*

Proof. This result follows because $p_\lambda = \frac{1}{\lambda}\left(\Phi_{X_0} a_0 - \Phi_X a\right)$ (the solution to $(\mathcal{D}_\lambda(y))$ converges to $p_V = \Gamma_{X_0}^{*,\dagger} \binom{\operatorname{sign}(a_0)}{0}$. We first establish this convergence under the assumption that g is differentiable: Given $\lambda \in [0, \lambda_0)$, let $(a, X) = g(\lambda)$. For all λ sufficiently small, we have $\operatorname{sign}(a) = \operatorname{sign}(a_0)$ by continuity of g. Therefore,

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

Note that $\Phi_X a = \Gamma_X \binom{a}{0}$. Applying $\Gamma_X^{*,\dagger} = \Gamma_X (\Gamma_X^* \Gamma_X)^\dagger$ to both sides gives

$$\Gamma_X \begin{pmatrix} a \\ 0 \end{pmatrix} - \Gamma_X \Gamma_X^\dagger \Gamma_{X_0} \begin{pmatrix} a_0 \\ 0 \end{pmatrix} + \lambda \Gamma_X^{*,\dagger} \begin{pmatrix} \operatorname{sign}(a_0) \\ 0_s \end{pmatrix} = 0. \tag{6}$$

Let Π_X be the projection onto $\operatorname{Im}(\Gamma_X)^\perp$. Then, $\Pi_X = (\mathrm{Id} - \Gamma_X \Gamma_X^\dagger)$, so we can rearrange (6) to obtain

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

Note that $p_\lambda = \frac{1}{\lambda}\left(\Phi_{X_0} a_0 - \Phi_X a\right)$ and

$$p_V = \Gamma_{X_0}^{*,\dagger} \begin{pmatrix} \operatorname{sign}(a_0) \\ 0_s \end{pmatrix} = \lim_{\lambda \to 0} \Gamma_X^{*,\dagger} \begin{pmatrix} \operatorname{sign}(a_0) \\ 0_s \end{pmatrix},$$

since $\|\Gamma_X^* \Gamma_X - \Gamma_{X_0} \Gamma_{X_0}\|$ converges to 0 as $\lambda \to 0$, by continuity of g. Therefore, $\lim_{\lambda \to 0} \|p_\lambda - p_V\| = 0$ provided that

$$\lim_{\lambda \to 0} \frac{1}{\lambda} \|\Pi_X \Phi_{X_0} a_0\| = 0.$$

By Taylor expansion,

$$\Phi_{X_0} a_0 = \sum_j a_{0,j} \varphi(x_{0,j}) = \sum_j a_{0,j} \left(\varphi(x_j) + \varphi'(x_j)(x_j - x_{0,j}) + (x_j - x_{0,j})^2 \int_0^1 \varphi''(t(x_j - x_{0,j}))\mathrm{d}t \right),$$

and since Π_X is a projection onto $\operatorname{Im}(\Gamma_X)^\perp$, we have $\Pi_X \varphi(x_j) = 0$ and $\Pi_X \varphi'(x_j) = 0$. Therefore,

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

since g is differentiable. TThis proves that $\lim_{\lambda\to 0} p_\lambda = p_V = p_0$ by uniqueness of limits.

It remains to show that g is differentiable. Define $f : \mathbb{R}^s \times \mathcal{X}^s \times \mathbb{R}_+ \to \mathbb{R}^{2s}$ by

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

with $u = (a, X)$. One can check that f is $\mathcal{C}^1$, $f((a_0, X_0), 0) = 0$, $f(g(\lambda), \lambda) = 0$ for all λ sufficiently small such that $\operatorname{sign}(a_\lambda) = \operatorname{sign}(a_0)$. Moreover, letting J_a be the diagonal matrix with diagonal $\binom{1_s}{\operatorname{sign}(a)}$, $\partial_u f((a_0, X_0), 0) = (\Gamma_{X_0}^* \Gamma_{X_0}) J_{a_0}$ is invertible.

Therefore, by the implicit function theorem, there exists a neighbourhood V of 0 in $[0, \lambda_0)$ and U of (a_0, X_0) in $\mathbb{R}^s \times \mathcal{X}^s$, and a unique continuously differentiable function g^* such that $g^* : V \to U$ satisfies $f(u, \lambda) = 0$ for $u \in U$ and $\lambda \in V$ if and only if $g^*(\lambda) = u$, and $g^*(0) = (a_0, X_0)$. In particular, g^* coincides with g on V. So, $f(g(\lambda), \lambda) = 0$ for $\lambda \in V$. □

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

5 Nondegenerate MNC for translation invariant kernels

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

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

Then,

$$K(x, x') \overset{\text{def.}}{=} \kappa(x - x'), \quad \text{where} \quad \kappa \overset{\text{def.}}{=} \psi \star \psi. \quad \text{and} \quad D_{K,X} \overset{\text{def.}}{=} \begin{pmatrix} (\kappa(x_i - x_j))_{i,j} & (\kappa'(x_i - x_j))_{i,j} \\ (\kappa'(x_j - x_i))_{i,j} & (-\kappa''(x_i - x_j))_{i,j} \end{pmatrix}.$$

For example, if $\psi = \frac{1}{\sqrt[4]{\pi}\sqrt{\sigma}} \exp\left(-t^2/(2\sigma^2)\right)$, then $\kappa(t) = \exp\left(-t^2/(4\sigma^2)\right)$.

First note that we can write we can write

$$\eta_V(x) = \sum_{i=1}^s \tilde{\alpha}_i \kappa(x_i - x) + \sum_{i=1}^N \tilde{\beta}_i \frac{\kappa'(x_i - x)}{\sqrt{|\kappa''(0)|}} \quad \text{where} \quad \begin{pmatrix} \tilde{\alpha} \\ \tilde{\beta} \end{pmatrix} = \tilde{D}_{K,X} \begin{pmatrix} \operatorname{sign}(a) \\ 0_S \end{pmatrix}$$

with

$$\tilde{D}_{K,X} \overset{\text{def.}}{=} \begin{pmatrix} (\kappa(x_i - x_j))_{i,j} & \left(|\kappa''(0)|^{-1/2} \kappa'(x_i - x_j)\right)_{i,j} \\ \left(|\kappa''(0)|^{-1/2} \kappa'(x_j - x_i)\right)_{i,j} & \left(- |\kappa''(0)|^{-1} \kappa''(x_i - x_j)\right)_{i,j} \end{pmatrix}$$

In the following theorem, we show that nondegeneracy of η_V can be guaranteed under a minimum separation condition on X. See [, Bendory] for further details.

Theorem 5. *Let $p > \frac{1}{2}$, $r, b > 0$ and assume that κ satisfies*

$$\frac{\kappa''(t)}{|\kappa''(0)|} < -b, \qquad \forall\, |t| < \frac{r}{\sqrt{|\kappa''(0)|}}$$

and for $k = 0, 1, 2, 3$,

$$\frac{\left|\kappa^{(k)}(t)\right|}{|\kappa''(0)|^{k/2}} \leqslant \frac{A_k}{(1 + Ct^2)^p}.$$

Choose $\gamma, \varepsilon \in (0, 1)$ such that $\varepsilon < b/(6 + 2b)$. and $\frac{1}{(1+r)^p} < \gamma(1 - 2\varepsilon) - 2\varepsilon$.

Let $|x_i - x_j| > \Delta$ for all $i \neq j$, with

$$\Delta \overset{\text{def.}}{=} \frac{1}{\sqrt{C}} \max_{k=0,1,2,3} \left(\frac{2p}{(2p - 1)\varepsilon} A_k\right)^{\frac{1}{2p}}.$$

Then, η_V is nondegenerate with

- $\frac{\operatorname{sign}(a_j)}{|\kappa''(0)|}\eta_V''(x) < -b/2$ *for all* $x \in \bigcup_{j=1}^s B_{|\kappa''(0)|^{-1/2}r}(x_j)$.
- $|\eta_V(x)| < \gamma$ *for all* $x \notin \bigcup_{j=1}^s B_{|\kappa''(0)|^{-1/2}r}(x_j)$.

Proof. By our choice of Δ and X,

$$\begin{aligned}\sum_{i\neq j} \frac{\left|\kappa^{(k)}(x_j - x_i)\right|}{|\kappa''(0)|^{k/2}} &\leqslant \sum_{j\geqslant 1} \frac{A_k}{(1+j^2\Delta^2)^p} \\ &\leqslant \frac{A_k}{\Delta^{2p}}\sum_{j\geqslant 1}\frac{1}{j^{2p}} \leqslant \frac{A_k}{\Delta^{2p}}\left(1 + \int_1^\infty x^{-2p}\right) \leqslant \frac{2pA_k}{(2p-1)\Delta^{2p}} = \varepsilon\end{aligned}$$

Step I, $D_{K,X}$ is invertible.
Let

$$\Upsilon_0 \overset{\text{def.}}{=} (\kappa(x_i - x_j))_{i,j}, \qquad \Upsilon_1 \overset{\text{def.}}{=} \left(\left|\kappa''(0)\right|^{-1/2}\kappa'(x_i - x_j)\right)_{i,j}, \qquad \Upsilon_2 \overset{\text{def.}}{=} \left(-\left|\kappa''(0)\right|^{-1}\kappa''(x_i - x_j)\right)_{i,j}.$$

Then,

$$\begin{aligned}\|\mathrm{Id} - \Upsilon_0\|_\infty &\leqslant \max_j \sum_{i\neq j} |\kappa(x_i - x_j)| \leqslant \varepsilon, \\ \|\Upsilon_1\|_\infty &\leqslant \frac{1}{\sqrt{|\kappa''(0)|}} \max_j \sum_i \left|\kappa'(x_i - x_j)\right| \leqslant \varepsilon, \\ \|\mathrm{Id} - \Upsilon_2\|_\infty &\leqslant \frac{1}{|\kappa''(0)|} \max_j \sum_{i\neq j} \left|\kappa''(x_i - x_j)\right| \leqslant \varepsilon,\end{aligned}$$

The Schur complement of the block Υ_2 is

$$\Upsilon_S \overset{\text{def.}}{=} \Upsilon_0 - \Upsilon_1\Upsilon_2^{-1}\Upsilon_1^\top$$

and Υ_S is invertible since

$$\|\mathrm{Id} - \Upsilon_S\|_\infty \leqslant \|\mathrm{Id} - \Upsilon_0\|_\infty + \|\Upsilon_1\|_\infty^2 \left\|\Upsilon_2^{-1}\right\|_\infty \leqslant \varepsilon + \frac{\varepsilon^2}{1-\varepsilon} = \frac{\varepsilon}{1-\varepsilon} \overset{\text{def.}}{=} \varepsilon_S.$$

Therefore, $\tilde{D}_{K,X}$ is invertible.

Step 2, bounds on the coefficients We have

$$\tilde{\alpha} = \Upsilon_S^{-1}\operatorname{sign}(a) \quad \text{and} \quad \tilde{\beta} = -\Upsilon_2^{-1}\Upsilon_1\Upsilon_S^{-1}\operatorname{sign}(a)$$

Therefore, $\|\tilde{\alpha}\|_\infty \leqslant 1/(1-\varepsilon_S) = (1-\varepsilon)/(1-2\varepsilon)$,

$$\|\tilde{\alpha} - \operatorname{sign}(a)\|_\infty \leqslant \left\|\mathrm{Id} - \Upsilon_S^{-1}\operatorname{sign}(a)\right\|_\infty \leqslant \left\|\Upsilon_S^{-1}\right\|_\infty \|\mathrm{Id} - \Upsilon_S\|_\infty \leqslant \frac{\varepsilon_S}{1-\varepsilon_S} = \frac{\varepsilon}{1-2\varepsilon}$$

and

$$\left\|\tilde{\beta}\right\|_\infty \leqslant \frac{\varepsilon}{(1-\varepsilon)(1-\varepsilon_S)} = \frac{\varepsilon}{1-2\varepsilon}.$$

Step 3, bounds on η_V Given any x, there is at most one element of X such that $|x_j - x| < \Delta/s$. So, given x s.t. $|\kappa''(0)|^{1/2}|x_j - x| \geqslant r$ for all j, we may asssume that $|x_j - x| \geqslant \Delta/2$ for all $j \neq i$ and we have

$$\begin{aligned}|\eta_V(x)| &\leqslant |\tilde{\alpha}_i\kappa(x_i - x)| + \|\tilde{\alpha}\|_\infty \sum_{j\neq i}|\kappa(x_j - x)| + \frac{\left|\tilde{\beta}_i\kappa'(x_i - x)\right|}{|\kappa''(0)|^{1/2}} + \left\|\tilde{\beta}\right\|_\infty \sum_{j\neq i}\frac{|\kappa'(x_j - x)|}{|\kappa''(0)|^{1/2}} \\ &\leqslant \frac{1-\varepsilon}{1-2\varepsilon}|\kappa(x_i - x)| + \frac{\varepsilon(1-\varepsilon)}{1-2\varepsilon} + \frac{\varepsilon}{1-2\varepsilon}\left(\left|\kappa''(0)\right|^{-1/2}\left|\kappa'(x_i - x)\right| + \varepsilon\right) \\ &\leqslant \frac{1-\varepsilon}{1-2\varepsilon}|\kappa(x_i - x)| + \frac{2\varepsilon}{1-2\varepsilon} \\ &\leqslant \frac{1-\varepsilon}{1-2\varepsilon}\frac{1}{(1+r)^p} + \frac{2\varepsilon}{1-2\varepsilon} < \gamma\end{aligned}$$

if

$$\frac{1}{(1+r)^p} < \gamma(1-2\varepsilon) - 2\varepsilon$$

If $|x_i - x| < r$, then

$$\begin{aligned}
\frac{\mathrm{sign}(a_i)\eta_V''(x)}{|\kappa''(0)|} &\leqslant \frac{1}{|\kappa''(0)|}\left(\mathrm{sign}(a_i)\tilde{\alpha}_i\kappa''(x_i-x) + \|\tilde{\alpha}\|_\infty \sum_{j\neq i}\left|\kappa''(x_j-x)\right| + \frac{\left|\tilde{\beta}_i\kappa'''(x_i-x)\right|}{|\kappa''(0)|^{1/2}} + \left\|\tilde{\beta}\right\|_\infty \sum_{j\neq i}\frac{|\kappa'''(x_j-x)|}{|\kappa''(0)|^{1/2}}\right)\\
&\leqslant \frac{\kappa''(x_i-x)}{|\kappa''(0)|} + \frac{\varepsilon}{1-2\varepsilon} + \frac{\varepsilon(1-\varepsilon)}{1-2\varepsilon} + \frac{\varepsilon}{1-2\varepsilon}(1+\varepsilon)\\
&\leqslant \frac{\kappa''(x_i-x)}{|\kappa''(0)|} + \frac{3\varepsilon}{1-2\varepsilon} < -b + \frac{3\varepsilon}{1-2\varepsilon} < -\frac{b}{2}
\end{aligned}$$

since $|\mathrm{sign}(a_i)\tilde{\alpha}_i - 1| = |\tilde{\alpha}_i - \mathrm{sign}(a_i)| \leqslant \frac{\varepsilon_S}{1-\varepsilon_S}$. and since $\varepsilon < b/(6+2b)$.

□

5.1 Examples

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

Cauchy kernel

Let $\kappa(t) = 1/(1+\sigma^{-2}t^2)$. Then,

$$\begin{aligned}
\kappa'(t) &= -\frac{2\sigma^{-2}t}{(\sigma^{-2}t^2+1)^2}\\
\kappa''(t) &= \frac{8\sigma^{-4}t^2}{(\sigma^{-2}t^2+1)^3} - \frac{2\sigma^{-2}}{(\sigma^{-2}t^2+1)^2}\\
\kappa'''(t) &= \frac{24\sigma^{-4}t}{(\sigma^{-2}t^2+1)^{p+2}} - \frac{(48\sigma^{-6}t^3)}{(\sigma^{-2}t^2+1)^4}
\end{aligned}$$

Normalising by $|\kappa''(0)| = 2\sigma^{-2}$, we have

$$\frac{\sigma}{\sqrt{2}}\kappa'(t) = -\frac{\sqrt{2}\sigma^{-1}t}{(\sigma^{-2}t^2+1)^2}$$

$$\begin{aligned}
\frac{\sigma^2}{2p}\kappa''(t) &= \frac{4\sigma^{-2}t^2}{(\sigma^{-2}t^2+1)^3} - \frac{1}{(\sigma^{-2}t^2+1)^2}\\
\frac{\sigma^3}{(2p)^{3/2}}\kappa'''(t) &= \frac{12p\sigma^{-1}t}{\sqrt{2}(\sigma^{-2}t^2+1)^3} - \frac{24\sigma^{-3}t^3}{\sqrt{2}(\sigma^{-2}t^2+1)^4}
\end{aligned}$$

Gaussian kernel Let $\kappa(t) = \exp(-t^2/(2\sigma^2))$. Then, $\kappa'(t) = \frac{-t}{\sigma^2}\exp(-t^2/\sigma)$.

$$\begin{aligned}
\kappa''(t) &= \frac{-1}{\sigma^2}\exp(-t^2/(2\sigma^2)) + \frac{t^2}{\sigma^4}\exp(-t^2/(2\sigma^2))\\
\kappa'''(t) &= \frac{3t\exp(-t^2/(2\sigma^2)))}{\sigma^4} - \frac{t^3\exp(-t^2/(2\sigma^2))}{\sigma^6}.
\end{aligned}$$

Normalising by $|\kappa''(0)|^{1/2} = 1/\sigma$, we have

$$\begin{aligned}
\sigma\kappa'(t) &= \frac{-t}{\sigma}\exp(-t^2/\sigma)\\
\sigma^2\kappa''(t) &= -\exp(-t^2/(2\sigma^2)) + \frac{t^2}{\sigma^2}\exp(-t^2/(2\sigma^2))\\
\sigma^3\kappa'''(t) &= \frac{3t\exp(-t^2/(2\sigma^2)))}{\sigma} - \frac{t^3\exp(-t^2/(2\sigma^2))}{\sigma^3}.
\end{aligned}$$

6 Some remarks on sampling the Fourier transform

Suppose we want to recover $\mu = \sum_j a_j \delta_{x_j}$ for $x_j \in \mathbb{T}$, from samples of its Fourier transform:

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

6.1 Squared Fejer kernel

Let $\varphi(x) = \left(\sqrt{g_{f_c}(k)} e^{-i2\pi k x} \right)_{|k| \leqslant f_c}$. Then the corresponding kernel is the squared Fejer kernel

$$K(x,x') = \kappa(x-x') = \sum_{|k| \leqslant f_c} g_{f_c}(k) e^{i2\pi k (x-x')} = \left(\frac{\sin(\pi t \left(\frac{f_c}{2}+1\right))}{\left(\frac{f_c}{2}+1\right)\sin(\pi t)} \right)^4.$$

Note that η_V defined via this kernel is an element of $\mathrm{Im}(\Phi^*)$

Theorem 6. *[CFG14] Suppose that $\Delta \geqslant \frac{C}{f_c}$. Then, η_V is nondegenerate and hence exact recovery is guaranteed.*

- Proving nondegeneracy of η_V with this kernel requires more refined arguments, however, the idea of the proof is similar to that of Theorem 5 above. One can show that for $t \gtrsim 1/f_c$,

$$f_c^{-j} \left| \kappa^{(j)}(t) \right| \lesssim \frac{1}{(1+t^2 f_c^2)^2}$$

  and $-\kappa''(0) = \frac{\pi^2 f_c (f_c+4)}{3} \sim f_c^2$.
- The choice of the weights is somewhat arbitrary and are chosen due to the easier-to-manipulate properties of the squared Fejer kernel. Note that without any weighting, the resultant kernel is the Dirichlet kernel which has decay like $1/(1+f_c\,|t|)$.

6.2 Necessity of the separation condition

Arbitrary signs We have so far imposed a separation condition to deduce that η_V is nondegenerate. We show here that in order to recover spikes of arbitrary signs, this is a necessary condition. See [Tan15] for further generalisations of this phenomenon to other measurement operators.

Suppose that $|x_j - x_i| = \Delta$, $\mathrm{sign}(a_j) = 1$, $\mathrm{sign}(a_i) = -1$. Then, by the mean value theorem, 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 $\Delta \geqslant 1/f_c$.

Remark 5. 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 [LF16].

All positive signs If the spikes are all positive, then ithe BLASSO does not require any separation. In particular, suppose we observe all Fourier measurements indexed by $\{k \in \mathbb{Z} \; ; \; |k| \leqslant f_c\}$ and suppose that $f_c \geqslant s$. Then,

$$\eta(x) \stackrel{\text{def.}}{=} 1 - \prod_{j=1}^{s} \sin^2(\pi(x_k - x)),$$

satisfies $\eta(x_j) = 1$, $|\eta(x)| < 1$ for all $x \notin X$ and $\eta''(x_j) \neq 0$. Therefore, η is a nondegenerate dual certificate for $\mathrm{sign}(a)$ and X if $a_j > 0$ for all j. In particular, $(\mathcal{P}_0(y))$ recovers $\mu_{a,X}$ as the unique solution. Although nondegeneracy of η can be used to deduce stability, note that the stability estimates will deteriorate as the points in X converge towards some $x_0 \in \mathcal{X}$, i.e. $\Delta(X) \to 0$, since the uniform limit of η is $1 - \sin^{2s}(\pi(x_0 - x))$ which has $2s-1$ vanishing derivatives at x_0.

See [DCG12] for generalisations of this to other measurement operators and [DDP17] for stability analysis of the BLASSO in the case of positive spikes.

6.3 Subsampling

Let us now consider the case where we observe draw m elements iid uniformly at random from

$$\left\{ \langle e^{-i2\pi k \cdot},\, \mu_0 \rangle \;;\; k \in \mathbb{Z}, |k| \leqslant \frac{f_c}{2} \right\}.$$

The following theorem is proved in [TBSR12].

Theorem 7. *Let $\mu_0 = \sum_j a_j \delta_{x_j}$. Suppose that* $\text{sign}(a)$ *is a Steinhaus sequence and*

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

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

We know that

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

is nondegenerate.

Let Ω denote the set of observed frequencies. We simply need to prove that

$$\hat{\eta}_V(x) \stackrel{\text{def.}}{=} \sum_{i=1}^N \hat{\alpha}_i \hat{K}(x_i, x) + \sum_{i=1}^N \hat{\beta}_i \partial_1 \hat{K}(x_i, x), \qquad \begin{pmatrix} \hat{\alpha} \\ \hat{\beta} \end{pmatrix} = D_{\hat{K},X}^{-1} \begin{pmatrix} \text{sign}(a) \\ 0_N \end{pmatrix}$$

where $\hat{K} = \sum_{k\in\Omega} g_{f_c}(k) e^{i2\pi k(x-x')}$, is nondegenerate. Note that $\mathbb{E}[\hat{K}] = K$ and $\mathbb{E}[D_{\hat{K},X}] = D_{K,X}$. The result can be proved by bounding the deviation of $\hat{K}$ from K, $D_{\hat{K},X}$ from $D_{K,X}$.

7 Numerical algorithm

$(\mathcal{P}_\lambda(y))$ 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 $(x_j)_{j=1}^N \subset \mathcal{X}$, that is, solve

$$\min_{a\in\mathbb{R}^N} \|a\|_1 + \|\Phi_{\text{discrete}} a - y\|^2$$

where $\Phi_{\text{discrete}} : \mathbb{R}^N \to \mathcal{H}$ is defined by $\Phi_{\text{discrete}} a = \sum_{j=1}^N a_j \varphi(x_j)$. 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. Other approach which are better aligned to the infinite dimensional nature of $(\mathcal{P}_\lambda(y))$ include SDP approaches (for Fourier measurements) or the Frank-Wolfe/conditional gradient algorithm.

7.1 SDP solver

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

$$\max_c \text{Re}\langle y,\, c\rangle - \frac{\lambda}{2} \|c\|^2 \text{ subject to } \|\mathcal{F}_n^* c\|_\infty \leqslant 1$$

where

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

Theorem 8. *[Dum07] A causal trigonometric polynomial (i.e. a trigonometric polynomial of the form $\sum_{k=0}^{n-1} c_k e^{i2\pi k t}$) with coefficients $c \in \mathbb{C}^n$ is bounded by 1 in magnitude iff there exists $Q \in \mathbb{C}^{n\times n}$ Hermitian s.t.*

$$0 \preceq \begin{pmatrix} Q & c \\ c^* & 1 \end{pmatrix} \quad \textit{and} \quad \sum_{i=1}^{n-j} Q_{i,i+j} = \delta_{0,j}, \quad j = 1, \ldots, n-1, \tag{7}$$

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

Remark 6. Note that one direction is easy to see: letting $z = (x^\top, -\langle x,\, c\rangle)$ where $x \in \mathbb{C}^n$, the positive semidefiniteness constraint implies that $x^* Q x \geqslant |\langle c,\, x\rangle|^2$. So, by choosing $x = (e^{2\pi i k t})_{k=0}^{n-1}$, we have $x^* Q x = 1$ by the constraint that all off-diagonals of Q sum to 0 and Q has trace 1. Therefore, $\left|\sum_{k=0}^{n-1} c_k e^{i2\pi k t}\right| \leqslant 1$.

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 [CFG14]: The dual problem becomes

Step I:

$$\max_{c,Q} \mathrm{Re}\langle y,\, c\rangle - \frac{\lambda}{2}\left\|c\right\|^2 \text{ subject to (7)}$$

This is a finite dimensional semidefinite program.

To find the solution to the primal problem, note that

$$p_{2n-2}(e^{i2\pi t}) = 1 - |(\mathcal{F}_n^* c)(t)|^2 = 1 - \sum_{|k| \leqslant 2f_c} u_k e^{i2\pi k t} \quad \text{where} \quad u_k = \sum_j c_j \bar{c}_{j-k}.$$

- $z^{2f_c} p_{2n-2}(z)$ 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.
- Any solution to the primal satisfies $|\mu|(\mathcal{X}) = \langle y,\, c\rangle = \langle \mu,\, \mathcal{F}_n^* c\rangle$. So, $\mathcal{F}_n^* c$ achieves its extremal points on the support of μ.

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 $\sum_{t \in \hat{X}} e^{-i2\pi k t} a_t = y_k$ to recover the amplitudes a.

7.1.1 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 $m \geqslant n = 2f_c + 1$:

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

$$\text{subject to } \begin{cases} 0 \preceq \begin{bmatrix} Q & p \\ p^* & 1 \end{bmatrix} \\ \mathrm{Trace}\Theta_k Q = \delta_{0,k}, \qquad k \in (-m, m)^d \cap \mathbb{Z}, \end{cases} \qquad (\hat{\mathcal{D}}_{\lambda,m}(y))$$

where $\Theta_k \overset{\text{def.}}{=} \theta_{k_d} \otimes \cdots \otimes \theta_{k_1}$ 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 $m \geqslant n$. For $d = 2$, it is known that we have finite convergence (and 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 $\Phi^* p$ which lie on the complex unit circle and amplitudes are found by solving the linear system of Step III above. If $\Phi^* p$ is a dual certificate to $\mu_{\lambda,m}$, then $\mu_{\lambda,m}$ is a solution to $(\mathcal{P}_\lambda(y))$.

7.2 Frank Wolfe

In this section, we present the Frank Wolfe approach to solving $(\mathcal{P}_\lambda(y))$. Unlike SDP approaches, this approach is much more general.

Frank-Wolfe algorithm aims to solve

$$\min_{m\in C} f(m) \tag{8}$$

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

In our setting, we are interested in recovering m as a measure, and $C \subseteq \mathcal{M}(\mathcal{X})$. 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 descent algorithm involves a proximal term which is often in terms of the Euclidean distance), and only uses directional derivatives of f. Indeed, in a Hilbert space setting, the proximal gradient descent algorithm

$$m^{k+1} = P_C(m^k - \gamma \mathrm{d}f(m^k))$$

where $P_C(m) \overset{\text{def.}}{=} \operatorname{argmin}_{s\in C} \|m-s\|$ is another approach for solving (8). However, this approach cannot be easily extended to the Banach space setting as projections are not necessarily well-defined.

The Frank-Wolfe algorithm is as follows

Algorithm 1 Frank-Wolfe

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

Let us make some remarks:

- 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.
- The assumption of weak compactness ensures that step 2 has a minimizer.

In our case, we are interested in applying Frank-Wolfe to

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

There are 2 immediate problems: the first is that f_λ is not differentiable and the second is that $\mathcal{M}(\mathcal{X})$ is unbounded. The following lemma allows us to rewrite minimisation of f_λ over $\mathcal{M}(\mathcal{X})$ into the form (8).

Lemma 5. *[DDPS18] μ_* is a minimiser of f_λ if and only if $(|\mu_*|\,(\mathcal{X}), \mu_*)$ minimises*

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

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

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 $\tilde{f}_\lambda$. □

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 6. *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*

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

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

Theorem 9. *Suppose that $\mu_{a,X} = \sum_i a_i \delta_{x_i}$ 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. 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, $a_j^k \overset{\text{def.}}{=} \mu^k(B_\varepsilon(x_j))$ satisfies $\left\|a^k - a\right\|^2 \lesssim \frac{1}{k}$.*

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

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

For the second claim, define

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

Note that for all ν, $R(\nu) \geqslant 0$ (since $-F'(\mu^*) \in \partial J(\mu^*)$ and $T(\nu) \geqslant 0$ by convexity of F. Also, $\lambda^{-1} r_k = J(\mu^k) + T(\mu^k) \geqslant T(\mu^k)$. 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 $\left\|\Phi_X a\right\|^2 \geqslant C\left\|\Phi_X a\right\|^2$, then

$$\begin{aligned}
\lambda^{-1} r_k \geqslant T(\mu^k) = \frac{\left\|\Phi(\mu^k - \mu^*)\right\|^2}{2} &\geqslant \frac{3}{8}\left\|\Phi(\hat{\mu}^k - \mu^*)\right\|^2 + \frac{3}{2}\left\|\Phi(\hat{\mu}^k - \mu^k)\right\|^2 \\
&\geqslant \frac{3}{8} C \sum_k \left|a_j^k - a_j\right|^2 - \frac{3}{2}\left\|\Phi(\hat{\mu}^k - \mu^k)\right\|^2,
\end{aligned}$$

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

$$\begin{aligned}
\left\|\Phi(\hat{\mu}^k - \mu^k)\right\|^2 &\leqslant \left\| \sum_j \int_{B_\varepsilon(x_j)} (\varphi(x) - \varphi(x_j))\mathrm{d}\mu^k(x) + \int_{\mathcal{X}^{far}} \varphi(x)\mathrm{d}\mu^k(x) \right\|^2 \\
&\leqslant 2\left(\sum_j \int_{B_\varepsilon(x_j)} \left\|\varphi'\right\|_\infty |x - x_j| \,\mathrm{d}\left|\mu^k\right|(x) \right)^2 + 2\left|\mu^k\right|(\mathcal{X}^{far})^2 \\
&\leqslant 2\left(\sum_j \left\|\varphi'\right\|_\infty \sqrt{\left|\mu^k\right|(B_\varepsilon(x_j)) \int_{B_\varepsilon(x_j)} |x - x_j|^2 \,\mathrm{d}\left|\mu^k\right|(x)} \right)^2 + 2\left|\mu^k\right|(\mathcal{X}^{far})^2 \\
&\leqslant 2\left\|\varphi'\right\|_\infty \left|\mu^k\right|(\mathcal{X}^{near}) \left(\sum_j \int_{B_\varepsilon(x_j)} |x - x_j|^2 \,\mathrm{d}\left|\mu^k\right|(x) \right) + 2\left|\mu^k\right|(\mathcal{X}^{far})^2 \\
&\lesssim \lambda^{-1} c_2^{-1} r_k + \lambda^{-2} c_0^{-2} r_k^2.
\end{aligned}$$

□

Let us now discuss lines the form of 2, 3, 4 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+1}, $s^k = (M, \sigma M\delta_x)$ where $\sigma\eta^k(x^{k+1}) = -\left\|\eta^k\right\|_\infty$. Let $\{x_j^k\}_{j=1}^k$ denote the support of μ^k.
- 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. Therefore, we in fact iteratively construct a dual certificate.
- 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 $\mu_\gamma = \sum_{j=1}^k a_j^k \delta_{x_j^k} + \sigma M \delta_{x^{k+1}}$. 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

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

  where $\mu_a \overset{\text{def.}}{=} \sum_{j=1}^k a_j \delta_{x_j^k} + a_{k+1}\delta_{x^{k+1}}$. This is a finite dimensional nonsmooth convex optimisation problem and can be tackled using a variety of algorithms such as Forward Backward or FISTA.

7.2.1 The sliding Frank-Wolfe algorithm

The observation that one can replace the update of step 6 by any value which improves the objective value is important. As observed in [BP13] and [BSR17], this can significantly improve the convergence properties of the algorithm. Building upon remark (iii) above, one can further improve this step by optimising over the positions and the amplitudes simultaneously. This idea is proposed and analysed in [DDPS18] and under certain nondegeneracy conditions, this update leads to *finite termination.* Moreover, it is observed empirically that under the nondegeneracy condition, one has convergence in s iterations. The algorithm of [DDPS18] is presented in Algorithm 2. Note that since t is an auxilliary variable, it is omitted in the presentation of the algorithm.

Theorem 10. *[DDPS18] 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. We discuss only the main ideas of the proof here. First note that μ^k converges to $\mu_{a,X}$ in the weak-$*$ topology. Since Φ is weak-$*$ to weak continuouse, 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 $\eta^k \overset{\text{def.}}{=} x \mapsto \langle\varphi(x), p^k\rangle$ 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.

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 $\left|\eta^k(x)\right| < 1$ for all $x \notin \cup_i I_{x_i,\varepsilon}$.

Finally, note that the optimality condition of step 8 is

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

So, $\eta^k = -\frac{1}{\lambda}\Phi^*(\Phi_{x^k}a^k - y)$ satisfies $\eta^k(x_j^k) = \mathrm{sign}(a_j^k)$ and $(\eta^k)'(x_j) = 0$. In particular, $\left|\eta^k(x)\right| < 1$ except at x^k. So, η^k is a valid certificate for all $k \geqslant k_1$. Hence, the algorithm terminates for $k \geqslant k_1$.

□

Algorithm 2 Sliding Frank-Wolfe [DDPS18]

```
 1: Initialise with $\mu^0 = 0$.
 2: for $k = 0, \dots, n$ do
 3:     $\mu^k = \sum_{i=1}^{N^k} a_i^k \delta_{x_i^k}$, $a_i^k \in \mathbb{R}$, $x_i^k \in \mathcal{X}$ distinct, find $x_*^k \in \mathcal{X}$ s.t.
            $x_*^k \in \operatorname{argmax}_{x\in\mathcal{X}} \left|\eta^k(x)\right| \quad \text{where} \quad \eta^k \overset{\text{def.}}{=} \frac{1}{\lambda}\Phi^*(y - \Phi\mu^k).$
 4:     if then $\left|\eta^k(x_*^k)\right| \leqslant 1$
 5:         $\mu^k$ is a solution. Stop.
 6:     else
 7:         $\mu^{k+\frac{1}{2}} = \sum_{i=1}^{N^k} a_i^{k+\frac{1}{2}} \delta_{x_i^k} + a_i^{k+\frac{1}{2}} \delta_{x_*^k}$ s.t.
                $a^{k+\frac{1}{2}} \in \operatorname{argmin}_{a\in\mathbb{R}^{N^k+1}} \frac{1}{2} \left\|\Phi_{x^{k+\frac{1}{2}}} a - y\right\|^2 + \lambda \left\|a\right\|_1$
    where $x^{k+\frac{1}{2}} = (x_1^k, \cdots, x_{N^k}^k, x_*^k)$.
 8:         $\mu^{k+1} = \sum_{i=1}^{N^k+1} a_i^{k+1} \delta_{x_i^{k+1}}$ s.t.
                $(a^{k+1}, x^{k+1}) \in \operatorname{argmin}_{(a,x)\in\mathbb{R}^{N^k}\times\mathcal{X}^{N^k+1}} \frac{1}{2} \left\|\Phi_x a - y\right\|^2 + \lambda \left\|a\right\|_1,$
    using a non-convex solver initialised with $(a^{k+\frac{1}{2}}, x^{k+\frac{1}{2}})$.
 9:     end if
10: end for
```

Remark 7. Step 8 of Algorithm 2 requires solving a nonconvex optimisation problem, however, the proof of Theorem 10 utilises only the optimality condition of the optimisation problem and hence, finite convergence still holds even if $a^{k+\frac{1}{2}}$ is merely a *stationary point*.

A Useful facts and definitions

Schur complement: Consider solving for x, y:

$$\begin{pmatrix} A & B \\ C & D \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix} \tag{9}$$

The Schur complement of the block D of this matrix is $S \overset{\text{def.}}{=} A - BD^{-1}C$, and provided that S and D are invertible, this system is solvable with

$$x = S^{-1}(a - BD^{-1}b) \quad \text{and} \quad y = D^{-1}(b - Cx).$$

A Banach space is a vector space over a field K ($\mathbb{R}$ or $\mathbb{C}$) equipped with a norm $\left\|\cdot\right\|_X$ which is complete with respect to this norm. A Hilbert space is a real or complex inner product space with inner product $\langle\cdot,\,\cdot\rangle$, complete with respect to the induced norm $\left\|x\right\| \overset{\text{def.}}{=} \sqrt{\langle x,\,x\rangle}$. A Hilbert space is a Banach space.

The dual space of a normed space X is denoted by X' and is the space of continuous linear functionals from X to K. X' is a Banach space for every normed space X.

The weak topology on X is the coarsest topology on X such that all elements $x' \in X'$ are continuous. In particular, a sequence $(x_n) \subset X$ is said to weakly converge to x if for all $y \in X'$, $y(x_n) \to y(x)$.

The weak* topology on X' is the coarsest topology on X' such that for all $x \in X$, the mapping $x' \mapsto x'(x)$ is continuous. In particular, a sequence $(x_n') \subset X'$ is said to weakly converge to x' if for all $y \in X$, $x_n(y) \to x(y)$.

Weakly convergent sequences are uniformly bounded (due to the uniform boundedness principle/Banach-Steinhaus theorem).

Every bounded sequence in a Hilbert space has a weakly convergent subsequence.

If u_n converges weakly to u in a Hilbert space, then $\liminf_n \|u_n\| \geqslant \|u\|$. Furthermore, $\lim_n \|u_n\| = u$ implies that u_n converges strongly to u.

$\mathcal{M}(\mathcal{X})$ is weak* compact, that is, given $\mu^n \in \mathcal{M}(\mathcal{X})$ such that $|\mu^n|(\mathcal{X}) \leqslant B$, there exists $\mu \in \mathcal{M}(\mathcal{X})$ and a subsequence such that μ^{n_k} weak* converges to μ.

Theorem 11 (Arzela-Ascoli)**.** *If X is a compact metric space and $f_n : X \to \mathbb{R}$ are equicontinuous (i.e. for every $\varepsilon > 0$, there exists $\delta > 0$ such that $|f_n(x) - f_n(y)| \leqslant \varepsilon$ for all x, y with $d(x, y) < \delta$ and all n) and equibounded (i.e. there exists $C > 0$ such that $|f_n(x)| \leqslant C$ for all $x \in X$ and all n), then there exists a subsequence f_{n_k} which converges uniformly to a continuous function $f : X \to \mathbb{R}$.*

B Duality

Let V be a real topological vector space and let V^* be its dual.

Definition 1. *Given $F : V \to (-\infty, +\infty]$, its convex conjugate is $F^* : V^* \to (-\infty, +\infty]$ defined by*

$$F^*(y) \overset{\text{def.}}{=} \sup_{x \in V} \{\langle x,\, y\rangle - F(x)\}.$$

- F^* is convex regardless of whether F is convex.
- We have the Fenchel Young inequality: $\langle x,\, y\rangle \leqslant F(x) + F^*(y)$,
- if F is convex and lower semi-continuous, then $F^{**} = F$.
- if F is convex, then $y \in \partial F(x)$ if and only if $F(x) + F^*(y) = \langle x,\, y\rangle$.

Examples:

(a) if $F(x) = \frac{1}{2}\|x\|^2$ and V is a Hilbert space, then $F^*(y) = \frac{1}{2}\|y\|^2$:
  - $F^*(y) = \sup_x \langle x,\, y\rangle - \frac{1}{2}\|x\|^2 \leqslant \frac{1}{2}\|y\|^2$.
  - Setting $x \overset{\text{def.}}{=} y$ in the supremum above yields $F^*(y) \geqslant \frac{1}{2}\|y\|^2$.

(b) If $F(x) = \|x\|$ and $\|\cdot\|_*$ is its dual norm, then

$$F^*(y) = \begin{cases} 0 & \|y\|_* \leqslant 1 \\ +\infty & \text{otherwise.} \end{cases}$$

(c) If $F = \iota_K$ (takes value 0 for $x \in K$ and $+\infty$ otherwise) with K being a convex set, then $F^*(y) = \sup_{x \in K} \langle x,\, y\rangle$.

B.1 Convex optimisation

Let V, Y be real topological vector spaces with duals V^* and Y^*. Let $y \in Y$ and $b_j \in \mathbb{R}$ for $j = 1, \ldots, M$. Consider

$$\min_{x \in V} F_0(x) \text{ subject to } Ax = y, \tag{10}$$

$$F_j(x) \leqslant b_j,\ j \in [M], \tag{11}$$

where $F_0 : V \to (-\infty, +\infty]$ is called the objective function and $F_j : V \to (-\infty, +\infty]$ for $j \in [M]$ are called the constraint functions. $A : V \to Y$ is a continuous linear functional. The set $K \overset{\text{def.}}{=} \{x \in V\ ;\ Ax = y, F_j(x) \leqslant b_j\}$ is called the admissible set.

The **Lagrange function** is defined for $x \in V$, $\xi \in Y^*$ and $\nu \in \mathbb{R}^M$ with $\nu_\ell \geqslant 0$ for all $\ell \in [M]$ by

$$L(x, \xi, \nu) \overset{\text{def.}}{=} F_0(x) + \langle \xi,\, Ax - y\rangle + \sum_{\ell=1}^{M} \nu \left(F_\ell(x) - b_\ell\right).$$

The variables ξ and ν are called the **Lagrange multipliers**.

The Lagrange dual function is defined as

$$H(\xi, \nu) \overset{\text{def.}}{=} \inf_{x \in V} L(x, \xi, \nu), \qquad \xi \in Y^*,\ \nu \in \mathbb{R}^M_{\geqslant 0}.$$

If $x \mapsto L(x, \xi, \nu)$ is unbounded from below, then we write $H(\xi, \nu) = -\infty$.

- The dual function is always concave since it is the pointwise infimum of a family of affine functions.
- We have $H(\xi,\nu) \leqslant \inf_{x\in K} F_0(x)$ for all $\xi \in Y^*$ and $\nu \in \mathbb{R}^M_{\geqslant 0}$. Indeed, we have $H(\xi,\nu) \leqslant \inf_{x\in K} L(x,\xi,\nu)$, and note that given any $x \in K$, we have $Ax - y = 0$ and $F_\ell(x) - b_\ell \leqslant 0$, so $L(x,\xi,\nu) \leqslant F_0(x)$.

So, $H(\xi,\nu)$ serves as a lower bound for the infimum of F_0 over K, and since we want this lower bound to be as tight as possible, it makes sense to consider

$$\sup_{\xi\in Y^*,\nu\in\mathbb{R}^M} H(\xi,\nu) \text{ subject to } \nu_\ell \geqslant 0,\ \ell \in [M]. \tag{12}$$

This optimisation problem is called the **dual problem** and (10) is called the **primal problem**.

- If D^* is the supremum of (12) and P^* is the infimum of (10), then we have in general $D^* \leqslant P^*$ (this is called **weak duality**). When $D^* = P^*$, then we say we have **strong duality**.

The following theorem (Slater's condition) gives a condition under which strong duality holds.

Theorem 12. *Let $F_0, F_1, \ldots, F_M$ be convex functions and suppose that* $\mathrm{dom}(F_0) = V$. *If there exists $x_0 \in V$ such that $Ax_0 = y$, $F_\ell(x_0) < b_\ell$ for all $\ell \in [M]$, then strong duality holds. In the absence of the inequality constraints, we have strong duality if there exists x_0 such that $Ax_0 = y$.*

Consider now $\inf_{x\in V} F(Ax) + G(x)$, where $F : Y \to (-\infty, +\infty]$ and $G : V \to (-\infty, +\infty]$ are convex functionals, and $A : V \to Y$ is a continuous linear operator. This is equivalent to

$$\inf_{z\in Y, x\in V} F(z) + G(x) \text{ subj. to } Ax = z$$

the Lagrange dual is for $\xi \in Y^*$ as

$$\begin{aligned} H(\xi) &= \inf_{x,z}\{F(z) + G(x) + \langle \xi,\, Ax - z\rangle\} \\ &= \inf_{x,z}\{F(z) + G(x) + \langle A^*\xi,\, x\rangle - \langle \xi,\, z\rangle\} \\ &= -\sup_{z\in Y}\langle \xi,\, z\rangle - F(z) - \sup_{x\in V}\langle -A^*\xi,\, x\rangle - G(x) \\ &= -F^*(\xi) - G^*(-A^*\xi). \end{aligned}$$

So, the dual problem is

$$\sup_{\xi\in Y^*} -F^*(\xi) - G^*(-A^*\xi)$$

Theorem 13. *Suppose that F and G are proper convex functionals, there exists $u_0 \in V$ such that $F(u_0) < \infty$, $G(Au_0) < \infty$ and G is continuous at Au_0. Then, strong duality holds and there exists at least one dual optimal solution. Moreover, if p^* is a primal optimal solution and d^* is a dual optimal solution, then*

$$Ap^* \in \partial F^*(d^*) \quad \textit{and} \quad A^*d^* \in -\partial G(p^*)$$

Deriving our dual problems: In our case, let $V = \mathcal{H}$, $Y = \mathcal{C}(\mathcal{X})$, $A = \Phi^*$ where $A : \mathcal{H} \to \mathcal{C}(\mathcal{X})$. So, $A^* = \Phi : \mathcal{M}(\mathcal{X}) \to \mathcal{H}$. Consider the primal problem as

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle p,\, f\rangle = -\inf_{p\in\mathcal{H}}\{\langle p,\, -f\rangle + \iota_{\|\cdot\|_\infty\leqslant 1}(\Phi^* p)\}.$$

So, $G(p) \overset{\text{def.}}{=} \langle p,\, -f\rangle$ and $F(z) \overset{\text{def.}}{=} \iota_{\|\cdot\|_\infty\leqslant 1}(z)$. We have $G^*(q) \overset{\text{def.}}{=} \iota_{\{-f\}}(q)$, and $F^*(\mu) = |\mu|(\mathcal{X})$. Therefore, the dual problem is

$$-\sup_{\mu\in\mathcal{M}(\mathcal{X})} -|\mu|(\mathcal{X}) + \iota_{\{\Phi\mu = f\}}(\mu) = \inf_{\mu\in\mathcal{M}(\mathcal{X})} |\mu|(\mathcal{X}) \text{ subject to } \Phi\mu = f.$$

Moreover, given primal solution p^* and dual solution μ^*, the optimality condition becomes

$$\Phi^* p^* \in |\mu^*|(\mathcal{X}) \quad \text{and} \quad \Phi\mu^* \in -\partial G^*(p^*) = \{f\}.$$

Similarly, for the case of $\lambda > 0$, consider the primal problem as

$$\sup_{\|\Phi^* p\|_\infty \leqslant 1} \langle p,\, f\rangle - \frac{\lambda}{2}\|p\|^2 = -\inf_{p\in\mathcal{H}}\{\langle p,\, -f\rangle + \frac{\lambda}{2}\|p\|^2 + \iota_{\|\cdot\|_\infty\leqslant 1}(\Phi^* p)\}$$

Let $G(p) = \langle p,\, -f\rangle + \frac{\lambda}{2}\left\|p\right\|^2$ and $F(z) = \iota_{\|\cdot\|_\infty \leqslant 1}(z)$. Then,

$$G^*(q) = \sup_q \langle p,\, q+f\rangle - \frac{\lambda}{2}\left\|p\right\|^2 = \lambda \sup_q \langle p,\, \frac{q+f}{\lambda}\rangle - \frac{1}{2}\left\|p\right\|^2 = \frac{1}{2\lambda}\left\|q+f\right\|^2.$$

Therefore, the dual problem is

$$\inf_{\mu\in\mathcal{M}(\mathcal{X})} \left|\mu\right|(\mathcal{X}) + \frac{1}{2\lambda}\left\|\Phi\mu - f\right\|^2.$$

C Proof of Theorem 4

Proof. Since η_0 is nondegenerate, there exists $c_0, c_2, \varepsilon > 0$ such that

$$\forall x \notin \bigcup_{j=1}^s B_\varepsilon(x_j),\ |\eta_0(x)| \leqslant 1 - c_0 \quad \text{and} \quad \forall x \in B_\varepsilon(x_j),\ \mathrm{sign}(a_{0,j})\eta_0''(x_j) < -c_2. \tag{13}$$

Let $\Phi_X : \mathbb{R}^s \to \mathcal{H}$ be defined by $\Phi_X a = \sum_{i=1}^s a_i \varphi(x_i)$. Recall that $\mu_{a,X}$ solves $(\mathcal{P}_\lambda(y))$ with $y = \Phi\mu_{a_0,X_0} + w$ if and only if

$$\eta_\lambda = \Phi^* p_\lambda \quad \text{where} \quad p_\lambda = \frac{1}{\lambda}(\Phi_{X_0}a_0 + w - \Phi_X a)$$

satisfies $\left\|\eta_\lambda\right\|_\infty \leqslant 1$ and $\eta(x_j) = \mathrm{sign}(a_j)$. Note that p_λ is the unique solution and hence, if η_λ saturates only at X and Φ_X is full rank, then $\mu_{a,X}$ must be unique. To see this, note that if the saturation points of η_λ are included in X, then any solution μ of $(\mathcal{P}_\lambda(y))$ must have support contained in X. Finally, injectivity of Φ_X implies that $\mu = \mu_{a,X}$.

Let $K(x, x') \overset{\text{def.}}{=} \langle\varphi(x),\, \varphi(x')\rangle$ and assume that $h_x \overset{\text{def.}}{=} \partial_1\partial_2 K(x,x) > 0$ for all x. Given $X = \{x_j\}_j$, let $\Gamma_X : \mathbb{R}^{2s} \to \mathcal{H}$ be defined by

$$\Gamma_X \begin{pmatrix} a \\ b \end{pmatrix} = \sum_j a_j \varphi(x_j) + \sum_j b_j \varphi'(x_j).$$

Then, given any $\sigma \in \mathbb{R}^s$, $\Gamma_X^* p = \binom{\sigma}{0_s}$ implies that for all j, $(\Phi^* p)(x_j) = \sigma_j$ and $(\Phi^* p)'(x_j) = 0$.

Construction of a candidate solution

Define $f : \mathbb{R}^s \times \mathcal{X}^s \times \mathbb{R} \times \mathcal{H} \to \mathbb{R}^{2s}$ be defined by

$$f(u,v) = \Gamma_X^*(\Phi_X a - \Phi_{X_0}a_0 - w) + \lambda\begin{pmatrix}\mathrm{sign}(a_0) \\ 0_s\end{pmatrix},$$

where $u = (a, X)$ and $v = (\lambda, w)$. Note that $f(u,v) = 0$ ensures that η_λ satisfies $\eta_\lambda(x_j) = \mathrm{sign}(a_{0,j})$ and $\eta_\lambda'(x_j) = 0$ for all j.

We first remark that f is continuously differentiable, writing $f = (f_\ell)_{\ell=1}^{2s}$ and $z \overset{\text{def.}}{=} \sum_j a_j\varphi(x_j) - y$, we have

$$\begin{aligned}
\partial_{a_k} f_\ell &= \langle\varphi(x_k),\, \varphi(x_\ell)\rangle, \quad \ell = 1,\ldots,s \\
\partial_{x_k} f_\ell &= a_k\langle\varphi'(x_k),\, \varphi(x_\ell)\rangle + \langle z,\, \varphi'(x_k)\rangle\delta_{k\ell}, \quad \ell = 1,\ldots,s \\
\partial_{a_k} f_{\ell+s} &= \langle\varphi(x_k),\, \varphi'(x_\ell)\rangle, \quad \ell = 1,\ldots,s \\
\partial_{x_k} f_{\ell+s} &= a_k\langle\varphi'(x_k),\, \varphi(x_\ell)\rangle + \langle z,\, \varphi''(x_k)\rangle\delta_{k\ell}, \quad \ell = 1,\ldots,s
\end{aligned}$$

Therefore,

$$\partial_u f = (\Gamma_X^*\Gamma_X + E)\,J_a, \quad \text{where} \quad E \overset{\text{def.}}{=} \begin{pmatrix} 0_{2s\times s}, & \begin{pmatrix}\mathrm{diag}\left(\langle z,\, \varphi'(x_j)/a_j\rangle\right)_j \\ \mathrm{diag}\left(\langle z,\, \varphi''(x_j)/a_j\rangle\right)_j\end{pmatrix}\end{pmatrix},\ J_a = \mathrm{diag}((1_s^\top, a^\top)).$$

and

$$\partial_v f(u,v) = \left(\binom{\mathrm{sign}(a_0)}{0_s},\quad \Gamma_X^*\right).$$

Note that $\partial_u f(u,v) = \Gamma_{X_0}^*\Gamma_{X_0}J_{a_0}$ is invertible when $u = u_0 = (a_0, X_0)$ and $v = (0,0)$. Hence, by the implicit function theorem, there exists $0 \in V \subset \mathbb{R}\times\mathcal{H}$ and $u_0 \in U \subset \mathbb{R}^s \times \mathcal{X}^s$ such that $g : V \to U$ is differentiable and $f(u,v) = 0$ if and only if $u = g(v)$. On V, we have

$$\mathrm{d}g(v) = -(\partial_u f(g(v), v))^{-1}\partial_v f(g(v), v).$$

Candidate solution is a true solution Given $(a, X) = g((\lambda, w))$ where $(\lambda, w) \in V$, we have a candidate certificate $\eta_{\lambda,w} = \Phi^* p_{\lambda,w}$ with

$$p_{\lambda,w} \overset{\text{def.}}{=} \frac{1}{\lambda}\left(\Phi_X a - \Phi_{X_0} a_0 - w\right).$$

Note that $\mu_{a,X}$ is indeed a solution of $(\mathcal{P}_\lambda(y))$ with $y = \Phi\mu_{a_0,X_0} + w$ if $\eta_{\lambda,w} \overset{\text{def.}}{=} \Phi^* p_{\lambda,w}$ is nondegenerate. To this end, we can show that

$$\eta_{\lambda,w} = \eta_0 + \Phi^*\Pi_X \frac{w}{\lambda} + \frac{1}{\lambda}\Phi^*\Pi_X\Phi_{X_0}a_0$$

where Π_X is the orthogonal projection onto $\mathrm{Im}(\Gamma_X)^\perp$. By Taylor expansion of $\varphi(x_{0,i})$ about x_i, we have

$$\varphi(x_{0,i}) = \varphi(x_i) - \varphi'(x_i)(x_i - x_{0,i}) + \int_0^1 \frac{1}{2}\varphi''(x_i + t(x_{0,i} - x_i))(x_{0,i} - x_i)^2 \mathrm{d}t$$

Therefore,

$$\Pi_X\Phi_X a = \Pi_X \sum_i a_i \int_0^1 \frac{1}{2}\varphi''(x_i + t(x_{0,i} - x_i))(x_{0,i} - x_i)^2 \mathrm{d}t$$

and $\|\Pi_X\Phi_X a\| \leqslant \frac{L_2}{2}\|a_0\|\|X - X_0\|^2$, and for $k = 0, 2$,

$$\left|\eta_{\lambda,w}^{(k)} - \eta_0^{(k)}\right| \leqslant \frac{L_0}{\lambda}\left(\|w\| + \frac{L_2}{2}\|a_0\|\|X - X_0\|^2\right)$$

and since g is differentiable, we have $\|X - X_0\| \lesssim \lambda + \|w\|$ and assuming that $\|w\| \leqslant c_*\lambda$ and $\lambda \leqslant \lambda_*$, we obtain

$$\left|\eta_{\lambda,w}^{(k)} - \eta_0^{(k)}\right| \lesssim \frac{L_0}{\lambda}\left(\|w\| + \frac{L_2}{2}\|a_0\|(\|w\| + \lambda)^2\right) \lesssim c_* + \lambda_*$$

Therefore, $\eta_{\lambda,w}$ is nondegenerate provided that c_* and λ_* are sufficiently small.

Although we shall not do this here, we remark that a quantitative version of this result can be obtained by bounding the size of the neighbourhood V on which g is well-defined.

□

References

[ADCG15] Jean-Marc Azais, Yohann De Castro, and Fabrice Gamboa. Spike detection from inaccurate samplings. *Applied and Computational Harmonic Analysis*, 38(2):177–195, 2015.

[Bac17] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[BO04] Martin Burger and Stanley Osher. Convergence rates of convex variational regularization. *Inverse problems*, 20(5):1411, 2004.

[BP13] Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.

[BSR17] Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.

[CFG13] Emmanuel J Candès and Carlos Fernandez-Granda. Super-resolution from noisy data. *Journal of Fourier Analysis and Applications*, 19(6):1229–1254, 2013.

[CFG14] Emmanuel J Candès and Carlos Fernandez-Granda. Towards a mathematical theory of super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.

[DCG12] Yohann De Castro and Fabrice Gamboa. Exact reconstruction using beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.

[DDP17] Quentin Denoyelle, Vincent Duval, and Gabriel Peyré. Support recovery for sparse super-resolution of positive measures. *Journal of Fourier Analysis and Applications*, 23(5):1153–1194, 2017.

[DDPS18] Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The sliding frank-wolfe algorithm and its application to super-resolution microscopy. *arXiv preprint arXiv:1811.06416*, 2018.

[DP15] Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.

[Dum07] Bogdan Dumitrescu. *Positive trigonometric polynomials and signal processing applications*, volume 103. Springer, 2007.

[LF16] Wenjing Liao and Albert Fannjiang. Music for single-snapshot spectral estimation: Stability and super-resolution. *Applied and Computational Harmonic Analysis*, 40(1):33–67, 2016.

[Tan15] Gongguo Tang. Resolution limits for atomic decompositions via markov-bernstein type inequalities. In *Sampling Theory and Applications (SampTA), 2015 International Conference on*, pages 548–552. IEEE, 2015.

[TBSR12] Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressive sensing off the grid. In *Communication, Control, and Computing (Allerton), 2012 50th Annual Allerton Conference on*, pages 778–785. IEEE, 2012.