Sparsity and High Dimensions

Lecture notes, 2027

Yohann De Castro
Institut Camille Jordan, École Centrale de Lyon

Version of 16 November 2026

Contents

1 Sparse recovery on a grid and dual certificates **3**
- 1.1 Two questions . . . 3
- 1.2 The sparsest solution . . . 4
- 1.3 Basis pursuit . . . 6
- 1.4 Dual certificates on the grid . . . 7
- 1.5 Classical results that we do not prove . . . 10

2 From the grid to measures **13**
- 2.1 Support recovery on the grid at low noise . . . 13
- 2.2 Why the grid fails . . . 14
- 2.3 The off-the-grid model . . . 16
- 2.4 Radon measures . . . 18
- 2.5 The BLASSO: existence, subdifferential, duality . . . 19

3 Exact recovery and the construction of certificates **25**
- 3.1 Exact recovery without noise . . . 25
- 3.2 The vanishing derivative pre-certificate . . . 27
- 3.3 Local positive curvature gives a certificate . . . 29

4 Noise: prediction, localisation, support stability **36**
- 4.1 Prediction and localisation . . . 36
- 4.2 The two examples . . . 38
- 4.3 Support stability at low noise . . . 40

5 The generalised BLASSO and Fenchel-Young losses **47**
- 5.1 Beyond the quadratic fidelity . . . 47
- 5.2 Fenchel-Young losses . . . 48
- 5.3 The duality gap . . . 50
- 5.4 Error bounds . . . 51
- 5.5 Examples . . . 54
- 5.6 Support stability for the generalised BLASSO . . . 55

6 Algorithms for the BLASSO **59**
- 6.1 The duality gap as a stopping test . . . 59
- 6.2 Frank-Wolfe and sliding Frank-Wolfe . . . 61
- 6.3 Conic particle gradient descent . . . 63
- 6.4 FastPart: stochastic particle descent . . . 65
- 6.5 Fast Spawn and Prune . . . 67

A Support stability without the implicit function theorem **70**

How to read these notes

The course has five lectures of three hours. Lectures 1 and 2 treat sparse recovery on a grid and then move to measures. Lectures 3 and 4 give the guarantees for the BLASSO: exact recovery, prediction and localisation with noise, support stability at low noise. Lecture 5 treats the generalised BLASSO with Fenchel-Young losses. Chapter 6 describes algorithms for the BLASSO. It is not taught in class and it supports the projects. The two running examples are weighted Fourier measurements on the torus and the estimation of the means of a Gaussian mixture.

The six chapters and the appendix are written in full.

Chapter 1

Sparse recovery on a grid and dual certificates

1.1 Two questions

The course starts from two questions.

The first question comes from astronomy. Two stars sit at positions x_1, x_2 of the torus $\mathbb{T} = \mathbb{R}/\mathbb{Z}$, with brightness $a_1, a_2 > 0$. The instrument has a cut-off frequency f_c. Up to noise, it gives the Fourier coefficients $\int_{\mathbb{T}} e^{-2i\pi kx} \, \mathrm{d}\mu^0(x)$, $|k| \leqslant f_c$, of the measure $\mu^0 = a_1\delta_{x_1} + a_2\delta_{x_2}$. The image built from these coefficients turns each star into a spot of width of order $1/f_c$. When $|x_1 - x_2|$ is below this width, the Rayleigh criterion says that the two spots look like one star. Can a method still find x_1, x_2, a_1, a_2? How does the error depend on the noise level and on the distance between the stars?

The second question comes from statistics. We observe n independent samples $z_1, \dots, z_n$ of the mixture $\sum_{i=1}^s a_i \mathcal{N}(x_i, \sigma^2)$, where σ is known, $a_i > 0$ and $\sum_i a_i = 1$. We want the means x_i and the weights a_i, that is the mixing measure $\mu^0 = \sum_i a_i \delta_{x_i}$. For every frequency $\omega \in \mathbb{R}$,

$$\mathbb{E}[e^{i\omega z_1}] = e^{-\sigma^2\omega^2/2} \sum_{i=1}^{s} a_i e^{i\omega x_i},$$

which is linear in μ^0. So the empirical characteristic function $\omega \mapsto \frac{1}{n}\sum_{k=1}^n e^{i\omega z_k}$ is a noisy linear observation of μ^0. The noise has a standard deviation at most $n^{-1/2}$ at each frequency. How many samples and how many frequencies are needed, and how close can two means be?

In both questions we observe $y = \Phi\mu^0 + w$. The unknown μ^0 is a measure with few atoms on a continuous set $\mathcal{X}$, the map Φ is linear in μ^0, and w is the noise. The data depend linearly on the amplitudes, but not on the positions. There are two ways to treat the positions.

- On a grid. We fix a finite set $G \subset \mathcal{X}$ of candidate positions and look only for the amplitudes on G. This is the recovery of a sparse vector of $\mathbb{R}^N$. This chapter treats this setting.
- Off the grid. We look for μ^0 among all measures on $\mathcal{X}$, and the ℓ^1 norm becomes the total variation norm. Lectures 2 to 5 follow this route.

The grid has two defects. The true positions are in general not grid points. If we refine the grid, neighbouring atoms become almost collinear, and the coherence and restricted isometry guarantees are lost (Remark 1.8 and Section 2.2). The dual certificates of Section 1.4 do not have this defect: their proofs carry over to measures in Lectures 2 to 4.

1.2 The sparsest solution

Let $G = \{t_1, \dots, t_N\}$ be a finite set, called the grid. A measure on G is $\mu = \sum_{j=1}^N a_j \delta_{t_j}$ with $a \in \mathbb{R}^N$. Its support is $\operatorname{supp}(\mu) = \{t_j \,;\, a_j \neq 0\}$, its total variation is $\|\mu\|_{\mathrm{TV}} = \sum_j |a_j|$, and $\|\mu\|_0$ is the cardinal of its support. The measure is s-sparse if $\|\mu\|_0 \leqslant s$, and Σ_s is the set of s-sparse measures. Let $\mathbb{H}$ be a real Hilbert space of finite dimension m and let $\varphi_t \in \mathbb{H}$ for $t \in G$. The forward operator is $\Phi\mu \overset{\text{def.}}{=} \sum_j a_j \varphi_{t_j}$. Its adjoint maps $p \in \mathbb{H}$ to the function $\Phi^\star p : t \in G \mapsto \langle \varphi_t, p \rangle_{\mathbb{H}}$.

We use the language of measures from the start, so that Lectures 2 to 4 can reuse it. The space $\mathcal{M}(G)$ of measures on G is identified with $\mathbb{R}^N$ through $\mu \mapsto a$. For $\mu = \sum_j a_j \delta_{t_j}$, $S \subset G$ and $\eta : G \to \mathbb{R}$, we set

$$|\mu| \overset{\text{def.}}{=} \sum_j |a_j| \, \delta_{t_j}, \qquad |\mu|(S) \overset{\text{def.}}{=} \sum_{t_j \in S} |a_j|, \qquad \mu_{|S} \overset{\text{def.}}{=} \sum_{t_j \in S} a_j \delta_{t_j},$$

$$\langle \eta, \mu \rangle_{\mathcal{M}} \overset{\text{def.}}{=} \sum_j \eta(t_j)\, a_j = \int_G \eta \, \mathrm{d}\mu, \qquad \|\eta\|_\infty \overset{\text{def.}}{=} \max_{t \in G} |\eta(t)|.$$

So $\|\mu\|_{\mathrm{TV}} = |\mu|(G)$. The sign of μ is the function σ_μ defined on $\operatorname{supp}(\mu)$ by $\sigma_\mu(t_j) = \operatorname{sign}(a_j)$, so that $\langle \eta, \mu \rangle_{\mathcal{M}} = \int \eta \, \sigma_\mu \, \mathrm{d}\,|\mu|$. Three facts are used all the time:

$$\langle \Phi^\star p, \mu \rangle_{\mathcal{M}} = \langle p, \Phi\mu \rangle_{\mathbb{H}}, \qquad |\langle \eta, \mu \rangle_{\mathcal{M}}| \leqslant \|\eta\|_\infty \, \|\mu\|_{\mathrm{TV}}, \tag{1.1}$$

$$\|\mu\|_{\mathrm{TV}} - \langle \eta, \mu \rangle_{\mathcal{M}} = \int_G (1 - \eta\, \sigma_\mu) \, \mathrm{d}\,|\mu| \geqslant \int_G (1 - |\eta|) \, \mathrm{d}\,|\mu|. \tag{1.2}$$

The first one is the definition of the adjoint, the second one is the triangle inequality, and the last one holds because $\eta\, \sigma_\mu \leqslant |\eta|$.

Definition 1.1 (Sparsest solution)**.** Given $y \in \mathbb{H}$, the problem $\mathcal{P}_{\ell^0}(y)$ is

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

The feasible set is empty or an affine subspace, and $\|\cdot\|_0$ takes finitely many values. So $\mathcal{P}_{\ell^0}(y)$ has a solution as soon as $y \in \operatorname{Im}\Phi$. The first question is when this solution is unique.

Theorem 1.1 (Uniqueness of sparse solutions)**.** *Let $s \geqslant 1$. The following are equivalent.*

1. *Every $\mu \in \Sigma_s$ is the unique s-sparse solution of $\Phi\nu = \Phi\mu$.*
2. $\operatorname{Ker}\Phi \cap \Sigma_{2s} = \{0\}$.
3. *For every $T \subset G$ with $|T| \leqslant 2s$, the family $(\varphi_t)_{t \in T}$ is linearly independent.*

Proof. (2) implies (1). Let $\mu, \nu \in \Sigma_s$ with $\Phi\nu = \Phi\mu$. Then $\nu - \mu \in \operatorname{Ker}\Phi$, and $\operatorname{supp}(\nu - \mu) \subset \operatorname{supp}(\nu) \cup \operatorname{supp}(\mu)$ has at most $2s$ points. So $\nu - \mu \in \operatorname{Ker}\Phi \cap \Sigma_{2s} = \{0\}$.

(1) implies (2). Let $\nu \in \operatorname{Ker}\Phi \cap \Sigma_{2s}$. Split $\operatorname{supp}(\nu) = T_1 \cup T_2$ with $T_1 \cap T_2 = \emptyset$ and $|T_1|, |T_2| \leqslant s$. Set $\mu \overset{\text{def.}}{=} \nu_{|T_1}$ and $\mu' \overset{\text{def.}}{=} -\nu_{|T_2}$. Both are s-sparse, $\nu = \mu - \mu'$, and $\Phi\mu' = \Phi\mu$ because $\Phi\nu = 0$. By (1), $\mu' = \mu$. Since μ and μ' have disjoint supports, $\mu = \mu' = 0$. Hence $\nu = 0$.

(2) is equivalent to (3). Let $T \subset G$. A measure ν with $\operatorname{supp}(\nu) \subset T$ reads $\nu = \sum_{t \in T} b_t \delta_t$, and $\Phi\nu = \sum_{t \in T} b_t \varphi_t$. So the family $(\varphi_t)_{t \in T}$ is linearly dependent if and only if there is a nonzero $\nu \in \operatorname{Ker}\Phi$ with $\operatorname{supp}(\nu) \subset T$. A measure is in Σ_{2s} if and only if its support is contained in a set T with $|T| \leqslant 2s$. So (2) fails if and only if (3) fails. □

Remark 1.1 (A lower bound on m). Under these conditions, every $\mu \in \Sigma_s$ is the unique solution of $\mathcal{P}_{\ell^0}(\Phi\mu)$. Indeed μ is feasible, a solution ν satisfies $\|\nu\|_0 \leqslant \|\mu\|_0 \leqslant s$, and (1) gives $\nu = \mu$. If $N \geqslant 2s$, item (3) gives $2s$ linearly independent vectors in $\mathbb{H}$, so $m \geqslant 2s$. If $N < 2s$, item (3) says that Φ is injective, so $m \geqslant N$. This last case is the reason why item (3) is written with $|T| \leqslant 2s$: for $N < 2s$ there is no family of $2s$ atoms.

The next result shows that $m = 2s$ is enough.

Theorem 1.2 ($2s$ measurements are enough)**.** *Let $N \geqslant 2s$. There exist $m = 2s$ and $(\varphi_t)_{t\in G}$ in $\mathbb{R}^{2s}$ such that every $\mu \in \Sigma_s$ is the unique s-sparse solution of $\Phi\nu = \Phi\mu$.*

Proof. Let $\mathbb{H} = \mathbb{R}^{2s}$ with the Euclidean scalar product. Take pairwise distinct reals $u_1, \dots, u_N$ and set $\varphi_{t_j} \stackrel{\text{def.}}{=} (1, u_j, u_j^2, \dots, u_j^{2s-1})$. Let $T = \{t_{j_1}, \dots, t_{j_{2s}}\}$ be a set of $2s$ grid points. The matrix with columns $\varphi_{t_{j_1}}, \dots, \varphi_{t_{j_{2s}}}$ is a Vandermonde matrix. Its determinant is $\prod_{k<l}(u_{j_l} - u_{j_k})$, which is nonzero. So these $2s$ vectors are linearly independent. Since $N \geqslant 2s$, every $T' \subset G$ with $|T'| \leqslant 2s$ is contained in a set of exactly $2s$ points, and a subfamily of a linearly independent family is linearly independent. So item (3) of Theorem 1.1 holds, hence item (1). □

Remark 1.2 (Prony's method). The same count holds without a grid. Let $\mu^0 = \sum_{i=1}^{s} a_i \delta_{x_i}$ be a measure on $\mathbb{T}$ with pairwise distinct x_i and $a_i \neq 0$. Its Fourier coefficients of indices $0, 1, \dots, 2s-1$ determine it, and Prony's method computes it from them [1]. The Vandermonde structure appears with $z_i = e^{-2\mathrm{i}\pi x_i}$ in place of u_j. The method is exact on noiseless data, but it is very sensitive to noise when the z_i are close [2].

So few measurements are enough in principle. Computing the sparsest solution is another matter. We recall the vocabulary of complexity theory, which we do not develop. A decision problem is a family of instances, each with the answer yes or no. A polynomial reduction from a problem A to a problem B maps the instances of A to instances of B, in time polynomial in the size of the instance, and keeps the answer. A problem B is NP-hard if every problem of the class NP has a polynomial reduction to B. If an NP-hard problem has a polynomial reduction to B, then B is NP-hard. We admit that the following problem is NP-complete, hence NP-hard.

> Exact Cover by 3-Sets. Instance: an integer $m \geqslant 1$ and subsets $C_1, \dots, C_N$ of $U = \{1, \dots, m\}$ with $|C_j| = 3$ for all j. Question: is there $J \subset \{1, \dots, N\}$ such that the sets C_j, $j \in J$, are pairwise disjoint and their union is U?

Such a J is called an exact cover.

Theorem 1.3 (Complexity)**.** *Consider the following decision problem. An instance is given by integers $m, N \geqslant 1$ and $k \geqslant 0$, vectors $\varphi_{t_1}, \dots, \varphi_{t_N} \in \mathbb{Q}^m$ and $y \in \mathbb{Q}^m$, with $\mathbb{H} = \mathbb{R}^m$ and its Euclidean scalar product. The question is whether there is $\mu \in \mathcal{M}(G)$ with $\Phi\mu = y$ and $\|\mu\|_0 \leqslant k$. This problem is NP-hard. In this sense, the problem $\mathcal{P}_{\ell^0}(y)$ is NP-hard.*

Proof. Let $(m, C_1, \dots, C_N)$ be an instance of Exact Cover by 3-Sets. Set $\varphi_{t_j} \stackrel{\text{def.}}{=} \mathbf{1}_{C_j} \in \mathbb{R}^m$, the indicator vector of C_j, $y \stackrel{\text{def.}}{=} (1, \dots, 1) \in \mathbb{R}^m$ and $k \stackrel{\text{def.}}{=} \lfloor m/3 \rfloor$. These data have $mN + m$ entries in $\{0, 1\}$, so they are computed in time polynomial in m and N. We show that the answer to the question of the theorem is yes if and only if an exact cover exists.

Let J be an exact cover. The sets C_j, $j \in J$, are pairwise disjoint, have 3 elements and cover U, so $3\,|J| = m$ and $|J| = k$. Set $\mu \stackrel{\text{def.}}{=} \sum_{j\in J} \delta_{t_j}$. For $i \in U$, the coordinate i of $\Phi\mu$ is the number of $j \in J$ with $i \in C_j$, which is 1. So $\Phi\mu = y$ and $\|\mu\|_0 = k$.

Conversely, let $\mu = \sum_j a_j \delta_{t_j}$ with $\Phi\mu = y$ and $\|\mu\|_0 \leqslant k$. Let $J \stackrel{\text{def.}}{=} \{j\ ;\ a_j \neq 0\}$. For $i \in U$, the coordinate i of $\Phi\mu = y$ reads $\sum_{j\in J, i\in C_j} a_j = 1$. This sum is not empty, so $i \in C_j$ for some $j \in J$. Hence $\bigcup_{j\in J} C_j = U$ and

$$m = \Big|\bigcup_{j\in J} C_j\Big| \leqslant \sum_{j\in J} |C_j| = 3\,|J| \leqslant 3k \leqslant m.$$

All these inequalities are equalities. The equality between the cardinal of the union and the sum of the cardinals means that the sets C_j, $j \in J$, are pairwise disjoint. So J is an exact cover. Only the support of μ was used: the amplitudes may have any sign. □

Remark 1.3. The theorem is a worst-case statement over all instances. An algorithm that returns a solution $\hat{\mu}$ of $\mathcal{P}_{\ell^0}(y)$ for rational data, or reports that $y \notin \operatorname{Im}\Phi$, answers the decision problem: compare $\|\hat{\mu}\|_0$ with k. So, unless P = NP, no such algorithm runs in time polynomial in the size of the data. The theorem says nothing about the structured operators of the course.

1.3 Basis pursuit

The difficulty of $\mathcal{P}_{\ell^0}(y)$ comes from the function $\|\cdot\|_0$, which is not convex. Basis pursuit replaces it by the total variation, which is a norm [3]. The LASSO is its penalised version for noisy data [4].

Definition 1.2 (Basis pursuit and LASSO on the grid)**.** Given $y \in \mathbb{H}$ and $\lambda > 0$, the problems $\mathcal{P}_0(y)$ and $\mathcal{P}_\lambda(y)$ are

$$\min_{\mu \in \mathcal{M}(G)} \|\mu\|_{\mathrm{TV}} \quad \text{subject to} \quad \Phi\mu = y, \qquad \text{and} \qquad \min_{\mu \in \mathcal{M}(G)} \tfrac{1}{2}\|\Phi\mu - y\|_{\mathbb{H}}^2 + \lambda\|\mu\|_{\mathrm{TV}}.$$

If $y \in \operatorname{Im}\Phi$, the problem $\mathcal{P}_0(y)$ has a solution. The feasible set is closed and nonempty, and the sublevel sets of $\|\cdot\|_{\mathrm{TV}}$ on it are compact, since $\mathcal{M}(G)$ is identified with $\mathbb{R}^N$. Existence for $\mathcal{P}_\lambda(y)$ is item 1 of Proposition 1.2.

Definition 1.3 (Null space property)**.** The operator Φ has the null space property of order s if, for every nonzero $\nu \in \operatorname{Ker}\Phi$ and every $S \subset G$ with $|S| \leqslant s$, $|\nu|(S) < |\nu|(G \setminus S)$.

If Φ is injective, the property holds for every s. Figure 1.1 shows the geometry in the smallest case, $N = 2$ and $m = 1$. The feasible set of $\mathcal{P}_0(\Phi\mu^0)$ is the line $\mu^0 + \operatorname{Ker}\Phi$. The measure μ^0 is the unique solution if and only if this line meets the ball $\{\nu\ ;\ \|\nu\|_{\mathrm{TV}} \leqslant \|\mu^0\|_{\mathrm{TV}}\}$ at μ^0 only.

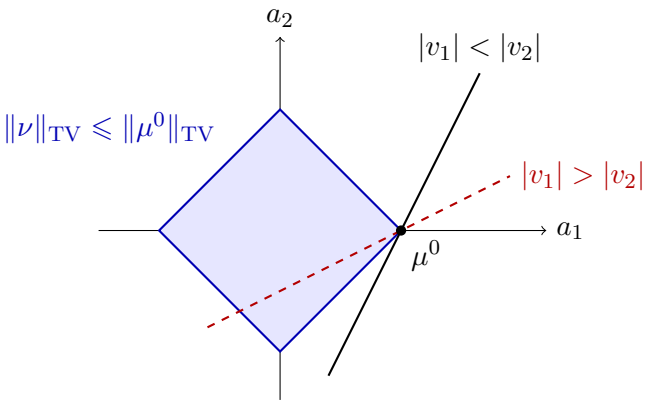


Figure 1.1: Grid with $N = 2$ points, one measurement ($m = 1$) and $\mu^0 = \delta_{t_1}$, in the coordinates (a_1, a_2) of $\nu = a_1\delta_{t_1} + a_2\delta_{t_2}$. The feasible set is the line $\mu^0 + \operatorname{Ker}\Phi$, with $\operatorname{Ker}\Phi = \mathbb{R}v$. If $|v_1| < |v_2|$ (solid line), it touches the ball $\{\nu\ ;\ \|\nu\|_{\mathrm{TV}} \leqslant \|\mu^0\|_{\mathrm{TV}}\}$ only at μ^0, and μ^0 is the unique solution of $\mathcal{P}_0(\Phi\mu^0)$. If $|v_1| > |v_2|$ (dashed line), it enters the ball and μ^0 is not a solution.

On the solid line, $|v|(\{t_1\}) < |v|(\{t_2\})$. The null space property asks this inequality for every vector of the kernel and every small set.

Theorem 1.4 (Exact recovery by basis pursuit)**.** *Every $\mu \in \Sigma_s$ is the unique solution of $\mathcal{P}_0(\Phi\mu)$ if and only if Φ has the null space property of order s.*

Proof. Sufficiency. Let $\mu \in \Sigma_s$ and $S \overset{\text{def.}}{=} \operatorname{supp}(\mu)$, so $|S| \leqslant s$. Let $\nu \neq \mu$ with $\Phi\nu = \Phi\mu$, and set $v \overset{\text{def.}}{=} \nu - \mu$. Then $v \in \operatorname{Ker}\Phi$ and $v \neq 0$. Since μ vanishes outside S, $\nu = \mu + v$ on S and $\nu = v$ on $G \setminus S$. The triangle inequality at each point of S gives

$$\|\nu\|_{\mathrm{TV}} = |\mu + v|(S) + |v|(G \setminus S) \geqslant |\mu|(S) - |v|(S) + |v|(G \setminus S) > |\mu|(S) = \|\mu\|_{\mathrm{TV}},$$

where the strict inequality is the null space property. So μ is the unique solution.

Necessity. Let $v \in \operatorname{Ker}\Phi$ be nonzero and $S \subset G$ with $|S| \leqslant s$. Set $\mu \overset{\text{def.}}{=} v_{|S}$, which is s-sparse, and $\nu \overset{\text{def.}}{=} -v_{|G\setminus S}$. Then $\Phi\mu - \Phi\nu = \Phi v = 0$, so ν is feasible for $\mathcal{P}_0(\Phi\mu)$. If $\nu = \mu$, these two measures have disjoint supports, so both vanish and $v = \mu - \nu = 0$, which is excluded. So $\nu \neq \mu$. Since μ is the unique solution, $\|\nu\|_{\mathrm{TV}} > \|\mu\|_{\mathrm{TV}}$, that is $|v|(G \setminus S) > |v|(S)$. □

Remark 1.4. The null space property of order s implies item (2) of Theorem 1.1. Indeed, let $\nu \in \operatorname{Ker}\Phi \cap \Sigma_{2s}$ be nonzero, and split its support into S_1, S_2 of size at most s. Then $|\nu|(S_1) < |\nu|(S_2)$ and $|\nu|(S_2) < |\nu|(S_1)$, a contradiction. In Figure 1.1 the null space property of order one never holds: it asks both $|v_1| < |v_2|$ and $|v_2| < |v_1|$. Still, on the solid line, the single measure δ_{t_1} is recovered. This is the point of view of the next section.

1.4 Dual certificates on the grid

The null space property is a condition on all s-sparse measures at once, and it is hard to check. A dual certificate is a function $\eta = \Phi^\star p$, with $p \in \mathbb{H}$, attached to one measure μ^0. It proves that this μ^0 is recovered, and it controls the error of the LASSO. This is the tool that survives without a grid. The first lemma explains the name: the certificates are the subgradients of the total variation that lie in the image of $\Phi^\star$.

Lemma 1.1 (Subdifferential of the total variation on the grid)**.** *Let $\mu \in \mathcal{M}(G)$ and $\eta : G \to \mathbb{R}$. Then $\|\nu\|_{\mathrm{TV}} \geqslant \|\mu\|_{\mathrm{TV}} + \langle \eta, \nu - \mu\rangle_{\mathcal{M}}$ for all $\nu \in \mathcal{M}(G)$ if and only if $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \mu\rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$, that is $\eta(t_j) = \operatorname{sign}(a_j)$ on $\operatorname{supp}(\mu)$. We write $\eta \in \partial\|\mu\|_{\mathrm{TV}}$.*

Proof. Assume the inequality for all ν. With $\nu = 0$ it gives $0 \geqslant \|\mu\|_{\mathrm{TV}} - \langle \eta, \mu\rangle_{\mathcal{M}}$. With $\nu = 2\mu$ it gives $2\|\mu\|_{\mathrm{TV}} \geqslant \|\mu\|_{\mathrm{TV}} + \langle \eta, \mu\rangle_{\mathcal{M}}$. Hence $\langle \eta, \mu\rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$, and the inequality becomes $\|\nu\|_{\mathrm{TV}} \geqslant \langle \eta, \nu\rangle_{\mathcal{M}}$ for all ν. With $\nu = \pm\delta_t$ it gives $1 \geqslant \pm\eta(t)$, so $|\eta(t)| \leqslant 1$ for all $t \in G$.

Conversely, assume $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \mu\rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$. For all ν, by (1.1), $\|\mu\|_{\mathrm{TV}} + \langle \eta, \nu - \mu\rangle_{\mathcal{M}} = \langle \eta, \nu\rangle_{\mathcal{M}} \leqslant \|\nu\|_{\mathrm{TV}}$.

Last, assume $\|\eta\|_\infty \leqslant 1$. By (1.2), $\|\mu\|_{\mathrm{TV}} - \langle \eta, \mu\rangle_{\mathcal{M}} = \sum_j |a_j|\,(1 - \eta(t_j)\operatorname{sign}(a_j))$, and each term is nonnegative. The sum vanishes if and only if $\eta(t_j)\operatorname{sign}(a_j) = 1$ whenever $a_j \neq 0$, that is $\eta(t_j) = \operatorname{sign}(a_j)$ on $\operatorname{supp}(\mu)$. □

For $\mu = 0$ the condition reduces to $\|\eta\|_\infty \leqslant 1$. The proof uses three kinds of test measures: 0, 2μ and $\pm\delta_t$. They exist in $\mathcal{M}(\mathcal{X})$ for any set $\mathcal{X}$, so the proof carries over to measures (Proposition 2.3).

Proposition 1.1 (A certificate gives exact recovery)**.** *Let $\mu^0 = \sum_{i=1}^s a_i\delta_{x_i}$ with $x_i \in G$ distinct and $a_i \neq 0$. Assume that $(\varphi_{x_i})_{i\leqslant s}$ is linearly independent and that there is $p \in \mathbb{H}$ such that $\eta = \Phi^\star p$ satisfies $\eta(x_i) = \operatorname{sign}(a_i)$ for all i and $|\eta(t)| < 1$ for $t \in G \setminus \operatorname{supp}(\mu^0)$. Then μ^0 is the unique solution of $\mathcal{P}_0(\Phi\mu^0)$.*

Proof. The assumptions give $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \mu^0\rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}$, so $\eta \in \partial\|\mu^0\|_{\mathrm{TV}}$. Let ν be feasible, that is $\Phi\nu = \Phi\mu^0$. By (1.1),

$$\|\nu\|_{\mathrm{TV}} \geqslant \langle \eta, \nu\rangle_{\mathcal{M}} = \langle p, \Phi\nu\rangle_{\mathbb{H}} = \langle p, \Phi\mu^0\rangle_{\mathbb{H}} = \langle \eta, \mu^0\rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}. \tag{1.3}$$

So μ^0, which is feasible, is a solution. Let ν be a solution. Then $\|\nu\|_{\mathrm{TV}} = \|\mu^0\|_{\mathrm{TV}}$, and all the terms of (1.3) are equal. By (1.2),

$$0 = \|\nu\|_{\mathrm{TV}} - \langle \eta, \nu\rangle_{\mathcal{M}} \geqslant \int_G (1 - |\eta|)\,\mathrm{d}\,|\nu| \geqslant 0.$$

So $|\eta(t)| = 1$ for every $t \in \operatorname{supp}(\nu)$. Since $|\eta| < 1$ outside $\operatorname{supp}(\mu^0)$, this gives $\operatorname{supp}(\nu) \subset \operatorname{supp}(\mu^0)$, and $\nu = \sum_i b_i\delta_{x_i}$ for some $b \in \mathbb{R}^s$. Then $0 = \Phi(\nu - \mu^0) = \sum_i (b_i - a_i)\varphi_{x_i}$, and linear independence gives $b = a$, that is $\nu = \mu^0$. □

The proof only uses the inequality (1.3), which is weak duality: $\|\nu\|_{\mathrm{TV}} \geqslant \langle p, \Phi\nu\rangle_{\mathbb{H}}$ for every ν when $\|\Phi^\star p\|_\infty \leqslant 1$. It does not need to know beforehand that $\mathcal{P}_0(\Phi\mu^0)$ has a solution.

For the LASSO, the next proposition gives the optimality condition and the dual problem. The constraint set of the dual problem is

$$C \stackrel{\text{def.}}{=} \{p \in \mathbb{H} \,;\, \|\Phi^\star p\|_\infty \leqslant 1\} = \bigcap_{t\in G} \{p \in \mathbb{H} \,;\, |\langle \varphi_t, p\rangle_{\mathbb{H}}| \leqslant 1\}\,.$$

It is an intersection of closed convex sets, so it is closed and convex, and it contains 0. The dual solution $\hat{p}$ is the main tool of Section 2.1 and of Lecture 4.

Proposition 1.2 (Optimality conditions and duality for the LASSO)**.** *Let $\lambda > 0$ and $y \in \mathbb{H}$. Let $C \stackrel{\text{def.}}{=} \{p \in \mathbb{H} \,;\, \|\Phi^\star p\|_\infty \leqslant 1\}$.*

1. *$\mathcal{P}_\lambda(y)$ has a solution.*
2. *$\hat{\mu}$ is a solution if and only if $\Phi^\star\hat{p} \in \partial\|\hat{\mu}\|_{\mathrm{TV}}$ with $\hat{p} \stackrel{\text{def.}}{=} (y - \Phi\hat{\mu})/\lambda$.*
3. *$\hat{p}$ is the projection of y/λ onto C. It is the same for all solutions.*
4. *$\min \mathcal{P}_\lambda(y) = \max_{p\in C} \lambda\langle p, y\rangle_{\mathbb{H}} - \frac{\lambda^2}{2}\|p\|^2_{\mathbb{H}}$, and the maximum is attained at $\hat{p}$ only.*

Proof. Let $J(\mu) \stackrel{\text{def.}}{=} \frac{1}{2}\|\Phi\mu - y\|^2_{\mathbb{H}} + \lambda\|\mu\|_{\mathrm{TV}}$ and $D(p) \stackrel{\text{def.}}{=} \lambda\langle p, y\rangle_{\mathbb{H}} - \frac{\lambda^2}{2}\|p\|^2_{\mathbb{H}}$. All norms and scalar products are those of $\mathbb{H}$.

(1) Through $\mu \mapsto a$, J is a continuous function on $\mathbb{R}^N$. The set $K \stackrel{\text{def.}}{=} \{\mu \,;\, J(\mu) \leqslant J(0)\}$ contains 0 and is closed. It is bounded, since $\lambda\|\mu\|_{\mathrm{TV}} \leqslant J(\mu) \leqslant J(0)$ on K. So K is compact, and J attains its minimum on K at some $\hat{\mu}$. Outside K, $J > J(0) \geqslant J(\hat{\mu})$. So $\hat{\mu}$ is a solution.

(2) Let $\hat{\mu}$ be a solution, $\nu \in \mathcal{M}(G)$ and $\tau \in (0, 1]$. Expanding the square, with $\Phi\hat{\mu} - y = -\lambda\hat{p}$,

$$\tfrac{1}{2}\|\Phi(\hat{\mu} + \tau\nu) - y\|^2 - \tfrac{1}{2}\|\Phi\hat{\mu} - y\|^2 = -\tau\lambda\langle\hat{p}, \Phi\nu\rangle + \tfrac{\tau^2}{2}\|\Phi\nu\|^2\,.$$

Since $\hat{\mu}+\tau\nu = (1-\tau)\hat{\mu}+\tau(\hat{\mu}+\nu)$, the convexity of the total variation gives $\|\hat{\mu}+\tau\nu\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}} \leqslant \tau(\|\hat{\mu}+\nu\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}})$. So the optimality $J(\hat{\mu} + \tau\nu) \geqslant J(\hat{\mu})$ gives, after division by $\tau\lambda$,

$$0 \leqslant -\langle\hat{p}, \Phi\nu\rangle + \tfrac{\tau}{2\lambda}\|\Phi\nu\|^2 + \|\hat{\mu} + \nu\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}}.$$

Let $\tau \to 0$ and use (1.1): $\|\hat{\mu} + \nu\|_{\mathrm{TV}} \geqslant \|\hat{\mu}\|_{\mathrm{TV}} + \langle\Phi^\star\hat{p}, \nu\rangle_{\mathcal{M}}$ for all $\nu \in \mathcal{M}(G)$. Replacing ν by $\nu - \hat{\mu}$, this is the inequality of Lemma 1.1, that is $\Phi^\star\hat{p} \in \partial\|\hat{\mu}\|_{\mathrm{TV}}$.

Conversely, let $\hat{\mu} \in \mathcal{M}(G)$ with $\Phi^\star\hat{p} \in \partial\|\hat{\mu}\|_{\mathrm{TV}}$, and let $\nu \in \mathcal{M}(G)$. The same expansion, with $\tau = 1$ and $\nu - \hat{\mu}$ in place of ν, gives

$$\tfrac{1}{2}\|\Phi\nu - y\|^2 - \tfrac{1}{2}\|\Phi\hat{\mu} - y\|^2 = -\lambda\langle\hat{p}, \Phi(\nu - \hat{\mu})\rangle + \tfrac{1}{2}\|\Phi(\nu - \hat{\mu})\|^2\,.$$

The subgradient inequality and (1.1) give $\lambda\|\nu\|_{\mathrm{TV}} - \lambda\|\hat{\mu}\|_{\mathrm{TV}} \geqslant \lambda\langle\Phi^\star\hat{p}, \nu - \hat{\mu}\rangle_{\mathcal{M}} = \lambda\langle\hat{p}, \Phi(\nu - \hat{\mu})\rangle$. Adding,

$$J(\nu) - J(\hat{\mu}) \geqslant \tfrac{1}{2}\|\Phi(\nu - \hat{\mu})\|^2 \geqslant 0. \tag{1.4}$$

So $\hat{\mu}$ is a solution.

(3) Let $\hat{\mu}$ be a solution. By (2) and Lemma 1.1, $\|\Phi^\star\hat{p}\|_\infty \leqslant 1$, so $\hat{p} \in C$, and $\langle\Phi^\star\hat{p}, \hat{\mu}\rangle_{\mathcal{M}} = \|\hat{\mu}\|_{\mathrm{TV}}$. Let $z \stackrel{\text{def.}}{=} y/\lambda$, so that $\lambda(z - \hat{p}) = \Phi\hat{\mu}$. For $q \in C$, by (1.1),

$$\lambda\langle z - \hat{p}, q - \hat{p}\rangle = \langle\Phi\hat{\mu}, q - \hat{p}\rangle = \langle\Phi^\star q, \hat{\mu}\rangle_{\mathcal{M}} - \|\hat{\mu}\|_{\mathrm{TV}} \leqslant \|\Phi^\star q\|_\infty\|\hat{\mu}\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}} \leqslant 0.$$

Hence, for all $q \in C$,

$$\|z - q\|^2 = \|z - \hat{p}\|^2 - 2\langle z - \hat{p}, q - \hat{p}\rangle + \|q - \hat{p}\|^2 \geqslant \|z - \hat{p}\|^2 + \|q - \hat{p}\|^2\,.$$

So $\hat{p}$ is the unique point of C at minimal distance from z, that is the projection of y/λ onto C. This point depends only on y, λ and C, not on the solution $\hat{\mu}$.

(4) Weak duality. Let $\mu \in \mathcal{M}(G)$ and $p \in \mathbb{H}$. Expanding the square and using (1.1), one checks the identity

$$J(\mu) - D(p) = \tfrac{1}{2}\,\|\Phi\mu - y + \lambda p\|^2 + \lambda(\|\mu\|_{\mathrm{TV}} - \langle \Phi^\star p,\, \mu\rangle_{\mathcal{M}}). \tag{1.5}$$

If $p \in C$, both terms are nonnegative by (1.1), so $J(\mu) \geqslant D(p)$.

Strong duality. Let $\hat{\mu}$ be a solution. In (1.5) for $(\hat{\mu}, \hat{p})$, the first term is zero by definition of $\hat{p}$, and the second one is zero by (3). So $D(\hat{p}) = J(\hat{\mu})$, and with weak duality, $D(p) \leqslant J(\hat{\mu}) = D(\hat{p})$ for all $p \in C$. Hence $\min \mathcal{P}_\lambda(y) = J(\hat{\mu}) = \max_C D$, attained at $\hat{p} \in C$.

Uniqueness of the maximiser. Let $p \in C$ with $D(p) = J(\hat{\mu})$. Then (1.5) for $(\hat{\mu}, p)$ gives $\|\Phi\hat{\mu} - y + \lambda p\| = 0$, so $p = (y - \Phi\hat{\mu})/\lambda = \hat{p}$. □

Remark 1.5 (Consequences). Let $\hat{\eta} \overset{\text{def.}}{=} \Phi^\star\hat{p}$. (a) By (3), $\Phi\hat{\mu} = y - \lambda\hat{p}$ is the same for all solutions, and so is $\|\hat{\mu}\|_{\mathrm{TV}} = (\min \mathcal{P}_\lambda(y) - \frac{1}{2}\,\|\lambda\hat{p}\|_{\mathbb{H}}^2)/\lambda$. (b) By (2) and Lemma 1.1, every solution satisfies $\mathrm{supp}(\hat{\mu}) \subset \{t \in G\,;\, |\hat{\eta}(t)| = 1\}$. If the atoms φ_t indexed by this set are linearly independent, then (a) gives a unique solution. (c) By (2) with $\hat{\mu} = 0$, the measure 0 is a solution if and only if $\|\Phi^\star y\|_\infty \leqslant \lambda$. (d) The projection onto a closed convex set is 1-Lipschitz. So (3) compares the dual solutions for two data y and y'. This is used in Corollary 2.1 and Lemma 4.2.

The last proposition is the noisy version of Proposition 1.1. It holds for any certificate p of the target, not only for the dual solution. Here μ^0 is any measure on G.

Proposition 1.3 (Basic inequality)**.** *Let $y = \Phi\mu^0 + w$ and let $p \in \mathbb{H}$ with $\eta = \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$. Let $\hat{\mu}$ be a solution of $\mathcal{P}_\lambda(y)$ and $\rho \overset{\text{def.}}{=} \|w\|_{\mathbb{H}} + \lambda\,\|p\|_{\mathbb{H}}$. Then*

$$\left\|\Phi(\hat{\mu} - \mu^0)\right\|_{\mathbb{H}} \leqslant 2\rho \quad \textit{and} \quad \|\hat{\mu}\|_{\mathrm{TV}} - \langle \eta,\, \hat{\mu}\rangle_{\mathcal{M}} \leqslant \frac{\rho^2}{2\lambda}.$$

If in addition $|\eta| \leqslant 1 - \varepsilon$ on $G \setminus \mathrm{supp}(\mu^0)$ for some $\varepsilon \in (0, 1]$, then $|\hat{\mu}|\,(G \setminus \mathrm{supp}(\mu^0)) \leqslant \rho^2/(2\lambda\varepsilon)$. If $p \neq 0$, $\|w\|_{\mathbb{H}} \leqslant \delta$ with $\delta > 0$, and $\lambda = \delta/\,\|p\|_{\mathbb{H}}$, then $\left\|\Phi(\hat{\mu} - \mu^0)\right\|_{\mathbb{H}} \leqslant 4\delta$ and, under the same assumption on η, $|\hat{\mu}|\,(G \setminus \mathrm{supp}(\mu^0)) \leqslant 2\delta\,\|p\|_{\mathbb{H}}/\varepsilon$.

Proof. Write $h \overset{\text{def.}}{=} \Phi(\hat{\mu} - \mu^0)$ and $D \overset{\text{def.}}{=} \|\hat{\mu}\|_{\mathrm{TV}} - \langle \eta,\, \hat{\mu}\rangle_{\mathcal{M}}$. Since $\|\eta\|_\infty \leqslant 1$, (1.2) gives $D \geqslant 0$.

Compare with μ^0. Since $\Phi\hat{\mu} - y = h - w$ and $\Phi\mu^0 - y = -w$, the inequality $J(\hat{\mu}) \leqslant J(\mu^0)$, where J is the objective of $\mathcal{P}_\lambda(y)$, reads

$$\tfrac{1}{2}\,\|h - w\|^2 + \lambda\|\hat{\mu}\|_{\mathrm{TV}} \leqslant \tfrac{1}{2}\,\|w\|^2 + \lambda\|\mu^0\|_{\mathrm{TV}}, \quad \text{that is} \quad \tfrac{1}{2}\,\|h\|^2 - \langle h,\, w\rangle + \lambda(\|\hat{\mu}\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}}) \leqslant 0.$$

Use the certificate. Since $\|\mu^0\|_{\mathrm{TV}} = \langle \eta,\, \mu^0\rangle_{\mathcal{M}}$ and, by (1.1), $\langle \eta,\, \hat{\mu} - \mu^0\rangle_{\mathcal{M}} = \langle p,\, h\rangle$,

$$\|\hat{\mu}\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}} = D + \langle \eta,\, \hat{\mu} - \mu^0\rangle_{\mathcal{M}} = D + \langle p,\, h\rangle.$$

Hence, with the Cauchy-Schwarz inequality,

$$\tfrac{1}{2}\,\|h\|^2 + \lambda D \leqslant \langle h,\, w - \lambda p\rangle \leqslant \|h\|\,(\,\|w\| + \lambda\,\|p\|\,) = \rho\,\|h\|\,. \tag{1.6}$$

First bound. Since $\lambda D \geqslant 0$, (1.6) gives $\frac{1}{2}\,\|h\|^2 \leqslant \rho\,\|h\|$. If $h \neq 0$, divide by $\|h\|$. In all cases $\|h\| \leqslant 2\rho$.

Second bound. By (1.6), $\lambda D \leqslant \rho\,\|h\| - \frac{1}{2}\,\|h\|^2 \leqslant \max_{u \geqslant 0}\,(\rho u - \frac{1}{2}u^2) = \frac{1}{2}\rho^2$, the maximum being attained at $u = \rho$. So $D \leqslant \rho^2/(2\lambda)$.

Mass off the support. By (1.2), since $1 - |\eta| \geqslant 0$ on G and $1 - |\eta| \geqslant \varepsilon$ on $G \setminus \mathrm{supp}(\mu^0)$,

$$D \geqslant \int_G (1 - |\eta|)\,\mathrm{d}\,|\hat{\mu}| \geqslant \int_{G\setminus\mathrm{supp}(\mu^0)} (1 - |\eta|)\,\mathrm{d}\,|\hat{\mu}| \geqslant \varepsilon\;|\hat{\mu}|\,(G \setminus \mathrm{supp}(\mu^0)).$$

So $|\hat{\mu}|(G \setminus \operatorname{supp}(\mu^0)) \leqslant D/\varepsilon \leqslant \rho^2/(2\lambda\varepsilon)$.

Choice of λ. Let $p \neq 0$, $\delta > 0$ and $\lambda = \delta/\|p\|$, which is positive. Then $\lambda\|p\| = \delta$ and $\rho \leqslant 2\delta$. The first bound gives $\|h\| \leqslant 4\delta$, and

$$\frac{\rho^2}{2\lambda\varepsilon} \leqslant \frac{4\delta^2}{2\varepsilon\,\delta/\|p\|} = \frac{2\delta\|p\|}{\varepsilon}. \qquad \square$$

Remark 1.6 (Hypotheses). (a) The proof uses only $J(\hat{\mu}) \leqslant J(\mu^0)$. So the bounds hold for every $\hat{\mu}$ with this property, not only for the solutions of $\mathcal{P}_\lambda(y)$. (b) If $\mu^0 \neq 0$, then $|\eta| = 1$ at the points of its support, so $\eta \neq 0$ and $p \neq 0$. If $p = 0$, then $\eta = 0$, and $\langle \eta, \mu^0\rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}$ forces $\mu^0 = 0$. In that case the first bounds hold for every $\lambda > 0$ with $\rho = \|w\|_{\mathbb{H}}$, and the choice $\lambda = \delta/\|p\|_{\mathbb{H}}$ is not defined. (c) The constant ε is a margin: the certificate must stay away from ± 1 outside the support. On a fine grid this margin is small, see Lemma 2.1.

Remark 1.7 (From the grid to measures). In Lectures 2 to 4, the proofs of Propositions 1.1, 1.2 and 1.3 are copied with $\langle \eta, \mu\rangle_{\mathcal{M}}$ read as $\int \eta \,\mathrm{d}\mu$ and σ_μ read as the density of μ with respect to $|\mu|$. Three points change. Existence uses weak-$\star$ compactness (Theorem 2.2) instead of compactness in $\mathbb{R}^N$. The function $\Phi^\star p$ must be continuous for Lemma 1.1 to make sense. The phrase "for every $t \in \operatorname{supp}(\nu)$" becomes "$|\nu|$-almost everywhere".

1.5 Classical results that we do not prove

This section lists classical results of compressed sensing on a fixed grid. They are all in the book of Foucart and Rauhut, *A Mathematical Introduction to Compressive Sensing*, in matrix notation; we write them for measures on G. For $S \subset G$ let $\Phi_S : b \in \mathbb{R}^S \mapsto \sum_{t\in S} b_t\varphi_t$. When Φ_S is injective, let $\Phi_S^+ \overset{\text{def.}}{=} (\Phi_S^\star\Phi_S)^{-1}\Phi_S^\star$, and for $u \in \mathbb{R}^S$ let $\|u\|_1 = \sum_{t\in S}|u_t|$.

Orthogonal matching pursuit. This greedy algorithm starts from $S^0 = \emptyset$ and $r^0 = y$. At step $k \geqslant 1$, it picks $t^k \in G$ with $\left|\Phi^\star r^{k-1}(t^k)\right| = \left\|\Phi^\star r^{k-1}\right\|_\infty$ and sets $S^k = S^{k-1} \cup \{t^k\}$. Then μ^k minimises $\|\Phi\mu - y\|_{\mathbb{H}}$ among the measures supported in S^k, and $r^k = y - \Phi\mu^k$. The algorithm stops when $r^k = 0$.

Theorem 1.5 (Exact recovery condition). *Let $S \subset G$ with $|S| = s \geqslant 1$ and Φ_S injective. Assume the exact recovery condition*

$$\max_{t\in G\setminus S} \left\|\Phi_S^+\varphi_t\right\|_1 < 1.$$

Then every nonzero μ with $\operatorname{supp}(\mu) \subset S$ is recovered from $y = \Phi\mu$ by orthogonal matching pursuit after at most s steps. Every μ with $\operatorname{supp}(\mu) = S$ is the unique solution of $\mathcal{P}_0(\Phi\mu)$.

For basis pursuit, this is Proposition 1.1. Given signs $\sigma \in \{-1, 1\}^S$, let $p \overset{\text{def.}}{=} \Phi_S(\Phi_S^\star\Phi_S)^{-1}\sigma$. Then $\Phi^\star p = \sigma$ on S and, for $t \notin S$, $|\Phi^\star p(t)| = \left|\langle \Phi_S^+\varphi_t, \sigma\rangle_{\mathbb{R}^S}\right| \leqslant \left\|\Phi_S^+\varphi_t\right\|_1 < 1$. With $\Phi_{x_0} = \Phi_S$, this p is the Fuchs pre-certificate of Definition 2.1.

Coherence. Assume $\|\varphi_t\|_{\mathbb{H}} = 1$ for all $t \in G$. The coherence of Φ is $M(\Phi) \overset{\text{def.}}{=} \max_{t\neq t'} |\langle \varphi_t, \varphi_{t'}\rangle_{\mathbb{H}}|$.

Theorem 1.6 (Welch bound and coherence guarantee). *(a) If $N > m$, then $M(\Phi) \geqslant \sqrt{(N-m)/(m(N-1))}$. (b) Let $s \geqslant 1$ with $(2s-1)M(\Phi) < 1$. Then the exact recovery condition holds for every $S \subset G$ with $|S| \leqslant s$. So every nonzero $\mu \in \Sigma_s$ is recovered by orthogonal matching pursuit after at most s steps, and is the unique solution of $\mathcal{P}_0(\Phi\mu)$.*

If $N \geqslant 2m$, then $(N-m)/(N-1) \geqslant 1/2$, and (a) gives $M(\Phi) \geqslant 1/\sqrt{2m}$. The condition of (b) then requires $2m > (2s-1)^2$. Coherence guarantees need a number of measurements of order s^2.

Stable and robust null space properties. For $\mu \in \mathcal{M}(G)$, let $\sigma_s(\mu) \stackrel{\text{def.}}{=} \min_{\nu \in \Sigma_s} \|\mu - \nu\|_{\mathrm{TV}}$, the error of the best s-sparse approximation. Let $0 < \varrho < 1$ and $\tau > 0$. The operator Φ has the robust null space property of order s with constants (ϱ, τ) if $|\nu|(S) \leqslant \varrho\, |\nu|(G \setminus S) + \tau \|\Phi\nu\|_{\mathbb{H}}$ for all $\nu \in \mathcal{M}(G)$ and all $S \subset G$ with $|S| \leqslant s$. It has the stable null space property of order s with constant ϱ if $|\nu|(S) \leqslant \varrho\, |\nu|(G \setminus S)$ for all $\nu \in \operatorname{Ker}\Phi$ and all $S \subset G$ with $|S| \leqslant s$.

Theorem 1.7 (Stable and robust recovery)**.** *(a) The null space property of order s holds if and only if the stable null space property of order s holds for some $\varrho \in (0,1)$. (b) Under the stable null space property with constant ϱ, every solution $\mu^\sharp$ of $\mathcal{P}_0(\Phi\mu)$ satisfies $\|\mu - \mu^\sharp\|_{\mathrm{TV}} \leqslant \frac{2(1+\varrho)}{1-\varrho}\, \sigma_s(\mu)$. (c) Under the robust null space property with constants (ϱ, τ), let $y = \Phi\mu + w$ with $\|w\|_{\mathbb{H}} \leqslant \delta$. Every solution $\mu^\sharp$ of $\min \|\nu\|_{\mathrm{TV}}$ subject to $\|\Phi\nu - y\|_{\mathbb{H}} \leqslant \delta$ satisfies $\|\mu - \mu^\sharp\|_{\mathrm{TV}} \leqslant \frac{2(1+\varrho)}{1-\varrho}\, \sigma_s(\mu) + \frac{4\tau}{1-\varrho}\, \delta$.*

Restricted isometry property. For $\mu = \sum_j a_j \delta_{t_j}$, let $\|a\|_2 = (\sum_j a_j^2)^{1/2}$. The restricted isometry constant δ_s of Φ is the smallest $\delta \geqslant 0$ such that $(1-\delta)\|a\|_2^2 \leqslant \|\Phi\mu\|_{\mathbb{H}}^2 \leqslant (1+\delta)\|a\|_2^2$ for all $\mu \in \Sigma_s$ [5]. In the next theorem the atoms are random and are not normalised.

Theorem 1.8 (Restricted isometry and random measurements)**.** *(a) If $\delta_{2s} < 1/3$, then Φ has the null space property of order s. (b) Let $\mathbb{H} = \mathbb{R}^m$ and let the coordinates of the φ_t, $t \in G$, be independent Gaussian variables with mean 0 and variance $1/m$. There are universal constants $C, c > 0$ such that, if $m \geqslant Cs\ln(eN/s)$, then $\delta_{2s} < 1/3$ with probability at least $1 - 2e^{-cm}$.*

With (a) and (b), m of order $s\ln(eN/s)$ random measurements are enough for the exact recovery of all s-sparse measures by basis pursuit, with high probability. The first results of this type are due to Candès and Tao [5], to Candès, Romberg, and Tao [6], and to Donoho [7].

Remark 1.8 (Why these results do not help on a fine grid). Let the grid G have step h in $\mathbb{T}$, and let φ_t be defined for all $t \in \mathbb{T}$, with $\|\varphi_t\|_{\mathbb{H}} = 1$ and $\|\varphi_t - \varphi_{t'}\|_{\mathbb{H}} \leqslant L\,|t - t'|$. For two neighbouring points $t, t+h$ of the grid,

$$\langle \varphi_t,\, \varphi_{t+h} \rangle_{\mathbb{H}} = 1 - \tfrac{1}{2}\,\|\varphi_{t+h} - \varphi_t\|_{\mathbb{H}}^2 \geqslant 1 - \tfrac{1}{2}L^2h^2.$$

So $M(\Phi) \to 1$ when $h \to 0$. If $L^2h^2 \leqslant 4/3$, then $M(\Phi) \geqslant 1/3$, and the condition of Theorem 1.6 (b) fails for every $s \geqslant 2$. In the same way, the measure $\mu = \delta_t - \delta_{t+h}$ has $\|a\|_2^2 = 2$ and $\|\Phi\mu\|_{\mathbb{H}}^2 = \|\varphi_t - \varphi_{t+h}\|_{\mathbb{H}}^2 \leqslant L^2h^2$, so $\delta_{2s} \geqslant \delta_2 \geqslant 1 - \frac{1}{2}L^2h^2$ for $s \geqslant 1$. If $L^2h^2 \leqslant 4/3$, then $\delta_{2s} \geqslant 1/3$ and Theorem 1.8 (a) does not apply. For the weighted Fourier measurements of Example 2.1, with the real scalar product $\Re\langle \cdot, \cdot \rangle_{\mathbb{H}}$, one has $\|\varphi_t\|_{\mathbb{H}} = 1$ and $\|\partial_t\varphi_t\|_{\mathbb{H}}^2 = \sum_\omega \Lambda(\omega) 4\pi^2\omega^2 \leqslant 4\pi^2 f_c^2$, so $L \leqslant 2\pi f_c$. The coherence and restricted isometry guarantees of this section are lost as soon as $h \leqslant 1/(\sqrt{3}\,\pi f_c)$, a fraction of the Rayleigh width $1/f_c$. This is the regime of interest. The margin ε of a certificate also degrades on a fine grid (Lemma 2.1), but certificates keep a meaning when the grid disappears.

Exercises

Exercise 1.1 (A certificate on three grid points)**.** Let $G = \{t_1, t_2, t_3\}$, $\mathbb{H} = \mathbb{R}^2$ with canonical basis (e_1, e_2), $\beta > 0$, and $\varphi_{t_1} = e_1$, $\varphi_{t_2} = \beta(e_1 + e_2)$, $\varphi_{t_3} = e_2$.

1. Compute $\operatorname{Ker}\Phi$. Show that Φ has the null space property of order one if and only if $\beta > 1/2$.

2. Let $\mu^0 = a_1\delta_{t_1} + a_3\delta_{t_3}$ with $a_1, a_3 \neq 0$. Show that there is exactly one $p \in \mathbb{H}$ with $(\Phi^\star p)(t_1) = \operatorname{sign}(a_1)$ and $(\Phi^\star p)(t_3) = \operatorname{sign}(a_3)$, and compute $(\Phi^\star p)(t_2)$.

3. Assume $\operatorname{sign}(a_1) = -\operatorname{sign}(a_3)$. Show that μ^0 is the unique solution of $\mathcal{P}_0(\Phi\mu^0)$ for every $\beta > 0$.

4. Assume $a_1, a_3 > 0$ and $\beta < 1/2$. Show that μ^0 is the unique solution of $\mathcal{P}_0(\Phi\mu^0)$.

5. Assume $0 < a_1 \leqslant a_3$ and $\beta > 1/2$. Find ν with $\Phi\nu = \Phi\mu^0$ and $\|\nu\|_{\mathrm{TV}} < \|\mu^0\|_{\mathrm{TV}}$. What happens for $\beta = 1/2$?

6. Show that Φ never has the null space property of order two. Compare with questions 3 and 4.

Exercise 1.2 (The LASSO with orthonormal atoms)**.** Let $\lambda > 0$ and $y \in \mathbb{H}$. Assume that $(\varphi_t)_{t\in G}$ is orthonormal, and let $z_j \overset{\text{def.}}{=} \langle \varphi_{t_j}, y\rangle_{\mathbb{H}}$.

1. Show that $\frac{1}{2}\|\Phi\mu - y\|_{\mathbb{H}}^2 = \frac{1}{2}\sum_j (a_j - z_j)^2 + c$ for $\mu = \sum_j a_j\delta_{t_j}$, where c does not depend on μ.

2. Deduce that $\mathcal{P}_\lambda(y)$ has a unique solution, with amplitudes $\hat{a}_j = \operatorname{sign}(z_j)\max(|z_j| - \lambda, 0)$.

3. Compute $\hat{p}$ and $\Phi^\star\hat{p}$. Check items (2) and (3) of Proposition 1.2 directly.

4. Let $y = \Phi\mu^0 + w$ with $\mu^0 = \sum_{i=1}^s a_i\delta_{x_i}$ as in Proposition 1.1, and let $p = \sum_i \operatorname{sign}(a_i)\varphi_{x_i}$. Show that $\Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$, with the margin $\varepsilon = 1$ and $\|p\|_{\mathbb{H}} = \sqrt{s}$. Compute $|\hat{\mu}|(G \setminus \operatorname{supp}(\mu^0))$ in terms of w, and compare it with the bound of Proposition 1.3.

Exercise 1.3 (One spike)**.** Assume $\|\varphi_t\|_{\mathbb{H}} = 1$ for all $t \in G$. Let $x \in G$, $a \neq 0$ and $\mu^0 = a\delta_x$. Let $\kappa \overset{\text{def.}}{=} \max_{t\neq x} |\langle \varphi_t, \varphi_x\rangle_{\mathbb{H}}|$.

1. Let $p = \operatorname{sign}(a)\varphi_x$ and $\eta = \Phi^\star p$. Show that $\eta \in \partial\|\mu^0\|_{\mathrm{TV}}$.

2. Assume $\kappa < 1$. Show that μ^0 is the unique solution of $\mathcal{P}_0(\Phi\mu^0)$.

3. Assume $\kappa < 1$. Let $y = \Phi\mu^0 + w$ with $\|w\|_{\mathbb{H}} \leqslant \delta$, $\delta > 0$, and $\lambda = \delta$. Show that every solution $\hat{\mu}$ of $\mathcal{P}_\lambda(y)$ satisfies $\|\Phi(\hat{\mu} - \mu^0)\|_{\mathbb{H}} \leqslant 4\delta$ and $|\hat{\mu}|(G \setminus \{x\}) \leqslant 2\delta/(1-\kappa)$.

4. In the setting of Remark 1.8, show that $1 - \kappa \leqslant \frac{1}{2}L^2h^2$ when x has a neighbour at distance h in G. What does the bound of question 3 become when $h \to 0$?

Exercise 1.4 (The dual solution at low noise)**.** For $\lambda > 0$ and $y \in \mathbb{H}$, let $\hat{p}_\lambda(y)$ be the dual solution of Proposition 1.2.

1. Show that $\|\hat{p}_\lambda(y) - \hat{p}_\lambda(y')\|_{\mathbb{H}} \leqslant \|y - y'\|_{\mathbb{H}}/\lambda$ for $y, y' \in \mathbb{H}$. Hint: write the inequality of the proof of item (3) twice and add.

2. Let $y_0 = \Phi\mu^0$ and
$$P \overset{\text{def.}}{=} \left\{p \in \mathbb{H}\ ;\ \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}\right\}.$$
Show that $P = \{p \in C\ ;\ \langle p, y_0\rangle_{\mathbb{H}} = \|\mu^0\|_{\mathrm{TV}}\}$, and that P is closed and convex.

3. Assume $P \neq \emptyset$. Show that P has a unique element p_0 of minimal norm. Show that $\|\hat{p}_\lambda(y_0)\|_{\mathbb{H}} \leqslant \|p\|_{\mathbb{H}}$ for all $p \in P$ and all $\lambda > 0$. Hint: compare the dual objective at $\hat{p}_\lambda(y_0)$ and at p.

4. Show that $\hat{p}_\lambda(y_0) \to p_0$ when $\lambda \to 0$. Hint: let p^* be a limit point; show that $\langle y_0, q\rangle_{\mathbb{H}} \leqslant \langle y_0, p^*\rangle_{\mathbb{H}}$ for all $q \in C$.

5. Assume that $|\Phi^\star p_0| < 1$ on $G \setminus \operatorname{supp}(\mu^0)$. Show that there are $\alpha, \lambda_0 > 0$ such that, for $0 < \lambda \leqslant \lambda_0$ and $\|w\|_{\mathbb{H}} \leqslant \alpha\lambda$, every solution of $\mathcal{P}_\lambda(y_0 + w)$ has its support in $\operatorname{supp}(\mu^0)$. Compare with Corollary 2.1.

Chapter 2

From the grid to measures

The first two sections stay on the grid of Lecture 1. They show that the LASSO recovers the support at low noise, and that the condition behind this result degrades when the grid gets finer. The last three sections remove the grid. The unknown becomes a measure on a continuous set $\mathcal{X}$, and the proofs of Section 1.4 are repeated on measures.

2.1 Support recovery on the grid at low noise

We keep the grid G of Lecture 1. Let $\mu^0 = \sum_{i=1}^s a_i \delta_{x_i}$ with $x_i \in G$, $y_0 \overset{\text{def.}}{=} \Phi\mu^0$, and $\Phi_{x_0} : b \in \mathbb{R}^s \mapsto \sum_i b_i \varphi_{x_i}$. For $\alpha, \lambda_0 > 0$ the low-noise regime is

$$D_{\alpha,\lambda_0} \overset{\text{def.}}{=} \{(\lambda, w) \in \mathbb{R}_+ \times \mathbb{H} \,;\, 0 < \lambda \leqslant \lambda_0,\ \|w\|_{\mathbb{H}} \leqslant \alpha\lambda\}\,. \tag{2.1}$$

As in Lecture 1, $\mathbb{H}$ is a real Hilbert space of finite dimension. Here $s \geqslant 1$, the points x_i are pairwise distinct, $a_i \neq 0$, $S \overset{\text{def.}}{=} \operatorname{supp}(\mu^0) = \{x_1, \dots, x_s\}$ and $\operatorname{sign}(a) \overset{\text{def.}}{=} (\operatorname{sign}(a_i))_{i \leqslant s}$. The adjoint of Φ_{x_0} is $\Phi_{x_0}^\star p = \big((\Phi^\star p)(x_i)\big)_{i \leqslant s}$, since $\langle \Phi_{x_0} b,\, p\rangle_{\mathbb{H}} = \sum_i b_i \langle \varphi_{x_i},\, p\rangle_{\mathbb{H}}$. Proposition 1.3 bounds the mass that the LASSO puts outside S. The next result says when this mass is zero.

Definition 2.1 (Fuchs pre-certificate)**.** Assume that Φ_{x_0} is injective. The Fuchs pre-certificate is $\eta_F \overset{\text{def.}}{=} \Phi^\star p_F$ with $p_F \overset{\text{def.}}{=} \Phi_{x_0}(\Phi_{x_0}^\star \Phi_{x_0})^{-1} \operatorname{sign}(a)$. The vector p_F has minimal norm among the $p \in \mathbb{H}$ with $(\Phi^\star p)(x_i) = \operatorname{sign}(a_i)$ for all i.

We check the definition. If $\Phi_{x_0}^\star \Phi_{x_0} b = 0$, then $\|\Phi_{x_0} b\|_{\mathbb{H}}^2 = \langle b,\, \Phi_{x_0}^\star \Phi_{x_0} b\rangle_{\mathbb{R}^s} = 0$, so $b = 0$. Hence the $s \times s$ matrix $\Phi_{x_0}^\star \Phi_{x_0}$ is invertible and p_F is well defined. The constraints read $\Phi_{x_0}^\star p = \operatorname{sign}(a)$, and $\Phi_{x_0}^\star p_F = \operatorname{sign}(a)$. If p satisfies them, then $\Phi_{x_0}^\star(p - p_F) = 0$, so $p - p_F$ is orthogonal to $\operatorname{Im} \Phi_{x_0}$, which contains p_F. By Pythagoras, $\|p\|_{\mathbb{H}}^2 = \|p_F\|_{\mathbb{H}}^2 + \|p - p_F\|_{\mathbb{H}}^2$. So p_F is the unique element of minimal norm.

Corollary 2.1 (Exact support recovery on a grid)**.** *Assume that Φ_{x_0} is injective and that $|\eta_F(t)| < 1$ for all $t \in G \setminus \operatorname{supp}(\mu^0)$. Then there are $\alpha, \lambda_0 > 0$ such that, for all $(\lambda, w) \in D_{\alpha,\lambda_0}$, the solution $\hat{\mu}_\lambda$ of $\mathcal{P}_\lambda(y_0 + w)$ is unique, $\operatorname{supp}(\hat{\mu}_\lambda) = \operatorname{supp}(\mu^0)$, and $\hat{a}_{\lambda,i} - a_i = \mathcal{O}(\lambda) + \mathcal{O}(\|w\|_{\mathbb{H}})$.*

On the support, the optimality condition of Proposition 1.2 is a linear system. The proof solves it, and then checks the optimality condition off the support.

Proof. All norms and scalar products are those of $\mathbb{H}$, and $\|\cdot\|_\infty$ is the max norm on $\mathbb{R}^s$. Let $\Gamma \overset{\text{def.}}{=} \Phi_{x_0}^\star \Phi_{x_0}$, $\sigma \overset{\text{def.}}{=} \operatorname{sign}(a)$ and $\Pi \overset{\text{def.}}{=} \Phi_{x_0} \Gamma^{-1} \Phi_{x_0}^\star$, so that $p_F = \Phi_{x_0} \Gamma^{-1} \sigma$. The operator Π is the orthogonal projection onto $\operatorname{Im} \Phi_{x_0}$: it is self-adjoint, $\Pi^2 = \Pi$ and $\Pi \Phi_{x_0} = \Phi_{x_0}$. Hence $\Phi_{x_0}^\star(\mathrm{Id} - \Pi) = 0$ and $\|(\mathrm{Id} - \Pi) w\| \leqslant \|w\|$ for $w \in \mathbb{H}$.

Constants. Let $\varepsilon \overset{\text{def.}}{=} 1 - \max_{t \in G \setminus S} |\eta_F(t)|$, with $\varepsilon \overset{\text{def.}}{=} 1$ if $G = S$. Since G is finite, $\varepsilon > 0$. Let $B \overset{\text{def.}}{=} \max_{t \in G} \|\varphi_t\|$, which is positive since $\varphi_{x_1} \neq 0$. Let $a_{\min} \overset{\text{def.}}{=} \min_i |a_i|$, $c \overset{\text{def.}}{=} \|\Gamma^{-1}\sigma\|_\infty > 0$, and let β be the norm of the linear map $\Gamma^{-1}\Phi_{x_0}^\star$ from $\mathbb{H}$ to $(\mathbb{R}^s, \|\cdot\|_\infty)$. Set

$$\alpha \overset{\text{def.}}{=} \frac{\varepsilon}{2B}, \qquad \lambda_0 \overset{\text{def.}}{=} \frac{a_{\min}}{2(\alpha\beta + c)}.$$

Fix $(\lambda, w) \in D_{\alpha,\lambda_0}$ and let $y \overset{\text{def.}}{=} y_0 + w$.

Candidate. Set $\hat{a} \overset{\text{def.}}{=} a + \Gamma^{-1}(\Phi_{x_0}^\star w - \lambda\sigma)$ and $\hat{\mu} \overset{\text{def.}}{=} \sum_i \hat{a}_i \delta_{x_i}$. For all i,

$$|\hat{a}_i - a_i| \leqslant \beta \|w\| + c\lambda \leqslant (\alpha\beta + c)\lambda_0 = \tfrac{1}{2} a_{\min} < |a_i|.$$

So $\hat{a}_i \neq 0$ and $\operatorname{sign}(\hat{a}_i) = \sigma_i$ for all i. In particular $\operatorname{supp}(\hat{\mu}) = S$.

Residual. Since $y_0 = \Phi_{x_0} a$, we get $\Phi\hat{\mu} = \Phi_{x_0}\hat{a} = y_0 + \Pi w - \lambda p_F$. Hence

$$\hat{p} \overset{\text{def.}}{=} \frac{y - \Phi\hat{\mu}}{\lambda} = p_F + \frac{(\mathrm{Id} - \Pi)w}{\lambda}. \tag{2.2}$$

Certificate. On the support, $\Phi_{x_0}^\star \hat{p} = \Phi_{x_0}^\star p_F = \sigma$, that is $(\Phi^\star \hat{p})(x_i) = \operatorname{sign}(\hat{a}_i)$ for all i. For $t \in G \setminus S$, by the Cauchy-Schwarz inequality,

$$|(\Phi^\star \hat{p})(t)| \leqslant |\eta_F(t)| + \frac{|\langle \varphi_t, (\mathrm{Id} - \Pi)w \rangle|}{\lambda} \leqslant 1 - \varepsilon + \frac{B\|w\|}{\lambda} \leqslant 1 - \varepsilon + B\alpha = 1 - \tfrac{1}{2}\varepsilon < 1.$$

So $\|\Phi^\star \hat{p}\|_\infty \leqslant 1$. By Lemma 1.1, $\Phi^\star \hat{p} \in \partial \|\hat{\mu}\|_{\mathrm{TV}}$, and by Proposition 1.2 (2), $\hat{\mu}$ is a solution of $\mathcal{P}_\lambda(y)$.

Uniqueness. Let ν be a solution of $\mathcal{P}_\lambda(y)$. By Proposition 1.2 (3), $(y - \Phi\nu)/\lambda = \hat{p}$, so $\Phi\nu = \Phi\hat{\mu}$. By Remark 1.5 (b) and the previous step, $\operatorname{supp}(\nu) \subset \{t \in G\,;\, |(\Phi^\star \hat{p})(t)| = 1\} = S$. So $\nu = \sum_i b_i \delta_{x_i}$ with $\Phi_{x_0} b = \Phi_{x_0}\hat{a}$, and injectivity gives $b = \hat{a}$, that is $\nu = \hat{\mu}$.

Error. We have $\hat{a} - a = \Gamma^{-1}\Phi_{x_0}^\star w - \lambda\Gamma^{-1}\sigma$, so $\|\hat{a} - a\|_\infty \leqslant \beta\|w\| + c\lambda$. □

Remark 2.1 (The dual point of view). Let $\hat{p}_\lambda(y)$ be the dual solution of Proposition 1.2 and $C = \{p \in \mathbb{H}\,;\, \|\Phi^\star p\|_\infty \leqslant 1\}$. (a) Let $P \overset{\text{def.}}{=} \{p \in \mathbb{H}\,;\, \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}\}$, as in Exercise 1.4. For $p \in C$, (1.1) gives $\langle p, y_0 \rangle_{\mathbb{H}} = \langle \Phi^\star p, \mu^0 \rangle_{\mathcal{M}} \leqslant \|\mu^0\|_{\mathrm{TV}}$, with equality if and only if $p \in P$ (Lemma 1.1). Under the assumption of Corollary 2.1, $p_F \in P$. So P is nonempty and it is the set of maximisers of $\langle \cdot, y_0 \rangle_{\mathbb{H}}$ on C. Every $p \in P$ satisfies $\Phi_{x_0}^\star p = \operatorname{sign}(a)$, so $\|p\|_{\mathbb{H}} \geqslant \|p_F\|_{\mathbb{H}}$. Hence p_F is the dual solution of minimal norm, that is $p_F = p_0$ in Exercise 1.4. (b) With $w = 0$, the proof gives $\hat{p}_\lambda(y_0) = p_F$ for all λ with $c\lambda < a_{\min}$. So $\hat{p}_\lambda(y_0) \to p_F$ when $\lambda \to 0$, and the limit is reached for λ small. (c) By (2.2), $\hat{p}_\lambda(y_0 + w) - \hat{p}_\lambda(y_0) = (\mathrm{Id} - \Pi)w/\lambda$, of norm at most $\|w\|_{\mathbb{H}}/\lambda \leqslant \alpha$. This is the bound that the Lipschitz property of the projection gives in general (Remark 1.5 (d)). On measures there is no closed form for the solution, and Lecture 4 follows this general route: non-expansiveness of $\hat{p}_\lambda$, then the limit $\hat{p}_\lambda(y_0) \to p_0$.

2.2 Why the grid fails

Corollary 2.1 and Proposition 1.3 need a margin: the certificate must stay away from ± 1 on $G \setminus \operatorname{supp}(\mu^0)$. When grid points come close to a spike, this margin shrinks. In this section $G \subset \mathbb{T}$ and $t \mapsto \varphi_t$ is defined on all of $\mathbb{T}$, as in Remark 1.8. Then $\Phi^\star p$ is a function on $\mathbb{T}$. If $t \mapsto \varphi_t$ is $\mathscr{C}^2$ from $\mathbb{T}$ to $\mathbb{H}$, then $\Phi^\star p \in \mathscr{C}^2(\mathbb{T})$ and $(\Phi^\star p)'' = \langle \partial_t^2 \varphi_t, p \rangle_{\mathbb{H}}$ (Remark 2.4).

Lemma 2.1 (Margin on a grid of step h). *Let $\eta \in \mathscr{C}^2(\mathbb{T})$, $x \in \mathbb{T}$ with $\eta(x) = \pm 1$, and $h > 0$.*

1. *If $\|\eta\|_\infty \leqslant 1$, then $\eta'(x) = 0$ and $1 - |\eta(x \pm h)| \leqslant \frac{1}{2}\|\eta''\|_\infty h^2$.*

2. *If only* $|\eta(x-h)| \leqslant 1$ *and* $|\eta(x+h)| \leqslant 1$*, for instance if* $|\eta| \leqslant 1$ *on a grid of step* h *that contains* x*, then* $|\eta'(x)| \leqslant \frac{1}{2}\|\eta''\|_\infty h$ *and*

$$1 - |\eta(x \pm h)| \leqslant |\eta'(x)|\, h + \tfrac{1}{2}\|\eta''\|_\infty h^2 \leqslant \|\eta''\|_\infty h^2.$$

Proof. Let $\varsigma \overset{\text{def.}}{=} \eta(x) \in \{-1, 1\}$ and $\psi \overset{\text{def.}}{=} \varsigma\eta$. Then $\psi(x) = 1$, $|\psi| = |\eta|$, $|\psi'(x)| = |\eta'(x)|$ and $\|\psi''\|_\infty = \|\eta''\|_\infty$. We see ψ as a 1-periodic function of class $\mathscr{C}^2$ on $\mathbb{R}$. The Taylor formula with Lagrange remainder gives, for all $u \in \mathbb{R}$,

$$|\psi(x+u) - 1 - \psi'(x)u| \leqslant \tfrac{1}{2}\|\eta''\|_\infty u^2. \tag{2.3}$$

(1) Since $\|\eta\|_\infty \leqslant 1$, $\psi \leqslant 1 = \psi(x)$ on $\mathbb{R}$. So x is a maximum of ψ on the open set $\mathbb{R}$, and $\psi'(x) = 0$. Then (2.3) with $u = \pm h$ gives $|\eta(x \pm h)| \geqslant \psi(x \pm h) \geqslant 1 - \frac{1}{2}\|\eta''\|_\infty h^2$.

(2) Let $\tau \in \{-1, 1\}$ with $\tau\psi'(x) = |\psi'(x)|$. By (2.3) with $u = \tau h$, and since $\psi(x + \tau h) \leqslant |\eta(x + \tau h)| \leqslant 1$,

$$1 \geqslant \psi(x + \tau h) \geqslant 1 + |\psi'(x)|\, h - \tfrac{1}{2}\|\eta''\|_\infty h^2.$$

Hence $|\eta'(x)| \leqslant \frac{1}{2}\|\eta''\|_\infty h$. By (2.3) with $u = \pm h$, $|\eta(x \pm h)| \geqslant \psi(x \pm h) \geqslant 1 - |\eta'(x)|\, h - \frac{1}{2}\|\eta''\|_\infty h^2$. The bound on $|\eta'(x)|$ gives the last inequality. □

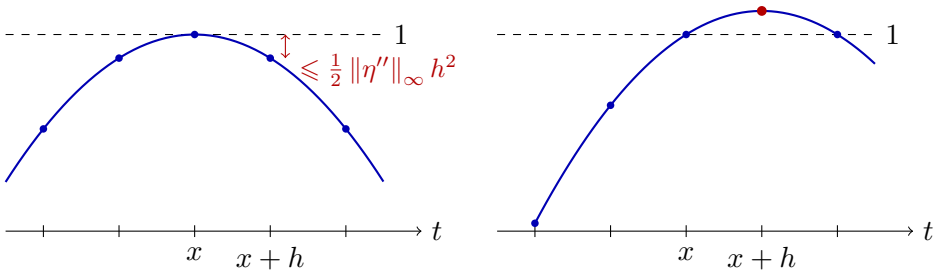


Figure 2.1: Lemma 2.1 with $\|\eta''\|_\infty = 6$ and $h = 0.2$. Left: $\|\eta\|_\infty \leqslant 1$ and $\eta(x) = 1$. Then $\eta'(x) = 0$, and the margin at the grid point $x + h$ is at most $\frac{1}{2}\|\eta''\|_\infty h^2$. Right: $\eta(x) = 1$ and $\eta'(x) = 1.2 > \frac{1}{2}\|\eta''\|_\infty h$. Then η exceeds 1 at the grid point $x + h$.

Remark 2.2 (Consequences on a fine grid). Let $t \mapsto \varphi_t$ be $\mathscr{C}^2$ from $\mathbb{T}$ to $\mathbb{H}$. Let $G \subset \mathbb{T}$ be a grid of step h that contains $S = \operatorname{supp}(\mu^0)$, with $h < \min_{i \neq j} |x_i - x_j|$, so that $x_i \pm h \in G \setminus S$ for all i. (a) Let $\eta = \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$, where the subdifferential is taken on G, and let ε be its margin as in Proposition 1.3. Then $|\eta| \leqslant 1$ on G and $|\eta(x_1)| = 1$, so item 2 of Lemma 2.1 gives $\varepsilon \leqslant 1 - |\eta(x_1 + h)| \leqslant \|\eta''\|_\infty h^2$. For the weighted Fourier measurements of Example 2.1, $\|\partial_t^2 \varphi_t\|_{\mathbb{H}} \leqslant 4\pi^2 f_c^2$, so $\|\eta''\|_\infty \leqslant 4\pi^2 f_c^2 \|p\|_{\mathbb{H}}$, and the bound $2\delta \|p\|_{\mathbb{H}} / \varepsilon$ of Proposition 1.3 is at least $\delta/(2\pi^2 f_c^2 h^2)$, whatever the certificate. The guarantee degrades like h^{-2}. This does not say that the LASSO itself fails. (b) The vector p_F depends only on $\varphi_{x_1}, \dots, \varphi_{x_s}$ and on $\operatorname{sign}(a)$, not on the grid. So η_F is one function on $\mathbb{T}$ for all grids that contain S. Assume $\eta_F'(x_i) \neq 0$ for some i. If $h < 2\,|\eta_F'(x_i)| \,/\, \|\eta_F''\|_\infty$, item 2 of Lemma 2.1 shows that $|\eta_F| > 1$ at $x_i + h$ or at $x_i - h$, and Corollary 2.1 does not apply. For one spike and a kernel $K(x, x') = \kappa(x - x')$ with κ even and $\kappa(0) = 1$, one finds $\eta_F = \operatorname{sign}(a_1)\kappa(\cdot - x_1)$ and $\eta_F'(x_1) = 0$. For several spikes, η_F only interpolates the signs, and in general $\eta_F'(x_i) \neq 0$ (Exercise 2.2). (c) By item 1 of Lemma 2.1, a certificate with $|\eta| \leqslant 1$ on all of $\mathbb{T}$ satisfies $\eta'(x_i) = 0$ for all i. The vanishing derivative pre-certificate η_V of Definition 3.2 adds these constraints to those of η_F. (d) Coherence and restricted isometry constants also tend to their worst values when $h \to 0$ (Remark 1.8).

Remark 2.3 (The LASSO on thin grids, admitted). The following facts are shown on slides and admitted. The setting is the following: $\mathcal{X} = \mathbb{T}$, $\mathbb{H} = \mathrm{L}^2(\mathbb{T})$, $\Phi\mu = \varphi \star \mu$ with $\varphi \in \mathscr{C}^3(\mathbb{T})$ real, the operator Γ_{x_0} of Section 3.2 is injective, μ^0 satisfies the non-degenerate source condition of

Definition 3.3, a structural inequality between two constants of (μ^0, Φ) holds, and G_h is the dyadic grid of step $h = 2^{-J}$, which contains $\mathrm{supp}(\mu^0)$. Let $\hat{\mu}_\lambda^{(h)}$ be the solution of the LASSO on G_h with data $y_0 + w$, and $\hat{a}_i^{\pm} \overset{\text{def.}}{=} \hat{\mu}_\lambda^{(h)}(\{x_i \pm h\})$. By [8, 9], there is $h_0 > 0$ such that, for every $h \leqslant h_0$, there are $\alpha_0(h), \lambda_0(h) > 0$ with the following properties for all $(\lambda, w) \in D_{\alpha_0(h),\lambda_0(h)}$.

(a) $\mathrm{supp}(\hat{\mu}_\lambda^{(h)}) \subset \mathrm{supp}(\mu^0) \cup \{x_i \pm h \ ; \ i = 1, \dots, s\}$.

(b) $\left|\hat{a}_i^+\right| + \left|\hat{a}_i^-\right| = \mathcal{O}(\,|a_i|\,(\lambda + \|w\|_{\mathbb{H}})/h)$. When λ is of order h, this upper bound is of order $|a_i|$.

So, at low noise, the LASSO on a thin grid can split one spike of μ^0 between two neighbouring grid points [8]. Since η_F does not depend on the grid (Remark 2.2 (b)), it does not converge to η_V when $h \to 0$, unless $\eta_F = \eta_V$.

2.3 The off-the-grid model

From now on $\mathcal{X}$ is the torus $\mathbb{T}$ or a closed subset of $\mathbb{R}^d$, measures are real, and $\mathbb{H}$ is a separable Hilbert space. When $\mathbb{H}$ is complex we use the real scalar product $\Re\langle \cdot, \cdot \rangle_{\mathbb{H}}$.

With the real scalar product, a complex Hilbert space is a real Hilbert space with the same norm, and it is still separable. Below, $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ always denotes this real scalar product. In both cases $\mathcal{X}$ is a locally compact metric space with a countable basis of open sets.

Definition 2.2 (Forward operator, model, BLASSO)**.** Let $x \in \mathcal{X} \mapsto \varphi_x \in \mathbb{H}$ be continuous and bounded. The forward operator is $\Phi\mu \overset{\text{def.}}{=} \int_{\mathcal{X}} \varphi_x \,\mathrm{d}\mu(x)$ and $\Phi^\star p \overset{\text{def.}}{=} \langle \varphi_\cdot, p \rangle_{\mathbb{H}}$. The target is $\mu^0 = \sum_{i=1}^s a_i \delta_{x_i}$ with $a_i \neq 0$ and pairwise distinct $x_i \in \mathcal{X}$. We observe $y = \Phi\mu^0 + w$ with $w \in \mathbb{H}$. For $\lambda > 0$ the BLASSO is

$$\hat{\mu}_\lambda \in \operatorname*{argmin}_{\mu \in \mathcal{M}(\mathcal{X})} \tfrac{1}{2} \|\Phi\mu - y\|_{\mathbb{H}}^2 + \lambda \|\mu\|_{\mathrm{TV}}, \tag{$\mathcal{P}_\lambda(y)$}$$

and the noiseless problem $\mathcal{P}_0(y_0)$ minimises $\|\mu\|_{\mathrm{TV}}$ subject to $\Phi\mu = y_0$. The kernel of Φ is $K(x, x') \overset{\text{def.}}{=} \langle \varphi_x, \varphi_{x'} \rangle_{\mathbb{H}}$.

The integral defining $\Phi\mu$ is a Bochner integral in $\mathbb{H}$ (Section 2.5). For a discrete measure it is a finite sum: $\Phi\mu^0 = \sum_i a_i \varphi_{x_i}$. The function $\Phi^\star p$ is continuous, as the composition of $x \mapsto \varphi_x$ with $\langle \cdot, p \rangle_{\mathbb{H}}$, and it is bounded by $\sup_x \|\varphi_x\|_{\mathbb{H}} \|p\|_{\mathbb{H}}$. With $p = \varphi_{x'}$, $\Phi^\star \varphi_{x'} = K(\cdot, x')$. So the functions $\sum_j c_j K(\cdot, x_j)$ lie in $\mathrm{Im}(\Phi^\star)$: certificates can be built from the kernel.

Remark 2.4 (Derivatives). Let $\mathcal{X} = \mathbb{T}$ or $\mathcal{X} = \mathbb{R}$, and assume that $x \mapsto \varphi_x$ is $\mathscr{C}^k$ from $\mathcal{X}$ to $\mathbb{H}$. Then, for all $p \in \mathbb{H}$, $\Phi^\star p$ is $\mathscr{C}^k$ and, for $j \leqslant k$,

$$(\Phi^\star p)^{(j)}(x) = \langle \partial_x^j \varphi_x, p \rangle_{\mathbb{H}}, \qquad \left|(\Phi^\star p)^{(j)}(x)\right| \leqslant \left\|\partial_x^j \varphi_x\right\|_{\mathbb{H}} \|p\|_{\mathbb{H}}\,.$$

Indeed, for $j = 1$, $((\Phi^\star p)(x + t) - (\Phi^\star p)(x))/t = \langle (\varphi_{x+t} - \varphi_x)/t, p \rangle_{\mathbb{H}}$ tends to $\langle \partial_x \varphi_x, p \rangle_{\mathbb{H}}$ when $t \to 0$, and this function of x is continuous. An induction gives $j \leqslant k$. In the same way, $K^{(ij)}(x, x') \overset{\text{def.}}{=} \partial_x^i \partial_{x'}^j K(x, x') = \langle \partial_x^i \varphi_x, \partial_{x'}^j \varphi_{x'} \rangle_{\mathbb{H}}$ for $i, j \leqslant k$. Lecture 3 uses these formulas with $k = 2$.

Example 2.1 (Weighted Fourier measurements). $\mathcal{X} = \mathbb{T}$, cut-off frequency f_c, weights $\Lambda(\omega) > 0$ for $|\omega| \leqslant f_c$, symmetric, of sum one, and $\varphi_x = (\sqrt{\Lambda(\omega)}\, e^{2\mathrm{i}\pi\omega x})_{|\omega| \leqslant f_c}$. Then $\Phi\mu$ is the vector of weighted Fourier coefficients of μ and $K(x, x') = \sum_\omega \Lambda(\omega) \cos(2\pi\omega(x - x'))$. With the Jackson weights, defined for f_c even, this is the Jackson kernel $J_{f_c}(x - x')$.

Here $\omega \in \mathbb{Z}$, $f_c \geqslant 1$ is an integer, and $\mathbb{H} = \mathbb{C}^{2f_c+1}$ with $\langle u, v\rangle_{\mathbb{H}} = \Re \sum_\omega u_\omega \overline{v_\omega}$. The map $x \mapsto \varphi_x$ is $\mathscr{C}^\infty$, since its coordinates are. The coordinates of $\Phi\mu$ are obtained by applying the continuous linear forms $u \mapsto u_\omega$ to the integral: $(\Phi\mu)_\omega = \sqrt{\Lambda(\omega)} \int_{\mathbb{T}} e^{2\mathrm{i}\pi\omega x}\, \mathrm{d}\mu(x)$. This is $\sqrt{\Lambda(\omega)}$ times the Fourier coefficient of index $-\omega$ in the convention of Section 1.1. Since Λ is symmetric, these are the weighted Fourier coefficients, listed in the reverse order.

The kernel. By definition,

$$K(x,x') = \Re \sum_{|\omega|\leqslant f_c} \Lambda(\omega)\, e^{2\mathrm{i}\pi\omega x}\, \overline{e^{2\mathrm{i}\pi\omega x'}} = \sum_{|\omega|\leqslant f_c} \Lambda(\omega) \cos\left(2\pi\omega(x-x')\right).$$

So $K(x,x') = \kappa(x-x')$ with $\kappa(u) \overset{\text{def.}}{=} \sum_\omega \Lambda(\omega)\cos(2\pi\omega u)$. The function κ is even and $\kappa(0) = \sum_\omega \Lambda(\omega) = 1$, so $\|\varphi_x\|_{\mathbb{H}} = 1$ for all x. Since Λ is symmetric, $\kappa(u) = \sum_\omega \Lambda(\omega) e^{2\mathrm{i}\pi\omega u}$.

The Jackson weights. Let f_c be even and $M \overset{\text{def.}}{=} f_c/2$. Let $g(k) \overset{\text{def.}}{=} (M+1-|k|)/(M+1)^2$ for $|k| \leqslant M$, and $g(k) \overset{\text{def.}}{=} 0$ otherwise. Since $\sum_{|k|\leqslant M}(M+1-|k|) = (M+1)^2$, g is a probability on $\mathbb{Z}$. The Jackson weights are $\Lambda \overset{\text{def.}}{=} g \star g$, that is $\Lambda(\omega) = \sum_k g(k)g(\omega-k)$. The term k of this sum is positive if and only if $|k| \leqslant M$ and $|\omega - k| \leqslant M$, and such a k exists if and only if $|\omega| \leqslant 2M = f_c$. So $\Lambda(\omega) > 0$ exactly for $|\omega| \leqslant f_c$. The change $k \to -k$ shows that Λ is symmetric, and $\sum_\omega \Lambda(\omega) = (\sum_k g(k))^2 = 1$. The Fourier series of a convolution is the product of the Fourier series, so

$$\kappa(u) = \sum_\omega \Lambda(\omega) e^{2\mathrm{i}\pi\omega u} = \Big(\sum_{|k|\leqslant M} g(k) e^{2\mathrm{i}\pi k u}\Big)^2.$$

For $|k| \leqslant M$, there are $M+1-|k|$ pairs $(j,j') \in \{0,\dots,M\}^2$ with $j - j' = k$. So, with $z = e^{2\mathrm{i}\pi u}$,

$$\sum_{|k|\leqslant M} (M+1-|k|)\, z^k = \sum_{j,j'=0}^{M} z^j\, \overline{z}^{j'} = \Big|\sum_{j=0}^{M} z^j\Big|^2.$$

For $u \notin \mathbb{Z}$, $\sum_{j=0}^M z^j = (z^{M+1}-1)/(z-1)$ and $\left|e^{\mathrm{i}\theta} - 1\right| = 2\,|\sin(\theta/2)|$. Hence

$$\kappa(u) = J_{f_c}(u) \overset{\text{def.}}{=} \left(\frac{\sin\left(\pi(\frac{f_c}{2}+1)u\right)}{(\frac{f_c}{2}+1)\sin(\pi u)}\right)^4, \qquad u \notin \mathbb{Z},$$

and $\kappa(0) = 1$ is the limit of $J_{f_c}(u)$ when $u \to 0$.

Example 2.2 (Gaussian mixture). $\mathcal{X} = \mathbb{R}$. The data are $z_1,\dots,z_n$ i.i.d. of law $\sum_i a_i \mathcal{N}(x_i,\sigma^2)$ with σ known, $a_i > 0$ and $\sum_i a_i = 1$. Let $\Lambda = \mathcal{N}(0,\sigma^{-2})$, $\mathbb{H} = \mathrm{L}^2(\Lambda)$, $C = 3^{1/4}$ and $\varphi_x(\omega) = C\, e^{\mathrm{i}\omega x} e^{-\sigma^2\omega^2/2}$. The observation is the empirical characteristic function $y_n(\omega) = \frac{C}{n}\sum_k e^{\mathrm{i}\omega z_k}$. Then $\mathbb{E}\left[y_n(\omega)\right] = (\Phi\mu^0)(\omega)$ for every $\omega \in \mathbb{R}$, and $K(x,x') = e^{-(x-x')^2/(6\sigma^2)}$.

Here $\mathbb{H}$ is the space of complex functions f with $\int_{\mathbb{R}} |f|^2\, \mathrm{d}\Lambda < \infty$, with $\langle f, f'\rangle_{\mathbb{H}} = \Re \int_{\mathbb{R}} f\, \overline{f'}\, \mathrm{d}\Lambda$. It is separable. The law Λ has the density $\rho_\sigma(\omega) \overset{\text{def.}}{=} \frac{\sigma}{\sqrt{2\pi}} e^{-\sigma^2\omega^2/2}$. We use the characteristic function of the law $\mathcal{N}(0,\tau^2)$, for $\tau > 0$:

$$\int_{\mathbb{R}} e^{\mathrm{i}\omega u}\, \frac{1}{\sqrt{2\pi}\,\tau}\, e^{-\omega^2/(2\tau^2)}\, \mathrm{d}\omega = e^{-\tau^2 u^2/2}, \qquad u \in \mathbb{R}. \tag{2.4}$$

The kernel. For $x, x' \in \mathbb{R}$,

$$\int_{\mathbb{R}} \varphi_x\, \overline{\varphi_{x'}}\, \mathrm{d}\Lambda = C^2\, \frac{\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\mathrm{i}\omega(x-x')}\, e^{-3\sigma^2\omega^2/2}\, \mathrm{d}\omega.$$

By (2.4) with $\tau^2 = 1/(3\sigma^2)$, the last integral is $\sqrt{2\pi}\,\tau\, e^{-(x-x')^2/(6\sigma^2)}$, and $\sigma\tau = 1/\sqrt{3}$. This value is real, so taking the real part changes nothing, and since $C^2 = \sqrt{3}$,

$$K(x,x') = \frac{C^2}{\sqrt{3}}\, e^{-(x-x')^2/(6\sigma^2)} = e^{-(x-x')^2/(6\sigma^2)}.$$

The constant $C = 3^{1/4}$ is the one that gives $K(x,x) = \|\varphi_x\|_{\mathbb{H}}^2 = 1$. Again $K(x,x') = \kappa(x-x')$, now with $\kappa(u) = e^{-u^2/(6\sigma^2)}$. The map $x \mapsto \varphi_x$ is continuous and bounded, since $\|\varphi_x\|_{\mathbb{H}} = 1$ and $\|\varphi_x - \varphi_{x'}\|_{\mathbb{H}}^2 = 2 - 2\kappa(x-x')$.

The mean. Each $|y_n(\omega)|$ is at most C, so $y_n \in \mathbb{H}$. The characteristic function of $\mathcal{N}(x_i, \sigma^2)$ is $\omega \mapsto e^{\mathrm{i}\omega x_i - \sigma^2\omega^2/2}$. So, for every $\omega \in \mathbb{R}$,

$$\mathbb{E}\,[y_n(\omega)] = C\,\mathbb{E}\,[e^{\mathrm{i}\omega z_1}] = C\sum_{i=1}^{s} a_i\, e^{\mathrm{i}\omega x_i} e^{-\sigma^2\omega^2/2} = \sum_{i=1}^{s} a_i \varphi_{x_i}(\omega) = (\Phi\mu^0)(\omega).$$

The noise is $w_n \overset{\text{def.}}{=} y_n - \Phi\mu^0 \in \mathbb{H}$. It is studied in Lecture 4.

In [10], the mixture lives in $\mathbb{R}^d$ and only m random frequencies of the empirical characteristic function are kept (a sketch); here $d = 1$ and we keep all the frequencies.

Remark 2.5 (Convolution on the torus). Section 4.3 uses a third instance. Let $\mathcal{X} = \mathbb{T}$, $\mathbb{H} = \mathrm{L}^2(\mathbb{T})$ real, $\varphi \in \mathscr{C}^0(\mathbb{T})$ real, and $\varphi_x \overset{\text{def.}}{=} \varphi(\cdot - x)$. The map $x \mapsto \varphi_x$ is continuous, since $\|\varphi_x - \varphi_{x'}\|_{\mathrm{L}^2} \leqslant \sup_t |\varphi(t-x) - \varphi(t-x')|$ and φ is uniformly continuous. For $p \in \mathbb{H}$, the identity of Proposition 2.1 and Fubini's theorem (the integrand is bounded and both measures are finite) give

$$\langle \Phi\mu,\, p\rangle_{\mathbb{H}} = \int_{\mathbb{T}}\int_{\mathbb{T}} \varphi(t-x)p(t)\,\mathrm{d}t\,\mathrm{d}\mu(x) = \int_{\mathbb{T}}\Big(\int_{\mathbb{T}} \varphi(t-x)\,\mathrm{d}\mu(x)\Big)p(t)\,\mathrm{d}t.$$

So $\Phi\mu$ is the function $\varphi \star \mu : t \mapsto \int \varphi(t-x)\,\mathrm{d}\mu(x)$, and $(\Phi^\star p)(x) = \int_{\mathbb{T}} \varphi(t-x)p(t)\,\mathrm{d}t$. The kernel is $K(x,x') = \int_{\mathbb{T}} \varphi(t)\varphi(t+x-x')\,\mathrm{d}t$.

2.4 Radon measures

This section recalls the facts on measures that replace the identification $\mathcal{M}(G) = \mathbb{R}^N$ of Lecture 1. The two theorems are admitted.

Definition 2.3 (Continuous functions and Radon measures)**.** $C_0(\mathcal{X})$ is the space of real continuous functions on $\mathcal{X}$ that vanish at infinity, with the norm $\|\eta\|_\infty = \sup_x |\eta(x)|$. $\mathcal{M}(\mathcal{X})$ is the space of finite regular Borel measures on $\mathcal{X}$. The total variation norm is

$$\|\mu\|_{\mathrm{TV}} \overset{\text{def.}}{=} \sup\{\textstyle\int_{\mathcal{X}} \eta\,\mathrm{d}\mu\ ;\ \eta \in C_0(\mathcal{X}),\ \|\eta\|_\infty \leqslant 1\} = |\mu|\,(\mathcal{X}).$$

A function η vanishes at infinity if the set $\{x\ ;\ |\eta(x)| \geqslant \varepsilon\}$ is compact for every $\varepsilon > 0$. If $\mathcal{X}$ is compact, $C_0(\mathcal{X})$ is the space of all continuous functions. The measures are signed: $\mu(E) \in \mathbb{R}$ for every Borel set E. The variation of μ is the positive measure $|\mu|$ defined by $|\mu|\,(E) \overset{\text{def.}}{=} \sup \sum_k |\mu(E_k)|$, where the supremum is over the finite partitions of E into Borel sets E_k. The second equality in the definition of $\|\mu\|_{\mathrm{TV}}$ is part of Theorem 2.1. We also admit the polar decomposition [11]: there is a Borel function $\sigma_\mu : \mathcal{X} \to \{-1, 1\}$ with $\mathrm{d}\mu = \sigma_\mu\,\mathrm{d}\,|\mu|$. It plays the role of the sign of Lecture 1. As in Lecture 1 we write $\langle \eta,\, \mu\rangle_{\mathcal{M}} \overset{\text{def.}}{=} \int_{\mathcal{X}} \eta\,\mathrm{d}\mu = \int_{\mathcal{X}} \eta\,\sigma_\mu\,\mathrm{d}\,|\mu|$ for a bounded Borel function η. The two facts of Lecture 1 hold with the same proofs:

$$|\langle \eta,\, \mu\rangle_{\mathcal{M}}| \leqslant \int_{\mathcal{X}} |\eta|\ \mathrm{d}\,|\mu| \leqslant \|\eta\|_\infty\ \|\mu\|_{\mathrm{TV}}, \tag{2.5}$$

$$\|\mu\|_{\mathrm{TV}} - \langle \eta,\, \mu\rangle_{\mathcal{M}} = \int_{\mathcal{X}} (1 - \eta\,\sigma_\mu)\,\mathrm{d}\,|\mu| \geqslant \int_{\mathcal{X}} (1 - |\eta|)\,\mathrm{d}\,|\mu|\,. \tag{2.6}$$

The equality uses $\|\mu\|_{\mathrm{TV}} = |\mu|\,(\mathcal{X})$, and the inequality holds because $\eta\,\sigma_\mu \leqslant |\eta|$. These are the analogues of the second part of (1.1) and of (1.2). The first part of (1.1) is Proposition 2.1.

Discrete measures. Let $\mu = \sum_{i=1}^{s} a_i \delta_{x_i}$ with pairwise distinct x_i and $a_i \neq 0$. For a Borel set E and a finite partition (E_k) of E, $\sum_k |\mu(E_k)| \leqslant \sum_{x_i \in E} |a_i|$, with equality when each $x_i \in E$ lies in its own E_k. Hence

$$|\mu| = \sum_{i=1}^{s} |a_i| \, \delta_{x_i}, \qquad \|\mu\|_{\mathrm{TV}} = \sum_{i=1}^{s} |a_i| \,, \qquad \langle \eta, \, \mu \rangle_{\mathcal{M}} = \sum_{i=1}^{s} a_i \eta(x_i). \tag{2.7}$$

In particular $\| \pm \delta_x \|_{\mathrm{TV}} = 1$ for every $x \in \mathcal{X}$.

Remark 2.6 (Support of a measure). The support $\mathrm{supp}(\mu)$ is the set of $x \in \mathcal{X}$ such that $|\mu|(U) > 0$ for every open set $U \ni x$. It is closed. By (2.7), the support of the discrete measure above is $\{x_1, \ldots, x_s\}$, since a finite set is closed. (i) $|\mu|(\mathcal{X} \backslash \mathrm{supp}(\mu)) = 0$. Indeed, every $x \notin \mathrm{supp}(\mu)$ has an open neighbourhood U_x with $|\mu|(U_x) = 0$. Since $\mathcal{X}$ has a countable basis of open sets, countably many U_x cover $\mathcal{X} \setminus \mathrm{supp}(\mu)$, and their union has measure zero. So, if $\mathrm{supp}(\mu) \subset \{x_1, \ldots, x_s\}$, then $\mu = \sum_i \mu(\{x_i\}) \, \delta_{x_i}$, since $|\mu(E)| \leqslant |\mu|(E)$ for every Borel set E. (ii) Let $\eta \in C_0(\mathcal{X})$ with $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \, \mu \rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$. Then $\mathrm{supp}(\mu) \subset \{x \in \mathcal{X} \, ; \, |\eta(x)| = 1\}$. Indeed, by (2.6), $\int (1 - |\eta|) \, \mathrm{d}\, |\mu| \leqslant 0$, and the integrand is nonnegative. Let x with $|\eta(x)| < 1$. By continuity, there are an open set $U \ni x$ and $c > 0$ with $1 - |\eta| \geqslant c$ on U. Then $c \, |\mu|(U) \leqslant \int (1 - |\eta|) \, \mathrm{d}\, |\mu| \leqslant 0$, so $|\mu|(U) = 0$ and $x \notin \mathrm{supp}(\mu)$.

Point (ii) is the measure version of the phrase "$|\eta(t)| = 1$ for every $t \in \mathrm{supp}(\nu)$" in the proof of Proposition 1.1. It needs the continuity of η.

Theorem 2.1 (Riesz-Markov-Kakutani). *The map $\mu \mapsto (\eta \mapsto \int \eta \, \mathrm{d}\mu)$ is an isometric isomorphism from $\mathcal{M}(\mathcal{X})$ with the total variation norm onto the dual of $C_0(\mathcal{X})$.*

So $\mathcal{M}(\mathcal{X})$ is a dual space, and the total variation norm is the dual norm of the uniform norm. This is why certificates are continuous functions.

Definition 2.4 (Weak-$\star$ convergence). $\mu_n \overset{\star}{\rightharpoonup} \mu$ if $\int \eta \, \mathrm{d}\mu_n \to \int \eta \, \mathrm{d}\mu$ for all $\eta \in C_0(\mathcal{X})$.

This is the convergence of sequences for the weak-$\star$ topology, the coarsest topology on $\mathcal{M}(\mathcal{X})$ for which the maps $\mu \mapsto \int \eta \, \mathrm{d}\mu$, $\eta \in C_0(\mathcal{X})$, are continuous.

Example 2.3. If $x_n \to x$ with $x_n \neq x$, then $\delta_{x_n} \overset{\star}{\rightharpoonup} \delta_x$ and $\|\delta_{x_n} - \delta_x\|_{\mathrm{TV}} = 2$.

Indeed, $\int \eta \, \mathrm{d}\delta_{x_n} = \eta(x_n) \to \eta(x)$ by continuity of η, and (2.7) gives $\|\delta_{x_n} - \delta_x\|_{\mathrm{TV}} = 1 + 1$. A small error on a position is small for the weak-$\star$ topology and maximal for the total variation. This is why the guarantees of Lectures 3 and 4 measure errors on positions with distances on $\mathcal{X}$.

Theorem 2.2 (Banach-Alaoglu). *For all $R \geqslant 0$ the ball $\{\mu \in \mathcal{M}(\mathcal{X}) \, ; \, \|\mu\|_{\mathrm{TV}} \leqslant R\}$ is weak-$\star$ compact and metrisable. Every bounded sequence of measures has a weak-$\star$ convergent subsequence.*

Compactness is the Banach-Alaoglu theorem in the dual of $C_0(\mathcal{X})$ [12]. Metrisability uses that $C_0(\mathcal{X})$ is separable, which holds because $\mathcal{X}$ has a countable basis of open sets. On a ball, closed and sequentially closed sets are then the same, and the last sentence follows: a compact metric space is sequentially compact.

2.5 The BLASSO: existence, subdifferential, duality

The proofs of this section are those of Section 1.4, with $\langle \eta, \, \mu \rangle_{\mathcal{M}} = \int \eta \, \mathrm{d}\mu$. Three points change (Remark 1.7): existence uses Theorem 2.2, the function $\Phi^\star p$ must lie in $C_0(\mathcal{X})$, and statements on supports use Remark 2.6.

We admit the following facts on the Bochner integral [13]. Let ν be a finite positive Borel measure on $\mathcal{X}$ and $f : \mathcal{X} \to \mathbb{H}$ be bounded and strongly measurable, for instance continuous, or continuous times a bounded real Borel function: since $\mathbb{H}$ is separable, measurability follows

from Pettis' theorem. Then the Bochner integral $\int f \,\mathrm{d}\nu \in \mathbb{H}$ is defined, it is linear in f, $\|\int f \,\mathrm{d}\nu\|_{\mathbb{H}} \leqslant \int \|f\|_{\mathbb{H}} \,\mathrm{d}\nu$, and $\ell(\int f \,\mathrm{d}\nu) = \int \ell \circ f \,\mathrm{d}\nu$ for every continuous linear form ℓ on $\mathbb{H}$. For $\mu \in \mathcal{M}(\mathcal{X})$ we set $\int f \,\mathrm{d}\mu \overset{\text{def.}}{=} \int f \,\sigma_\mu \,\mathrm{d}\,|\mu|$. With $f(x) = \varphi_x$ this defines $\Phi\mu$, and $\|\Phi\mu\|_{\mathbb{H}} \leqslant B\|\mu\|_{\mathrm{TV}}$ with $B \overset{\text{def.}}{=} \sup_x \|\varphi_x\|_{\mathbb{H}}$.

Proposition 2.1 (Weak-$\star$ to weak continuity)**.** *Assume that $x \mapsto \langle \varphi_x,\, p\rangle_{\mathbb{H}}$ is in $C_0(\mathcal{X})$ for all $p \in \mathbb{H}$. Then $\Phi^\star p \in C_0(\mathcal{X})$, $\langle \Phi\mu,\, p\rangle_{\mathbb{H}} = \int \Phi^\star p \,\mathrm{d}\mu$, and $\mu_n \overset{\star}{\rightharpoonup} \mu$ implies $\Phi\mu_n \rightharpoonup \Phi\mu$ weakly in $\mathbb{H}$.*

Proof. The first claim is the assumption, since $\Phi^\star p = \langle \varphi_\cdot,\, p\rangle_{\mathbb{H}}$. Let $p \in \mathbb{H}$ and $\mu \in \mathcal{M}(\mathcal{X})$. The map $\ell : q \mapsto \langle q,\, p\rangle_{\mathbb{H}}$ is a continuous linear form on $\mathbb{H}$, so

$$\langle \Phi\mu,\, p\rangle_{\mathbb{H}} = \ell\Big(\int \varphi_x \,\sigma_\mu(x) \,\mathrm{d}\,|\mu|\,(x)\Big) = \int \langle \varphi_x,\, p\rangle_{\mathbb{H}} \,\sigma_\mu(x) \,\mathrm{d}\,|\mu|\,(x) = \langle \Phi^\star p,\, \mu\rangle_{\mathcal{M}}. \tag{2.8}$$

Let $\mu_n \overset{\star}{\rightharpoonup} \mu$ and $p \in \mathbb{H}$. Definition 2.4 applied to $\eta = \Phi^\star p \in C_0(\mathcal{X})$ gives $\langle \Phi\mu_n,\, p\rangle_{\mathbb{H}} = \int \Phi^\star p \,\mathrm{d}\mu_n \to \int \Phi^\star p \,\mathrm{d}\mu = \langle \Phi\mu,\, p\rangle_{\mathbb{H}}$. This holds for every p, which is the weak convergence $\Phi\mu_n \rightharpoonup \Phi\mu$. □

The identity (2.8) is the first part of (1.1), now on measures. One can also avoid the Bochner integral. The linear form $p \mapsto \int \Phi^\star p \,\mathrm{d}\mu$ is bounded by $B\|\mu\|_{\mathrm{TV}}\,\|p\|_{\mathbb{H}}$ by (2.5). The Riesz representation theorem in $\mathbb{H}$ gives a unique $\Phi\mu \in \mathbb{H}$ that satisfies (2.8) for all p, and one can take this as the definition. Only (2.8) is used below.

Remark 2.7 (The assumption in the examples). (a) If $\mathcal{X}$ is compact, $C_0(\mathcal{X})$ is the space of continuous functions, and $\Phi^\star p$ is continuous (Section 2.3). So the assumption holds for Example 2.1, where $\Phi^\star p$ is even a trigonometric polynomial, and for Remark 2.5. (b) In Example 2.2, $\mathcal{X} = \mathbb{R}$ is not compact, and we also need the decay at infinity. Let $p \in \mathbb{H}$. Then $(\Phi^\star p)(x) = \Re \int_{\mathbb{R}} e^{\mathrm{i}\omega x} g(\omega) \,\mathrm{d}\omega$ with $g(\omega) \overset{\text{def.}}{=} Ce^{-\sigma^2\omega^2/2}\,\overline{p(\omega)}\,\rho_\sigma(\omega)$. Since Λ is a probability, the Cauchy-Schwarz inequality gives $\int_{\mathbb{R}} |g| \,\mathrm{d}\omega \leqslant C \int_{\mathbb{R}} |p| \,\mathrm{d}\Lambda \leqslant C\,\|p\|_{\mathbb{H}}$, so g is integrable. By the Riemann-Lebesgue lemma, which we admit [11], the Fourier transform of an integrable function is continuous and tends to 0 at infinity. So $\Phi^\star p \in C_0(\mathbb{R})$. (c) If $\mathcal{X}$ is not compact, the assumption is a true restriction. For $\mathcal{X} = \mathbb{R}$ and $\varphi_x = e$ a fixed unit vector, $\Phi^\star e = 1$ is not in $C_0(\mathbb{R})$, and $\delta_n \overset{\star}{\rightharpoonup} 0$ while $\Phi\delta_n = e$ for all n. Exercise 2.4 shows that the BLASSO can then have no solution.

From now on, the assumption of Proposition 2.1 holds.

Proposition 2.2 (Existence)**.** *Under the assumption of Proposition 2.1, for all $y \in \mathbb{H}$ and $\lambda > 0$ the sublevel sets of $J_\lambda(\mu) \overset{\text{def.}}{=} \frac{1}{2}\,\|\Phi\mu - y\|_{\mathbb{H}}^2 + \lambda\|\mu\|_{\mathrm{TV}}$ are weak-$\star$ compact and $\mathcal{P}_\lambda(y)$ has a solution.*

Proof. Lower semi-continuity. Let $\mu_n \overset{\star}{\rightharpoonup} \mu$. We show $J_\lambda(\mu) \leqslant \liminf_n J_\lambda(\mu_n)$.

First, the total variation. Let $\eta \in C_0(\mathcal{X})$ with $\|\eta\|_\infty \leqslant 1$. By (2.5), $\int \eta \,\mathrm{d}\mu_n \leqslant \|\mu_n\|_{\mathrm{TV}}$ for all n. Letting $n \to \infty$ gives $\int \eta \,\mathrm{d}\mu \leqslant \liminf_n \|\mu_n\|_{\mathrm{TV}}$. The supremum over such η is $\|\mu\|_{\mathrm{TV}}$ by Definition 2.3. So $\|\mu\|_{\mathrm{TV}} \leqslant \liminf_n \|\mu_n\|_{\mathrm{TV}}$: a supremum of weak-$\star$ continuous linear forms is weak-$\star$ lower semi-continuous.

Second, the data term. By Proposition 2.1, $u_n \overset{\text{def.}}{=} \Phi\mu_n - y$ converges weakly to $u \overset{\text{def.}}{=} \Phi\mu - y$. Then $\|u\|_{\mathbb{H}}^2 = \lim_n \langle u_n,\, u\rangle_{\mathbb{H}} \leqslant \liminf_n \|u_n\|_{\mathbb{H}}\,\|u\|_{\mathbb{H}}$. If $u \neq 0$, divide by $\|u\|_{\mathbb{H}}$. In all cases $\|u\|_{\mathbb{H}} \leqslant \liminf_n \|u_n\|_{\mathbb{H}}$. Adding the two inequalities gives the claim, since the lim inf of a sum is at least the sum of the lim inf.

Compactness. Let $c \in \mathbb{R}$ and $S_c \overset{\text{def.}}{=} \{\mu\,;\, J_\lambda(\mu) \leqslant c\}$. If $c < 0$, S_c is empty. Otherwise, $\lambda\|\mu\|_{\mathrm{TV}} \leqslant J_\lambda(\mu) \leqslant c$ on S_c, so S_c is contained in the ball B_R of radius $R = c/\lambda$. By Theorem 2.2, B_R with the weak-$\star$ topology is a compact metric space. If $\mu_n \in S_c$ and $\mu_n \overset{\star}{\rightharpoonup} \mu$, then $J_\lambda(\mu) \leqslant \liminf_n J_\lambda(\mu_n) \leqslant c$, so $\mu \in S_c$. So S_c is sequentially closed in the metric space B_R, hence closed in B_R, hence compact.

Existence. Let $m \overset{\text{def.}}{=} \inf J_\lambda$. Then $0 \leqslant m \leqslant J_\lambda(0) < \infty$. Let (μ_n) be a minimising sequence. For n large, $J_\lambda(\mu_n) \leqslant m+1$, so $\mu_n \in S_{m+1}$. By compactness, a subsequence (μ_{n_k}) converges weak-$\star$ to some $\hat{\mu}$. By lower semi-continuity, $J_\lambda(\hat{\mu}) \leqslant \liminf_k J_\lambda(\mu_{n_k}) = m$. So $\hat{\mu}$ is a solution. □

The total variation is not weak-$\star$ continuous. On $\mathbb{T}$, $\delta_{1/n} - \delta_0 \overset{\star}{\rightharpoonup} 0$ while $\|\delta_{1/n} - \delta_0\|_{\mathrm{TV}} = 2$ (Example 2.3). Lower semi-continuity is all the proof needs.

Proposition 2.3 (Subdifferential of the total variation)**.** *Let $\mu \in \mathcal{M}(\mathcal{X})$ and $\eta \in C_0(\mathcal{X})$. Then $\|\nu\|_{\mathrm{TV}} \geqslant \|\mu\|_{\mathrm{TV}} + \int \eta \,\mathrm{d}(\nu - \mu)$ for all $\nu \in \mathcal{M}(\mathcal{X})$ if and only if $\|\eta\|_\infty \leqslant 1$ and $\int \eta \,\mathrm{d}\mu = \|\mu\|_{\mathrm{TV}}$. We write $\eta \in \partial\|\mu\|_{\mathrm{TV}}$. For $\mu = \sum_{i=1}^s a_i \delta_{x_i}$ with pairwise distinct x_i and $a_i \neq 0$, the second condition reads $\eta(x_i) = \mathrm{sign}(a_i)$ for all i.*

Proof. Assume the inequality for all ν. With $\nu = 0$ it gives $0 \geqslant \|\mu\|_{\mathrm{TV}} - \langle \eta, \mu \rangle_{\mathcal{M}}$. With $\nu = 2\mu$ it gives $2\|\mu\|_{\mathrm{TV}} \geqslant \|\mu\|_{\mathrm{TV}} + \langle \eta, \mu \rangle_{\mathcal{M}}$. Hence $\langle \eta, \mu \rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$, and the inequality becomes $\|\nu\|_{\mathrm{TV}} \geqslant \langle \eta, \nu \rangle_{\mathcal{M}}$ for all ν. Let $x \in \mathcal{X}$. The Dirac mass δ_x lies in $\mathcal{M}(\mathcal{X})$ and $\|\pm\delta_x\|_{\mathrm{TV}} = 1$ by (2.7). With $\nu = \pm\delta_x$ the inequality gives $1 \geqslant \pm\eta(x)$. So $|\eta(x)| \leqslant 1$ for all $x \in \mathcal{X}$, that is $\|\eta\|_\infty \leqslant 1$.

Conversely, assume $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \mu \rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$. For all ν, by (2.5), $\|\mu\|_{\mathrm{TV}} + \langle \eta, \nu - \mu \rangle_{\mathcal{M}} = \langle \eta, \nu \rangle_{\mathcal{M}} \leqslant \|\nu\|_{\mathrm{TV}}$.

Last, let $\mu = \sum_i a_i \delta_{x_i}$ as in the statement, and assume $\|\eta\|_\infty \leqslant 1$. By (2.7), $\|\mu\|_{\mathrm{TV}} - \langle \eta, \mu \rangle_{\mathcal{M}} = \sum_i |a_i| \,(1 - \eta(x_i)\,\mathrm{sign}(a_i))$, and each term is nonnegative. The sum vanishes if and only if $\eta(x_i)\,\mathrm{sign}(a_i) = 1$ for all i, that is $\eta(x_i) = \mathrm{sign}(a_i)$ for all i. □

The tests $\pm\delta_x$ use every point $x \in \mathcal{X}$, not only the points of a grid. This is why the condition $|\eta| \leqslant 1$ now holds on all of $\mathcal{X}$. The same tests give the conjugate of the total variation, used in Lecture 5. Let $\eta \in C_0(\mathcal{X})$. If $\|\eta\|_\infty \leqslant 1$, then $\langle \eta, \mu \rangle_{\mathcal{M}} - \|\mu\|_{\mathrm{TV}} \leqslant 0$ for all μ, with equality at $\mu = 0$. If $|\eta(x)| > 1$ for some x, the measures $\mu_t \overset{\text{def.}}{=} t\,\mathrm{sign}(\eta(x))\delta_x$, $t > 0$, give $\langle \eta, \mu_t \rangle_{\mathcal{M}} - \|\mu_t\|_{\mathrm{TV}} = t(|\eta(x)| - 1) \to \infty$. So $\sup_\mu \left(\langle \eta, \mu \rangle_{\mathcal{M}} - \|\mu\|_{\mathrm{TV}}\right)$ is 0 if $\|\eta\|_\infty \leqslant 1$ and $+\infty$ otherwise.

Definition 2.5 (Dual certificate)**.** A dual certificate for μ is a function $\eta \in \mathrm{Im}(\Phi^\star) \cap \partial\|\mu\|_{\mathrm{TV}}$.

For $\mu^0 = \sum_i a_i \delta_{x_i}$, a dual certificate is a function $\eta = \Phi^\star p$, with $p \in \mathbb{H}$, such that $\|\eta\|_\infty \leqslant 1$ and $\eta(x_i) = \mathrm{sign}(a_i)$ for all i. Its existence is the source condition of Definition 3.1.

Proposition 2.4 (Optimality conditions and strong duality)**.** *Under the assumption of Proposition 2.1, let $\lambda > 0$, $y \in \mathbb{H}$ and $C \overset{\text{def.}}{=} \{p \in \mathbb{H} \,;\, \|\Phi^\star p\|_\infty \leqslant 1\}$.*

1. *$\hat{\mu}$ is a solution of $\mathcal{P}_\lambda(y)$ if and only if $\Phi^\star \hat{p}_\lambda \in \partial\|\hat{\mu}\|_{\mathrm{TV}}$ with $\hat{p}_\lambda \overset{\text{def.}}{=} (y - \Phi\hat{\mu})/\lambda$.*
2. *$\hat{p}_\lambda$ is the projection of y/λ onto the closed convex set C. In particular $\Phi\hat{\mu}$ is the same for all solutions.*
3. *$\min \mathcal{P}_\lambda(y) = \max_{p \in C} \lambda\langle p, y \rangle_{\mathbb{H}} - \frac{\lambda^2}{2}\|p\|_{\mathbb{H}}^2$, and the maximum is attained at $\hat{p}_\lambda$ only.*
4. *$\mathrm{supp}(\hat{\mu}) \subset \{x \in \mathcal{X} \,;\, |\hat{\eta}_\lambda(x)| = 1\}$ with $\hat{\eta}_\lambda \overset{\text{def.}}{=} \Phi^\star \hat{p}_\lambda$.*

Proof. Let $J(\mu) \overset{\text{def.}}{=} \frac{1}{2}\|\Phi\mu - y\|^2 + \lambda\|\mu\|_{\mathrm{TV}}$ and $D(p) \overset{\text{def.}}{=} \lambda\langle p, y \rangle - \frac{\lambda^2}{2}\|p\|^2$. All norms and scalar products are those of $\mathbb{H}$. By assumption $\Phi^\star p \in C_0(\mathcal{X})$ for every p, so Proposition 2.3 applies to $\eta = \Phi^\star p$, and (2.8) holds. By Proposition 2.2, $\mathcal{P}_\lambda(y)$ has a solution. We write $\hat{p}$ for $\hat{p}_\lambda$.

The set C. For $x \in \mathcal{X}$, the map $p \mapsto \langle \varphi_x, p \rangle$ is a continuous linear form, so $C_x \overset{\text{def.}}{=} \{p \,;\, |\langle \varphi_x, p \rangle| \leqslant 1\}$ is closed and convex, as the inverse image of $[-1, 1]$. So $C = \bigcap_{x \in \mathcal{X}} C_x$ is closed and convex, as an intersection of closed convex sets, and $0 \in C$. The intersection is infinite, but this plays no role.

(1) Let $\hat{\mu}$ be a solution, $\nu \in \mathcal{M}(\mathcal{X})$ and $\tau \in (0,1]$. Expanding the square, with $\Phi\hat{\mu} - y = -\lambda\hat{p}$,

$$\tfrac{1}{2}\left\|\Phi(\hat{\mu}+\tau\nu)-y\right\|^2 - \tfrac{1}{2}\left\|\Phi\hat{\mu}-y\right\|^2 = -\tau\lambda\langle\hat{p},\,\Phi\nu\rangle + \tfrac{\tau^2}{2}\left\|\Phi\nu\right\|^2.$$

Since $\hat{\mu}+\tau\nu = (1-\tau)\hat{\mu}+\tau(\hat{\mu}+\nu)$, the convexity of the total variation gives $\|\hat{\mu}+\tau\nu\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}} \leqslant \tau(\|\hat{\mu}+\nu\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}})$. So the optimality $J(\hat{\mu}+\tau\nu) \geqslant J(\hat{\mu})$ gives, after division by $\tau\lambda$,

$$0 \leqslant -\langle\hat{p},\,\Phi\nu\rangle + \tfrac{\tau}{2\lambda}\left\|\Phi\nu\right\|^2 + \|\hat{\mu}+\nu\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}}.$$

Let $\tau \to 0$ and use (2.8): $\|\hat{\mu}+\nu\|_{\mathrm{TV}} \geqslant \|\hat{\mu}\|_{\mathrm{TV}} + \langle\Phi^\star\hat{p},\,\nu\rangle_{\mathcal{M}}$ for all $\nu \in \mathcal{M}(\mathcal{X})$. Replacing ν by $\nu - \hat{\mu}$, this is the inequality of Proposition 2.3, that is $\Phi^\star\hat{p} \in \partial\|\hat{\mu}\|_{\mathrm{TV}}$.

Conversely, let $\hat{\mu} \in \mathcal{M}(\mathcal{X})$ with $\Phi^\star\hat{p} \in \partial\|\hat{\mu}\|_{\mathrm{TV}}$, and let $\nu \in \mathcal{M}(\mathcal{X})$. The same expansion, with $\tau = 1$ and $\nu - \hat{\mu}$ in place of ν, gives

$$\tfrac{1}{2}\left\|\Phi\nu-y\right\|^2 - \tfrac{1}{2}\left\|\Phi\hat{\mu}-y\right\|^2 = -\lambda\langle\hat{p},\,\Phi(\nu-\hat{\mu})\rangle + \tfrac{1}{2}\left\|\Phi(\nu-\hat{\mu})\right\|^2.$$

The subgradient inequality and (2.8) give $\lambda\|\nu\|_{\mathrm{TV}} - \lambda\|\hat{\mu}\|_{\mathrm{TV}} \geqslant \lambda\langle\Phi^\star\hat{p},\,\nu-\hat{\mu}\rangle_{\mathcal{M}} = \lambda\langle\hat{p},\,\Phi(\nu-\hat{\mu})\rangle$. Adding,

$$J(\nu) - J(\hat{\mu}) \geqslant \tfrac{1}{2}\left\|\Phi(\nu-\hat{\mu})\right\|^2 \geqslant 0. \tag{2.9}$$

So $\hat{\mu}$ is a solution.

(2) Let $\hat{\mu}$ be a solution. By (1) and Proposition 2.3, $\left\|\Phi^\star\hat{p}\right\|_\infty \leqslant 1$, so $\hat{p} \in C$, and $\langle\Phi^\star\hat{p},\,\hat{\mu}\rangle_{\mathcal{M}} = \|\hat{\mu}\|_{\mathrm{TV}}$. Let $z \overset{\text{def.}}{=} y/\lambda$, so that $\lambda(z-\hat{p}) = \Phi\hat{\mu}$. For $q \in C$, by (2.8) and (2.5),

$$\lambda\langle z-\hat{p},\,q-\hat{p}\rangle = \langle\Phi\hat{\mu},\,q-\hat{p}\rangle = \langle\Phi^\star q,\,\hat{\mu}\rangle_{\mathcal{M}} - \|\hat{\mu}\|_{\mathrm{TV}} \leqslant \left\|\Phi^\star q\right\|_\infty\|\hat{\mu}\|_{\mathrm{TV}} - \|\hat{\mu}\|_{\mathrm{TV}} \leqslant 0.$$

Hence, for all $q \in C$,

$$\|z-q\|^2 = \|z-\hat{p}\|^2 - 2\langle z-\hat{p},\,q-\hat{p}\rangle + \|q-\hat{p}\|^2 \geqslant \|z-\hat{p}\|^2 + \|q-\hat{p}\|^2.$$

So $\hat{p}$ is the unique point of C at minimal distance from z, that is the projection of y/λ onto C. This argument proves the existence and the uniqueness of this projection directly, from the existence of $\hat{\mu}$; the projection theorem in Hilbert spaces is not needed. The point $\hat{p}$ depends only on y, λ and C, not on the solution $\hat{\mu}$. So $\Phi\hat{\mu} = y - \lambda\hat{p}$ is the same for all solutions.

(3) Weak duality. Let $\mu \in \mathcal{M}(\mathcal{X})$ and $p \in \mathbb{H}$. Expanding the square $\frac{1}{2}\left\|\Phi\mu - y + \lambda p\right\|^2$ and using $\lambda\langle p,\,\Phi\mu\rangle = \lambda\langle\Phi^\star p,\,\mu\rangle_{\mathcal{M}}$ gives the identity

$$J(\mu) - D(p) = \tfrac{1}{2}\left\|\Phi\mu - y + \lambda p\right\|^2 + \lambda(\|\mu\|_{\mathrm{TV}} - \langle\Phi^\star p,\,\mu\rangle_{\mathcal{M}}). \tag{2.10}$$

If $p \in C$, both terms are nonnegative by (2.5), so $J(\mu) \geqslant D(p)$.

Strong duality. Let $\hat{\mu}$ be a solution. In (2.10) for $(\hat{\mu},\hat{p})$, the first term is zero by definition of $\hat{p}$, and the second one is zero by (2). So $D(\hat{p}) = J(\hat{\mu})$, and with weak duality, $D(p) \leqslant J(\hat{\mu}) = D(\hat{p})$ for all $p \in C$. Hence $\min\mathcal{P}_\lambda(y) = J(\hat{\mu}) = \max_C D$, attained at $\hat{p} \in C$.

Uniqueness of the maximiser. Let $p \in C$ with $D(p) = \max_C D = J(\hat{\mu})$. Then (2.10) for $(\hat{\mu},p)$ is a sum of two nonnegative terms equal to zero. So $\left\|\Phi\hat{\mu} - y + \lambda p\right\| = 0$, that is $p = (y - \Phi\hat{\mu})/\lambda = \hat{p}$.

(4) By (2), $\hat{\eta}_\lambda = \Phi^\star\hat{p} \in C_0(\mathcal{X})$ satisfies $\left\|\hat{\eta}_\lambda\right\|_\infty \leqslant 1$ and $\langle\hat{\eta}_\lambda,\,\hat{\mu}\rangle_{\mathcal{M}} = \|\hat{\mu}\|_{\mathrm{TV}}$. Remark 2.6 (ii) gives $\mathrm{supp}(\hat{\mu}) \subset \{x \in \mathcal{X}\ ;\ |\hat{\eta}_\lambda(x)| = 1\}$. □

Remark 2.8 (Notation and consequences). We write $\hat{p}_\lambda(y)$ for the projection of y/λ onto C, and $\hat{\eta}_\lambda = \Phi^\star\hat{p}_\lambda(y)$. Lectures 3 to 5 use the following facts. (a) By (2), $\Phi\hat{\mu} = y - \lambda\hat{p}_\lambda(y)$ is the same for all solutions, and so is $\|\hat{\mu}\|_{\mathrm{TV}} = \big(\min\mathcal{P}_\lambda(y) - \frac{1}{2}\left\|\lambda\hat{p}_\lambda(y)\right\|_{\mathbb{H}}^2\big)/\lambda$. (b) By (1) with $\hat{\mu} = 0$, the measure 0 is a solution if and only if $\left\|\Phi^\star y\right\|_\infty \leqslant \lambda$. (c) The projection onto a closed convex set is 1-Lipschitz [14]. With (2), this gives Lemma 4.2. (d) If the set $\{x\ ;\ |\hat{\eta}_\lambda(x)| = 1\}$ is finite, then (4) and Remark 2.6 (i) show that every solution is a discrete measure supported in this set. If the atoms φ_x indexed by this set are linearly independent, (a) gives a unique solution. Theorem 4.2 and Appendix A use item (4) in this way.

Remark 2.9 (The general frame). Items (1) and (3) are an instance of Fenchel-Rockafellar duality [15]. We only use weak duality in general form: for $\Phi\mu = y_0$ and $p \in C$, $\langle p,\, y_0\rangle_{\mathbb{H}} = \int \Phi^\star p \,\mathrm{d}\mu \leqslant \|\mu\|_{\mathrm{TV}}$.

The three lines are $\langle p,\, y_0\rangle_{\mathbb{H}} = \langle p,\, \Phi\mu\rangle_{\mathbb{H}}$, then (2.8), then (2.5) with $\|\Phi^\star p\|_\infty \leqslant 1$. By Proposition 2.3, equality holds if and only if $\Phi^\star p \in \partial\|\mu\|_{\mathrm{TV}}$, and then Remark 2.6 (ii) gives $\mathrm{supp}(\mu) \subset \{x\ ;\ |(\Phi^\star p)(x)| = 1\}$. This is what Proposition 3.1 and Lemma 3.1 use.

Exercises

Exercise 2.1 (The kernels of the two examples)**.** Let κ be the function with $K(x,x') = \kappa(x-x')$ in Example 2.1 or in Example 2.2.

1. In Example 2.1, compute $\partial_x\varphi_x$ and show that $K^{(ij)}(x,x') = (-1)^j\kappa^{(i+j)}(x-x')$ for $i,j \leqslant 2$. Deduce that $\|\partial_x\varphi_x\|^2_{\mathbb{H}} = -\kappa''(0) = 4\pi^2\sum_\omega \omega^2\Lambda(\omega)$.
2. Still in Example 2.1, show that $\kappa(u) < 1$ for $u \in \mathbb{T}\setminus\{0\}$ and that $\kappa(u) > -1$ for all u. Hint: $\Lambda(0) > 0$ and $\Lambda(1) > 0$.
3. For the Jackson weights, show that $\sum_k k^2 g(k) = M(M+2)/6$. Deduce that $-\kappa''(0) = \frac{4\pi^2}{3}M(M+2) = \frac{\pi^2}{3}f_c(f_c+4)$.
4. In Example 2.2, show that $x \mapsto \varphi_x$ is differentiable from $\mathbb{R}$ to $\mathbb{H}$ with $\partial_x\varphi_x(\omega) = \mathrm{i}\omega\varphi_x(\omega)$. Hint: $\left|e^{\mathrm{i}t} - 1 - \mathrm{i}t\right| \leqslant \min(2\,|t|\,, t^2/2)$ for $t \in \mathbb{R}$, and dominated convergence.
5. Compute $\|\partial_x\varphi_x\|^2_{\mathbb{H}}$ in Example 2.2 in two ways: with the moments of a Gaussian law, and with $-\kappa''(0)$. Check that both give $1/(3\sigma^2)$.
6. Show that both kernels satisfy the assumptions of Example 3.1: κ is even, $\kappa(0) = 1$, $\kappa''(0) < 0$ and $|\kappa(u)| < 1$ for $u \neq 0$.

Exercise 2.2 (The Fuchs pre-certificate of two spikes)**.** Let Φ be as in Example 2.1, on $\mathcal{X} = \mathbb{T}$, with the function κ of Exercise 2.1. Let $0 < d < 1/2$, $a_1, a_2 > 0$ and $\mu^0 = a_1\delta_0 + a_2\delta_d$. Let $G \subset \mathbb{T}$ be a finite grid that contains 0 and d.

1. Show that Φ_{x_0} is injective. Compute p_F and show that $\eta_F = \big(\kappa(\cdot) + \kappa(\cdot - d)\big)/\big(1 + \kappa(d)\big)$.
2. Show that $\eta_F'(0) = -\kappa'(d)/(1+\kappa(d))$.
3. Assume $\kappa'(d) \neq 0$. Show that there is $h_0 > 0$ such that the assumption of Corollary 2.1 fails for every grid G that contains $-h$ and h, with $0 < h < h_0$.
4. Let $\eta = \Phi^\star p$ be any certificate of μ^0 on $\mathbb{T}$, that is $\|\eta\|_\infty \leqslant 1$ on $\mathbb{T}$ and $\eta(0) = \eta(d) = 1$. Show that $\eta'(0) = \eta'(d) = 0$. Deduce that $\eta \neq \eta_F$ when $\kappa'(d) \neq 0$.
5. Let η be as in question 4, and let G contain h with $0 < h < d$. Show that the margin ε of η on $G\setminus\{0,d\}$ is at most $2\pi^2 f_c^2\,\|p\|_{\mathbb{H}}\,h^2$.

Exercise 2.3 (The Gaussian mixture kernel and the choice of frequencies)**.** In Example 2.2, replace Λ by $\Lambda_\tau \overset{\text{def.}}{=} \mathcal{N}(0,\tau^2)$ for some $\tau > 0$, and set $\varphi_x(\omega) \overset{\text{def.}}{=} C_\tau e^{\mathrm{i}\omega x}e^{-\sigma^2\omega^2/2}$ in $\mathbb{H}_\tau \overset{\text{def.}}{=} \mathrm{L}^2(\Lambda_\tau)$, with the real scalar product.

1. Show that $K(x,x') = C_\tau^2(1+2\sigma^2\tau^2)^{-1/2}\exp\big(-\tau^2(x-x')^2/(2(1+2\sigma^2\tau^2))\big)$. Find C_τ such that $K(x,x) = 1$, and check Example 2.2 for $\tau = 1/\sigma$.
2. The kernel is $e^{-(x-x')^2/(2\ell_\tau^2)}$. Compute the width ℓ_τ. Show that $\tau \mapsto \ell_\tau$ is decreasing and find its limit when $\tau \to \infty$. Can the kernel be narrower than the components of the mixture?

3. Let $y_n(\omega) \overset{\text{def.}}{=} \frac{C_\tau}{n} \sum_{k=1}^n e^{\mathrm{i}\omega z_k}$ and $w_n \overset{\text{def.}}{=} y_n - \Phi\mu^0$. Show that $\mathbb{E}\,\|w_n\|_{\mathbb{H}_\tau}^2 \leqslant C_\tau^2/n$. Hint: compute $\mathbb{E}\,|w_n(\omega)|^2$ for each ω, then use Fubini's theorem.

4. Express C_τ^2 in terms of $\sigma\tau$. Describe the trade-off between the width of the kernel and the noise level when τ grows.

Exercise 2.4 (A BLASSO without solution)**.** Let $\mathcal{X} = \mathbb{R}$, $\mathbb{H} = \mathbb{R}$, $\varphi_x \overset{\text{def.}}{=} x^2/(1+x^2)$, $y \overset{\text{def.}}{=} 1$ and $0 < \lambda < 1$.

1. Check that $x \mapsto \varphi_x$ is continuous and bounded. Compute $\Phi\mu$ and $\Phi^\star p$. For which p is $\Phi^\star p \in C_0(\mathbb{R})$?

2. Let $\mu \neq 0$. Show that $|\Phi\mu| < \|\mu\|_{\mathrm{TV}}$.

3. Show that $J_\lambda(\mu) > \lambda - \lambda^2/2$ for all $\mu \in \mathcal{M}(\mathbb{R})$. Hint: minimise $t \mapsto \frac{1}{2}(t-1)^2 + \lambda\,|t|$ on $\mathbb{R}$, and treat $\mu = 0$ apart.

4. Compute $\lim_{n\to\infty} J_\lambda\big((1-\lambda)\delta_n\big)$. Conclude that $\mathcal{P}_\lambda(1)$ has no solution.

5. Which step of the proof of Proposition 2.2 fails? Compare the weak-$\star$ limit of δ_n with the limit of $\Phi\delta_n$.

Exercise 2.5 (The BLASSO for one spike)**.** Let $\mathcal{X} = \mathbb{T}$ or $\mathcal{X} = \mathbb{R}$. Assume the setting of Proposition 2.4 and $K(x,x') = \kappa(x-x')$, with $\kappa(0) = 1$ and $|\kappa(u)| < 1$ for $u \neq 0$. Let $x_0 \in \mathcal{X}$, $a \neq 0$ and $y = a\varphi_{x_0}$.

1. Compute $\Phi^\star y$ and $\|\Phi^\star y\|_\infty$. Show that 0 is a solution of $\mathcal{P}_\lambda(y)$ if and only if $\lambda \geqslant |a|$.

2. Let $0 < \lambda < |a|$ and $\hat{\mu} \overset{\text{def.}}{=} (a - \lambda\,\mathrm{sign}(a))\delta_{x_0}$. Compute $\hat{p}_\lambda$ and $\hat{\eta}_\lambda$, and show that $\hat{\mu}$ is a solution of $\mathcal{P}_\lambda(y)$.

3. Show that $\hat{\mu}$ is the unique solution. Hint: use item (4) of Proposition 2.4, then item (2).

4. Let $\mathcal{X} = \mathbb{R}$ and let $y \in \mathbb{H}$ be arbitrary. Show that every solution of $\mathcal{P}_\lambda(y)$ has a compact support.

Chapter 3

Exact recovery and the construction of certificates

Throughout, $\mu^0 = \sum_{i=1}^s a_i \delta_{x_i}$, $y_0 = \Phi\mu^0$, and the assumption of Proposition 2.1 holds.

As in Lecture 2, $s \geqslant 1$, the points $x_i \in \mathcal{X}$ are pairwise distinct, $a_i \neq 0$, $\operatorname{sign}(a) \overset{\text{def.}}{=} (\operatorname{sign}(a_i))_{i \leqslant s}$ and $\operatorname{supp}(\mu^0) = \{x_1, \dots, x_s\}$ (Remark 2.6). Measures are real and $\langle \cdot, \cdot \rangle_{\mathbb{H}}$ is the real scalar product of Section 2.3. So $\Phi^\star p \in C_0(\mathcal{X})$ for every $p \in \mathbb{H}$, and (2.8) holds.

The lecture answers two questions. When is μ^0 the unique solution of $\mathcal{P}_0(y_0)$? Which certificate should we build, and how? The first answer is that of Lecture 1: a certificate that is strictly below one off the support. For the second one, the certificate that matters at low noise is the minimal norm certificate η_0 (Lecture 4). Section 3.2 gives a candidate for η_0 in closed form, and Section 3.3 shows that this candidate is a non-degenerate certificate when the spikes are separated.

3.1 Exact recovery without noise

The two results of this section are the proof of Proposition 1.1 on measures and the definition of the minimal norm certificate. Both use weak duality only (Remark 2.9).

Definition 3.1 (Source condition)**.** μ^0 satisfies the source condition (SC) if there is $p \in \mathbb{H}$ with $\Phi^\star p \in \partial \|\mu^0\|_{\mathrm{TV}}$, that is $\|\Phi^\star p\|_\infty \leqslant 1$ and $(\Phi^\star p)(x_i) = \operatorname{sign}(a_i)$ for all i.

The two forms are equivalent by Proposition 2.3. In the words of Definition 2.5, (SC) says that μ^0 has a dual certificate. We write

$$C \overset{\text{def.}}{=} \{p \in \mathbb{H} \,;\, \|\Phi^\star p\|_\infty \leqslant 1\}, \qquad P \overset{\text{def.}}{=} \left\{p \in \mathbb{H} \,;\, \Phi^\star p \in \partial \|\mu^0\|_{\mathrm{TV}}\right\},$$

as in Remark 2.1. So (SC) means $P \neq \emptyset$.

Proposition 3.1 (Uniqueness under a strict certificate)**.** *Assume that $(\varphi_{x_i})_{i \leqslant s}$ is linearly independent and that there is $\eta = \Phi^\star p \in \partial \|\mu^0\|_{\mathrm{TV}}$ with $|\eta(x)| < 1$ for all $x \in \mathcal{X} \setminus \operatorname{supp}(\mu^0)$. Then μ^0 is the unique solution of $\mathcal{P}_0(y_0)$.*

Proof. By Proposition 2.3, $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \mu^0 \rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}$. Let $\nu \in \mathcal{M}(\mathcal{X})$ be feasible, that is $\Phi\nu = y_0$. By (2.5) and by (2.8), used twice,

$$\|\nu\|_{\mathrm{TV}} \geqslant \langle \eta, \nu \rangle_{\mathcal{M}} = \langle p, \Phi\nu \rangle_{\mathbb{H}} = \langle p, \Phi\mu^0 \rangle_{\mathbb{H}} = \langle \eta, \mu^0 \rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}. \tag{3.1}$$

So μ^0, which is feasible, is a solution. Let ν be a solution. Then $\|\nu\|_{\mathrm{TV}} = \|\mu^0\|_{\mathrm{TV}}$, and all the terms of (3.1) are equal. In particular $\langle \eta, \nu \rangle_{\mathcal{M}} = \|\nu\|_{\mathrm{TV}}$. Since $\eta \in C_0(\mathcal{X})$ and $\|\eta\|_\infty \leqslant 1$, Remark 2.6 (ii) gives

$$\operatorname{supp}(\nu) \subset \{x \in \mathcal{X} \,;\, |\eta(x)| = 1\} = \{x_1, \dots, x_s\},$$

where the equality holds because $|\eta(x_i)| = 1$ and $|\eta| < 1$ elsewhere. By Remark 2.6 (i), $\nu = \sum_i b_i \delta_{x_i}$ with $b_i \overset{\text{def.}}{=} \nu(\{x_i\})$. Then $0 = \Phi\nu - \Phi\mu^0 = \sum_i (b_i - a_i)\varphi_{x_i}$, and linear independence gives $b = a$, that is $\nu = \mu^0$. □

Remark 3.1 (The hypotheses). (a) The proof does not assume that $\mathcal{P}_0(y_0)$ has a solution: the certificate proves that μ^0 is one. (b) Under (SC) alone, the first half of the proof still shows that μ^0 is a solution. The strict inequality off the support and the linear independence give uniqueness. Neither can be dropped (Exercise 3.1).

The next lemma gives the dual problem of $\mathcal{P}_0(y_0)$ and its solution of minimal norm. It replaces strong duality, which Proposition 2.4 gives for $\lambda > 0$ only.

Lemma 3.1 (Noiseless duality under the source condition)**.** *Let $C = \{p \in \mathbb{H} \,;\, \|\Phi^\star p\|_\infty \leqslant 1\}$ and assume (SC). Then $\sup_{p \in C} \langle p,\, y_0\rangle_{\mathbb{H}} = \|\mu^0\|_{\mathrm{TV}} = \min \mathcal{P}_0(y_0)$. The supremum is attained exactly at the p with $\Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$. Among them there is a unique element p_0 of minimal norm. We set $\eta_0 \overset{\text{def.}}{=} \Phi^\star p_0$, the minimal norm certificate.*

Proof. All norms and scalar products are those of $\mathbb{H}$.

Weak duality. Let $p \in C$ and let ν be feasible for $\mathcal{P}_0(y_0)$. By (2.8) and (2.5),

$$\langle p,\, y_0\rangle = \langle p,\, \Phi\nu\rangle = \langle \Phi^\star p,\, \nu\rangle_{\mathcal{M}} \leqslant \|\nu\|_{\mathrm{TV}}. \tag{3.2}$$

With $\nu = \mu^0$ this gives $\langle p,\, y_0\rangle \leqslant \|\mu^0\|_{\mathrm{TV}}$ for all $p \in C$.

Values. Let $\bar{p} \in P$, given by (SC). Then $\bar{p} \in C$ and $\langle \bar{p},\, y_0\rangle = \langle \Phi^\star \bar{p},\, \mu^0\rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}$. So the supremum equals $\|\mu^0\|_{\mathrm{TV}}$ and it is attained at $\bar{p}$. By (3.2) with $p = \bar{p}$, every feasible ν satisfies $\|\nu\|_{\mathrm{TV}} \geqslant \|\mu^0\|_{\mathrm{TV}}$. So μ^0 is a solution and $\min \mathcal{P}_0(y_0) = \|\mu^0\|_{\mathrm{TV}}$.

Maximisers. Let $p \in C$. By (2.8), $\langle p,\, y_0\rangle = \|\mu^0\|_{\mathrm{TV}}$ if and only if $\langle \Phi^\star p,\, \mu^0\rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}$. Since $\Phi^\star p \in C_0(\mathcal{X})$ and $\|\Phi^\star p\|_\infty \leqslant 1$, Proposition 2.3 says that this holds if and only if $\Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$. Hence the set of maximisers is P, and

$$P = C \cap \left\{p \in \mathbb{H} \,;\, \langle p,\, y_0\rangle = \|\mu^0\|_{\mathrm{TV}}\right\}.$$

The set C is closed and convex (proof of Proposition 2.4), and so is the affine hyperplane. So P is closed, convex and nonempty.

Minimal norm. Let $m \overset{\text{def.}}{=} \inf_{p \in P} \|p\|$. For $p, q \in P$, the midpoint $(p+q)/2$ lies in P, so $\|(p+q)/2\| \geqslant m$, and the parallelogram identity gives

$$\|p - q\|^2 = 2\,\|p\|^2 + 2\,\|q\|^2 - 4\left\|\tfrac{p+q}{2}\right\|^2 \leqslant 2\,\|p\|^2 + 2\,\|q\|^2 - 4m^2. \tag{3.3}$$

Let (p_n) be a sequence of P with $\|p_n\| \to m$. By (3.3), $\|p_n - p_k\|^2 \to 0$ when $n, k \to \infty$, so (p_n) is a Cauchy sequence. Since $\mathbb{H}$ is complete, it converges to some p_0, which lies in P because P is closed, and $\|p_0\| = m$. If $q \in P$ also has $\|q\| = m$, then (3.3) gives $\|p_0 - q\|^2 \leqslant 0$, so $q = p_0$. □

Remark 3.2 (On the lemma). (a) On a grid, p_0 is the vector p_0 of Exercise 1.4, and under the assumption of Corollary 2.1 it is p_F (Remark 2.1). (b) Lecture 4 uses three facts from the lemma: $p_0 \in C$, p_0 maximises $\langle \cdot,\, y_0\rangle_{\mathbb{H}}$ on C, and p_0 has minimal norm among the maximisers. The dual solutions $\hat{p}_\lambda(y_0)$ of the BLASSO converge to p_0 when $\lambda \to 0$ (Proposition 4.2). This is why η_0 is the certificate that governs the low noise regime. (c) Without (SC) the lemma says nothing, and η_0 is not defined. For a general kernel the dual problem may then have no solution; we admit this.

3.2 The vanishing derivative pre-certificate

From now on $\mathcal{X} = \mathbb{T}$ or $\mathcal{X} = \mathbb{R}$ and $x \mapsto \varphi_x$ is $\mathscr{C}^2$. For $x = (x_1, \dots, x_s)$ let $\Phi_x b \overset{\text{def.}}{=} \sum_i b_i \varphi_{x_i}$, $\Phi'_x b \overset{\text{def.}}{=} \sum_i b_i \partial_x \varphi_{x_i}$ and $\Gamma_x \overset{\text{def.}}{=} (\Phi_x, \Phi'_x) : \mathbb{R}^s \times \mathbb{R}^s \to \mathbb{H}$.

Here $\Gamma_x(b, c) = \Phi_x b + \Phi'_x c$, and $\mathbb{R}^s \times \mathbb{R}^s = \mathbb{R}^{2s}$ carries the Euclidean scalar product. As in Lecture 2, $x_0 \overset{\text{def.}}{=} (x_1, \dots, x_s)$ denotes the positions of μ^0. By Remark 2.4, every $\Phi^\star p$ is $\mathscr{C}^2$, and $\langle \Gamma_x(b, c),\, p\rangle_{\mathbb{H}} = \sum_i b_i \langle \varphi_{x_i},\, p\rangle_{\mathbb{H}} + \sum_i c_i \langle \partial_x \varphi_{x_i},\, p\rangle_{\mathbb{H}}$. So

$$\Gamma_x^\star p = \Big(\big((\Phi^\star p)(x_i)\big)_{i \leqslant s}, \big((\Phi^\star p)'(x_i)\big)_{i \leqslant s} \Big) \in \mathbb{R}^s \times \mathbb{R}^s. \tag{3.4}$$

The matrix $\Gamma_x^\star \Gamma_x$ is the $2s \times 2s$ matrix of the scalar products of the vectors φ_{x_i} and $\partial_x \varphi_{x_i}$. With the notation $K^{(ij)}(x, x') = \langle \partial_x^i \varphi_x,\, \partial_{x'}^j \varphi_{x'} \rangle_{\mathbb{H}}$ of Remark 2.4, its blocks are

$$\Gamma_x^\star \Gamma_x = \begin{pmatrix} \big(K(x_i, x_j)\big)_{i,j} & \big(K^{(01)}(x_i, x_j)\big)_{i,j} \\ \big(K^{(10)}(x_i, x_j)\big)_{i,j} & \big(K^{(11)}(x_i, x_j)\big)_{i,j} \end{pmatrix}. \tag{3.5}$$

Remark 3.3 (Certificates are flat at the spikes). Let $\eta \in \mathscr{C}^2(\mathcal{X})$ with $\|\eta\|_\infty \leqslant 1$, and let $x \in \mathcal{X}$ with $\eta(x) = \pm 1$. Set $\psi \overset{\text{def.}}{=} \eta(x)\, \eta$. Then $\psi \leqslant 1 = \psi(x)$. On $\mathbb{R}$, and on $\mathbb{T}$ seen as 1-periodic functions on $\mathbb{R}$, the point x is an interior maximum, so $\psi'(x) = 0$, as in item 1 of Lemma 2.1. Then the Taylor-Young formula $\psi(x + v) = 1 + \frac{1}{2}\psi''(x) v^2 + o(v^2) \leqslant 1$ gives $\psi''(x) \leqslant 0$. Hence every $\eta = \Phi^\star p$ with $p \in P$ satisfies

$$\eta(x_i) = \operatorname{sign}(a_i), \qquad \eta'(x_i) = 0, \qquad \operatorname{sign}(a_i)\, \eta''(x_i) \leqslant 0, \qquad i = 1, \dots, s.$$

By (3.4), the first two conditions read $\Gamma_{x_0}^\star p = (\operatorname{sign}(a), 0_s)$.

Definition 3.2 (Vanishing derivative pre-certificate)**.** $\eta_V \overset{\text{def.}}{=} \Phi^\star p_V$ where p_V has minimal norm among the $p \in \mathbb{H}$ such that $(\Phi^\star p)(x_i) = \operatorname{sign}(a_i)$ and $(\Phi^\star p)'(x_i) = 0$ for all i, when such p exist.

The constraint set is $A \overset{\text{def.}}{=} \{p \in \mathbb{H}\,;\, \Gamma_{x_0}^\star p = (\operatorname{sign}(a), 0_s)\}$. It is closed and convex, as the inverse image of a point by a continuous linear map. When it is nonempty, the minimal norm step of the proof of Lemma 3.1 shows that it has a unique element of minimal norm. The name says that η_V interpolates the signs with zero slope. The prefix "pre" says that nothing guarantees $\|\eta_V\|_\infty \leqslant 1$.

Proposition 3.2 (Closed form)**.** *If Γ_{x_0} is injective, then* $p_V = \Gamma_{x_0} (\Gamma_{x_0}^\star \Gamma_{x_0})^{-1} \begin{pmatrix} \operatorname{sign}(a) \\ 0_s \end{pmatrix}$.

Proof. Write $\Gamma \overset{\text{def.}}{=} \Gamma_{x_0}$ and $u \overset{\text{def.}}{=} (\operatorname{sign}(a), 0_s) \in \mathbb{R}^{2s}$. If $\Gamma^\star \Gamma c = 0$, then $\|\Gamma c\|_{\mathbb{H}}^2 = \langle c,\, \Gamma^\star \Gamma c \rangle_{\mathbb{R}^{2s}} = 0$, so $\Gamma c = 0$ and $c = 0$. Hence the $2s \times 2s$ matrix $\Gamma^\star \Gamma$ is invertible, and $q \overset{\text{def.}}{=} \Gamma (\Gamma^\star \Gamma)^{-1} u$ is well defined. It satisfies $\Gamma^\star q = u$, so $q \in A$ by (3.4). Let $p \in A$. Then $\Gamma^\star (p - q) = 0$, so $p - q$ is orthogonal to $\operatorname{Im} \Gamma$, which contains q. By Pythagoras, $\|p\|_{\mathbb{H}}^2 = \|q\|_{\mathbb{H}}^2 + \|p - q\|_{\mathbb{H}}^2$. So q is the unique element of minimal norm of A, that is $q = p_V$. □

This is the check of Definition 2.1 with Γ_{x_0} in place of Φ_{x_0}. The Fuchs pre-certificate only asks $\Phi_{x_0}^\star p = \operatorname{sign}(a)$. The pre-certificate η_V also asks the zero slope that every certificate has (Remark 3.3).

Remark 3.4 (Consequences of the closed form). Assume that $\Gamma \overset{\text{def.}}{=} \Gamma_{x_0}$ is injective, and let $u \overset{\text{def.}}{=} (\operatorname{sign}(a), 0_s)$. (a) Let $\alpha = (\alpha_1, \alpha_2) \overset{\text{def.}}{=} (\Gamma^\star \Gamma)^{-1} u \in \mathbb{R}^s \times \mathbb{R}^s$. Then $p_V = \Gamma \alpha = \sum_j \alpha_{1,j} \varphi_{x_j} + \sum_j \alpha_{2,j} \partial_x \varphi_{x_j}$. Since $\Phi^\star \varphi_{x_j} = K(x_j, \cdot)$ and $\Phi^\star \partial_x \varphi_{x_j} = \langle \partial_x \varphi_{x_j},\, \varphi_\cdot \rangle_{\mathbb{H}} = K^{(10)}(x_j, \cdot)$,

$$\eta_V = \sum_{j=1}^s \alpha_{1,j} K(x_j, \cdot) + \sum_{j=1}^s \alpha_{2,j} K^{(10)}(x_j, \cdot). \tag{3.6}$$

So η_V is computed from the kernel and its derivatives at the spikes only. (b) $\|p_V\|_{\mathbb{H}}^2 = \langle \Gamma\alpha, \Gamma\alpha\rangle_{\mathbb{H}} = \langle \alpha, \Gamma^\star\Gamma\alpha\rangle = \langle \alpha, u\rangle = \langle \alpha_1, \operatorname{sign}(a)\rangle_{\mathbb{R}^s}$. (c) The map $\Gamma^\star : \mathbb{H} \to \mathbb{R}^{2s}$ is onto: for $e \in \mathbb{R}^{2s}$, $q_e \overset{\text{def.}}{=} \Gamma(\Gamma^\star\Gamma)^{-1}e$ satisfies $\Gamma^\star q_e = e$. With e a vector of the canonical basis, this gives functions $\Phi^\star q_e$ that interpolate one value or one slope at one spike and vanish with zero slope at the others. Appendix A uses them (Exercise 3.4). (d) Φ_{x_0} is injective, since $\Phi_{x_0} b = \Gamma(b, 0)$.

Proposition 3.3 (η_V and the minimal norm certificate)**.** *Assume that Γ_{x_0} is injective. Then $\|\eta_V\|_\infty \leqslant 1$ if and only if (SC) holds and $\eta_V = \eta_0$.*

Proof. By Proposition 3.2, p_V exists. Assume $\|\eta_V\|_\infty \leqslant 1$. Since $\eta_V(x_i) = \operatorname{sign}(a_i)$ for all i and $\eta_V \in C_0(\mathcal{X})$, Proposition 2.3 gives $\eta_V \in \partial\|\mu^0\|_{\mathrm{TV}}$. So (SC) holds and $p_V \in P$. By Lemma 3.1, p_0 exists, and $p_0 \in P$. By Remark 3.3, $p_0 \in A$, so $\|p_V\|_{\mathbb{H}} \leqslant \|p_0\|_{\mathbb{H}}$. Since $p_V \in P$ and p_0 is the unique element of minimal norm of P, this gives $p_V = p_0$, hence $\eta_V = \eta_0$.

Conversely, if (SC) holds and $\eta_V = \eta_0$, then $\|\eta_V\|_\infty = \|\eta_0\|_\infty \leqslant 1$, since $p_0 \in P$. □

The interior position of the x_i matters. On an interval $[0, 1]$, a spike at 0 can have a certificate with a nonzero derivative at 0. When $\|\eta_V\|_\infty > 1$, either (SC) fails or $\eta_0 \neq \eta_V$. The non-degeneracy that Lecture 4 needs is a property of η_0.

Definition 3.3 (Non-degenerate source condition)**.** μ^0 satisfies (NDSC) if (SC) holds, $|\eta_0(x)| < 1$ for all $x \notin \{x_1, \dots, x_s\}$, and $\operatorname{sign}(a_i)\,\eta_0''(x_i) < 0$ for all i.

By Remark 3.3, every certificate satisfies the two conditions of (NDSC) with non-strict inequalities. (NDSC) asks them strictly. With the linear independence of the φ_{x_i}, it gives exact recovery by Proposition 3.1. The next remark turns (NDSC) into a computation.

Remark 3.5 (Checking (NDSC) with η_V). Assume that Γ_{x_0} is injective. Then μ^0 satisfies (NDSC) if and only if

$$|\eta_V(x)| < 1 \text{ for all } x \notin \{x_1, \dots, x_s\} \quad \text{and} \quad \operatorname{sign}(a_i)\,\eta_V''(x_i) < 0 \text{ for all } i. \tag{3.7}$$

The recipe is: compute η_V with (3.6) and check these two conditions. In particular, if $\|\eta_V\|_\infty > 1$, then (NDSC) fails.

Proof. If (3.7) holds, then $\|\eta_V\|_\infty \leqslant 1$, since $|\eta_V(x_i)| = 1$, and Proposition 3.3 gives (SC) and $\eta_0 = \eta_V$. So (NDSC) holds.

Conversely, assume (NDSC). It is enough to show $p_V = p_0$. Let $\sigma_i \overset{\text{def.}}{=} \operatorname{sign}(a_i)$, $q \overset{\text{def.}}{=} p_V - p_0$ and $g \overset{\text{def.}}{=} \Phi^\star q$, and assume $q \neq 0$. By Remark 3.3, $p_0 \in A$, so $\Gamma_{x_0}^\star q = 0$, that is $g(x_i) = g'(x_i) = 0$, and $\|p_V\|_{\mathbb{H}} \leqslant \|p_0\|_{\mathbb{H}}$. For $t \in [0, 1]$ let $p_t \overset{\text{def.}}{=} p_0 + tq$ and $\eta_t \overset{\text{def.}}{=} \Phi^\star p_t = \eta_0 + tg$.

Let $k \overset{\text{def.}}{=} \frac{1}{2}\min_i(-\sigma_i\eta_0''(x_i)) > 0$. Since η_0 and η_0'' are continuous, there is $\varepsilon > 0$, smaller than half the distance between two spikes and than $\frac{1}{4}$ on $\mathbb{T}$, such that on each interval $I_i \overset{\text{def.}}{=} [x_i - \varepsilon, x_i + \varepsilon]$, $\sigma_i\eta_0'' \leqslant -k$ and $\sigma_i\eta_0 \geqslant \frac{1}{2}$. Let U be the union of the open intervals $(x_i - \varepsilon, x_i + \varepsilon)$. We claim that $r \overset{\text{def.}}{=} 1 - \sup_{x\notin U} |\eta_0(x)| > 0$. On $\mathbb{T}$, $\mathbb{T} \setminus U$ is compact and $|\eta_0| < 1$ there, so the supremum is a maximum and it is less than 1. On $\mathbb{R}$, the set $L \overset{\text{def.}}{=} \left\{x\ ;\ |\eta_0(x)| \geqslant \frac{1}{2}\right\}$ is compact since $\eta_0 \in C_0(\mathbb{R})$. The maximum of $|\eta_0|$ on the compact set $L \setminus U$ is less than 1, and $|\eta_0| < \frac{1}{2}$ off L.

Let $B \overset{\text{def.}}{=} \sup_x \|\varphi_x\|_{\mathbb{H}}$, so that $\|g\|_\infty \leqslant B\,\|q\|_{\mathbb{H}}$, and let $M \overset{\text{def.}}{=} \max_i \max_{I_i} |g''|$, which is finite since g'' is continuous. Choose $t \in (0, 1)$ with $tM \leqslant k$, $tB\,\|q\|_{\mathbb{H}} \leqslant r$ and $tB\,\|q\|_{\mathbb{H}} \leqslant \frac{1}{2}$. Off U, $|\eta_t| \leqslant 1 - r + tB\,\|q\|_{\mathbb{H}} \leqslant 1$. On I_i, the function $\psi \overset{\text{def.}}{=} \sigma_i\eta_t$ satisfies $\psi(x_i) = 1$, $\psi'(x_i) = 0$ and $\psi'' = \sigma_i\eta_0'' + t\sigma_i g'' \leqslant -k + tM \leqslant 0$. The Taylor formula with integral remainder, $\psi(x_i + v) = 1 + v^2\int_0^1 (1-\theta)\psi''(x_i + \theta v)\,\mathrm{d}\theta$ for $|v| \leqslant \varepsilon$, gives $\psi \leqslant 1$ on I_i. Also $\psi \geqslant \sigma_i\eta_0 - t\,\|g\|_\infty \geqslant \frac{1}{2} - \frac{1}{2} = 0$ on I_i. So $|\eta_t| \leqslant 1$ on I_i. Hence $\|\eta_t\|_\infty \leqslant 1$ and $\eta_t(x_i) = \sigma_i$, that is $p_t \in P$.

Last, $\|p_0 + q\|_{\mathbb{H}}^2 = \|p_V\|_{\mathbb{H}}^2 \leqslant \|p_0\|_{\mathbb{H}}^2$ gives $2\langle p_0, q\rangle_{\mathbb{H}} \leqslant -\|q\|_{\mathbb{H}}^2$, so

$$\|p_t\|_{\mathbb{H}}^2 = \|p_0\|_{\mathbb{H}}^2 + 2t\langle p_0, q\rangle_{\mathbb{H}} + t^2\|q\|_{\mathbb{H}}^2 \leqslant \|p_0\|_{\mathbb{H}}^2 - t(1-t)\|q\|_{\mathbb{H}}^2 < \|p_0\|_{\mathbb{H}}^2.$$

This contradicts the minimality of p_0 in P (Lemma 3.1). So $q = 0$. □

Example 3.1 (One spike). Let $s = 1$ and $K(x, x') = \kappa(x - x')$ with κ even, $\kappa(0) = 1$, $\kappa''(0) < 0$. Then $\eta_V = \operatorname{sign}(a_1)\, K(x_1, \cdot)$, and (NDSC) holds if and only if $|\kappa(u)| < 1$ for $u \neq 0$. This covers Examples 2.1 and 2.2.

Here $u \neq 0$ means $u \in \mathbb{T} \setminus \{0\}$ when $\mathcal{X} = \mathbb{T}$. Let $\sigma \overset{\text{def.}}{=} \operatorname{sign}(a_1)$ and $\mathfrak{g} \overset{\text{def.}}{=} -\kappa''(0) > 0$. By Remark 2.4, $K^{(ij)}(x, x') = \partial_x^i \partial_{x'}^j \kappa(x - x') = (-1)^j \kappa^{(i+j)}(x - x')$. Since κ is even and differentiable, $\kappa'(0) = 0$. By (3.5),

$$\Gamma_{x_0}^\star \Gamma_{x_0} = \begin{pmatrix} \kappa(0) & -\kappa'(0) \\ \kappa'(0) & -\kappa''(0) \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & \mathfrak{g} \end{pmatrix}.$$

This matrix is invertible, so Γ_{x_0} is injective (if $\Gamma_{x_0} c = 0$ then $\Gamma_{x_0}^\star \Gamma_{x_0} c = 0$). Proposition 3.2 gives $p_V = \Gamma_{x_0}(\sigma, 0) = \sigma \varphi_{x_1}$, hence $\eta_V(x) = \sigma \langle \varphi_{x_1}, \varphi_x \rangle_{\mathbb{H}} = \sigma K(x_1, x) = \sigma \kappa(x - x_1)$, since κ is even.

By the Cauchy-Schwarz inequality, $|\kappa(x - x')| \leqslant \|\varphi_x\|_{\mathbb{H}} \|\varphi_{x'}\|_{\mathbb{H}} = \kappa(0) = 1$. So $\|\eta_V\|_\infty \leqslant 1$ for every such kernel, and Proposition 3.3 gives (SC) and $\eta_0 = \eta_V$. The curvature condition always holds: $\sigma \eta_0''(x_1) = \kappa''(0) < 0$. The remaining condition, $|\eta_0(x)| < 1$ for $x \neq x_1$, reads $|\kappa(u)| < 1$ for $u \neq 0$, with $u = x - x_1$.

Weighted Fourier measurements (Example 2.1). Here $\kappa(u) = \sum_{|\omega| \leqslant f_c} \Lambda(\omega) \cos(2\pi\omega u)$, with $\Lambda > 0$ symmetric of sum one and $f_c \geqslant 1$. So κ is even, $\kappa(0) = 1$ and $\kappa''(0) = -4\pi^2 \sum_\omega \omega^2 \Lambda(\omega) < 0$. All the terms below are nonnegative. Keeping the terms $\omega = \pm 1$, then the term $\omega = 0$,

$$1 - \kappa(u) = \sum_\omega \Lambda(\omega)(1 - \cos(2\pi\omega u)) \geqslant 2\Lambda(1)(1 - \cos(2\pi u)), \qquad 1 + \kappa(u) \geqslant 2\Lambda(0).$$

For $u \in \mathbb{T} \setminus \{0\}$, $\cos(2\pi u) < 1$. So $|\kappa(u)| < 1$ and (NDSC) holds. With the Jackson weights, $\eta_0(x) = \operatorname{sign}(a_1) J_{f_c}(x - x_1)$ and $\operatorname{sign}(a_1)\eta_0''(x_1) = -\frac{\pi^2}{3} f_c(f_c + 4)$ (Exercise 2.1): the curvature grows like f_c^2.

Gaussian mixture (Example 2.2). Here $\kappa(u) = e^{-u^2/(6\sigma^2)}$ is even, $\kappa(0) = 1$, $\kappa''(0) = -1/(3\sigma^2) < 0$ and $0 < \kappa(u) < 1$ for $u \neq 0$. So $\eta_0(x) = \operatorname{sign}(a_1) e^{-(x - x_1)^2/(6\sigma^2)}$ and (NDSC) holds. The map $x \mapsto \varphi_x$ is $\mathscr{C}^2$, as the section requires. Indeed, for $k \leqslant 2$ the map $x \mapsto (\mathrm{i}\omega)^k \varphi_x$ is continuous from $\mathbb{R}$ to $\mathbb{H}$ by dominated convergence, since $\left|\omega^k \varphi_x(\omega)\right|^2 \leqslant C^2 \omega^{2k}$ is integrable for Λ. For $k \leqslant 1$, the bound $\left|e^{\mathrm{i}\theta} - 1 - \mathrm{i}\theta\right| \leqslant \theta^2/2$ gives

$$\left\| \frac{(\mathrm{i}\omega)^k (\varphi_{x+t} - \varphi_x)}{t} - (\mathrm{i}\omega)^{k+1} \varphi_x \right\|_{\mathbb{H}}^2 \leqslant \frac{C^2 t^2}{4} \int_{\mathbb{R}} \omega^{2k+4} \, \mathrm{d}\Lambda(\omega) \xrightarrow[t \to 0]{} 0.$$

So $\partial_x \varphi_x = \mathrm{i}\omega \varphi_x$ and $\partial_x^2 \varphi_x = -\omega^2 \varphi_x$.

For several spikes, η_V is no longer a translate of κ, and the recipe can fail.

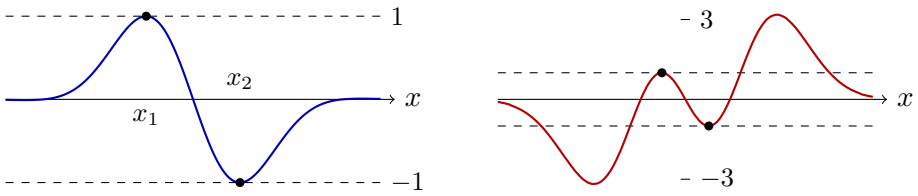


Figure 3.1: The pre-certificate η_V for $\kappa(u) = e^{-u^2/2}$ (Example 2.2 with $\sigma^2 = 1/3$), $a_1 > 0 > a_2$ and $x_2 - x_1 = d$, computed with (3.6). Left: $d = 2$. Numerically, $|\eta_V| < 1$ off $\{x_1, x_2\}$ and η_V curves at both spikes, so (NDSC) holds. Right: $d = 1$, with a compressed vertical scale. Then $\|\eta_V\|_\infty \approx 3.18$, and (NDSC) fails by Remark 3.5. The coefficients are those of Exercise 3.2.

3.3 Local positive curvature gives a certificate

Example 3.1 gives (NDSC) for one spike. For several spikes we need a quantitative argument. If the spikes are far apart, the matrix $\Gamma_{x_0}^\star \Gamma_{x_0}$ is close to a diagonal matrix, and η_V is close to

$\sum_i \operatorname{sign}(a_i) K(x_i, \cdot)$, a sum of one-spike certificates. To make this precise, the kernel must stay below 1 away from the diagonal and curve near it, and its tails must be small. This section follows [16], in dimension one.

Setting. In this section $\mathcal{X} = \mathbb{T}$ or $\mathcal{X} = \mathbb{R}$, $x \mapsto \varphi_x$ is $\mathscr{C}^2$, and the kernel of Φ is $K(x, x') = \kappa(x - x')$ with $\kappa(0) = 1$ and

$$\mathfrak{g} \overset{\text{def.}}{=} -\kappa''(0) = K^{(11)}(x, x) = \|\partial_x \varphi_x\|_{\mathbb{H}}^2 > 0.$$

The function κ is even, since K is symmetric, so $\kappa'(0) = 0$. We use the distance $\mathfrak{d}_{\mathfrak{g}}(x, x') \overset{\text{def.}}{=} \sqrt{\mathfrak{g}}\,|x - x'|$, where $|x - x'|$ is the distance on $\mathbb{T}$ when $\mathcal{X} = \mathbb{T}$. By Remark 2.4, $K^{(ij)}(x, x') = (-1)^j \kappa^{(i+j)}(x - x')$ for $i, j \leqslant 2$. A number q that is a derivative of total order k, such as $K^{(ij)}(x, x')$ with $i + j = k$ or $\eta^{(k)}(x)$, is measured in the metric: $\|q\| \overset{\text{def.}}{=} \mathfrak{g}^{-k/2}\,|q|$. So $\left\|K^{(ij)}(x, x')\right\| = \mathfrak{g}^{-(i+j)/2}\left|\kappa^{(i+j)}(x - x')\right|$ depends only on $i + j$ and on $|x - x'|$. In Example 2.2, $\mathfrak{g} = 1/(3\sigma^2)$ and $K(x, x') = e^{-\mathfrak{d}_{\mathfrak{g}}(x,x')^2/2}$: in the metric, the kernel is the standard Gaussian.

Definition 3.4 (Local curvature constants)**.** For $r > 0$,

$$\varepsilon_0(r) \overset{\text{def.}}{=} \sup\left\{\varepsilon\ ;\ K(x, x') \leqslant 1 - \varepsilon \text{ whenever } \mathfrak{d}_{\mathfrak{g}}(x, x') \geqslant r\right\},$$
$$\varepsilon_2(r) \overset{\text{def.}}{=} \sup\left\{\varepsilon\ ;\ -K^{(02)}(x', x) \geqslant \varepsilon\,\mathfrak{g} \text{ whenever } \mathfrak{d}_{\mathfrak{g}}(x, x') < r\right\}.$$

Since $K^{(02)}(x', x) = \kappa''(x' - x)$ and κ'' is even,

$$\varepsilon_0(r) = 1 - \sup\left\{\kappa(v)\ ;\ \sqrt{\mathfrak{g}}\,|v| \geqslant r\right\}, \qquad \varepsilon_2(r) = \inf\left\{-\kappa''(v)/\mathfrak{g}\ ;\ \sqrt{\mathfrak{g}}\,|v| < r\right\}.$$

A large ε_0 means that the kernel is far below 1 outside the ball of radius r. A large ε_2 means that it is strictly concave inside. These are plain suprema, as in [16]. Another convention in the literature divides them by 2 and 4 [17].

Definition 3.5 (Kernel width)**.** For $h > 0$ and $s \in \mathbb{N}$, $\Delta(h, s)$ is the infimum of the Δ such that $\sum_{k \geqslant 2} \left\|K^{(ij)}(x_1, x_k)\right\| \leqslant h$ for all $(i, j) \in \{0, 1, 2\} \times \{0, 1\}$ and all families of at most s points with $\mathfrak{d}_{\mathfrak{g}}(x_k, x_\ell) \geqslant \Delta$ for $k \neq \ell$.

If the property holds for some Δ, it holds for every $\Delta_2 \geqslant \Delta$, since a Δ_2-separated family is Δ-separated. So it holds for every family of at most s points whose separations are all larger than $\Delta(h, s)$. Here the six pairs (i, j) give four conditions, one for each order $i + j \in \{0, 1, 2, 3\}$, and any point of the family can be listed first.

Assumption 3.1 (Kernel assumption)**.** For $i, j \leqslant 2$, $B_{ij} \overset{\text{def.}}{=} \sup_{x,x'} \left\|K^{(ij)}(x, x')\right\| < \infty$, and $B_i \overset{\text{def.}}{=} B_{0i} + B_{1i} + 1$. There are $r_{\text{near}} < B_{02}^{-1/2}$ with $\varepsilon_0 \overset{\text{def.}}{=} \varepsilon_0(r_{\text{near}}) > 0$ and $\varepsilon_2 \overset{\text{def.}}{=} \varepsilon_2(r_{\text{near}}) > 0$, an integer s, and $h \leqslant \frac{1}{64}\min(\varepsilon_0/B_0, \varepsilon_2/B_2)$ with $\Delta(h, s) < \infty$. The admissible separation is $\Delta_{\text{adm}} \overset{\text{def.}}{=} 2\Delta(h, s)$.

Remark 3.6 (Constants that are always known). By the Cauchy-Schwarz inequality in $\mathbb{H}$, $|K(x, x')| \leqslant \|\varphi_x\|_{\mathbb{H}}\,\|\varphi_{x'}\|_{\mathbb{H}} = 1$, $\left|K^{(10)}(x, x')\right| \leqslant \sqrt{\mathfrak{g}}$ and $\left|K^{(11)}(x, x')\right| \leqslant \mathfrak{g}$, with equality in the first and last ones at $x = x'$. Since $\left|K^{(02)}\right| = |\kappa''| = \left|K^{(11)}\right|$, we get $B_{00} = B_{02} = 1$ and $B_{10} \leqslant 1$. So the condition $r_{\text{near}} < B_{02}^{-1/2}$ reads $r_{\text{near}} < 1$. Taking $x = x'$ in the definition of $\varepsilon_2(r)$ gives $\varepsilon_2(r) \leqslant 1$. Hence $B_2 \geqslant 2$ and $h \leqslant \varepsilon_2/(64B_2) \leqslant 1/128$.

Definition 3.6 (Near and far regions, non-degenerate certificate)**.** Let $\mathcal{X}_i^{\text{near}} \overset{\text{def.}}{=} \{x\ ;\ \mathfrak{d}_{\mathfrak{g}}(x, x_i) < r_{\text{near}}\}$ and $\mathcal{X}^{\text{far}} \overset{\text{def.}}{=} \mathcal{X} \setminus \bigcup_i \mathcal{X}_i^{\text{near}}$. A function $\eta \in \operatorname{Im}(\Phi^\star)$ is an $(\varepsilon_0, \varepsilon_2, r_{\text{near}})$-non-degenerate certificate if $\eta(x_i) = \operatorname{sign}(a_i)$ for all i, $|\eta| \leqslant 1 - \varepsilon_0$ on $\mathcal{X}^{\text{far}}$, and $|\eta(x)| \leqslant 1 - \varepsilon_2\,\mathfrak{d}_{\mathfrak{g}}(x, x_i)^2$ on $\mathcal{X}_i^{\text{near}}$.

Remark 3.7 (On non-degenerate certificates). (a) If $\mathfrak{d}_{\mathfrak{g}}(x_j, x_k) \geqslant 2r_{\text{near}}$ for $j \neq k$, the near regions are pairwise disjoint: a point x in $\mathcal{X}_j^{\text{near}} \cap \mathcal{X}_k^{\text{near}}$ would give $\mathfrak{d}_{\mathfrak{g}}(x_j, x_k) < 2r_{\text{near}}$. This is the assumption of item 2 of Proposition 4.1. (b) A non-degenerate certificate with $\varepsilon_0, \varepsilon_2 > 0$ satisfies $\|\eta\|_\infty \leqslant 1$, with $|\eta(x)| = 1$ only at the spikes, since $\mathcal{X} = \mathcal{X}^{\text{far}} \cup \bigcup_i \mathcal{X}_i^{\text{near}}$. So it lies in $\partial\|\mu^0\|_{\text{TV}}$, and it is strict in the sense of Proposition 3.1, with explicit margins. Lecture 4 turns these margins into error bounds.

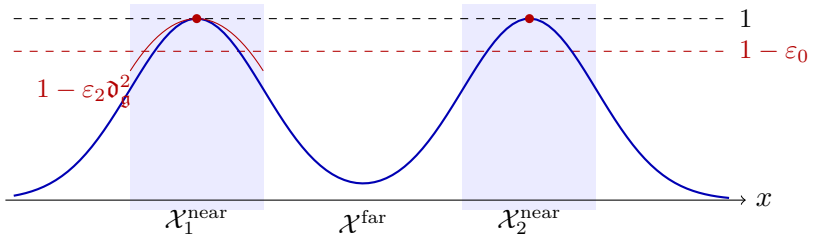


Figure 3.2: A schematic non-degenerate certificate for two positive spikes. In the near regions (shaded), $|\eta|$ stays below the parabola $1 - \varepsilon_2\mathfrak{d}_{\mathfrak{g}}(x, x_i)^2$. In the far region it stays below $1 - \varepsilon_0$.

The next lemma turns curvature into the quadratic decay of Definition 3.6.

Lemma 3.2 (Curvature gives quadratic decay)**.** *Let Assumption 3.1 hold, and let $\eta : \mathcal{X} \to \mathbb{R}$ be $\mathscr{C}^2$. Let $a \in \{-1, 1\}$, $0 \leqslant \delta < \varepsilon_2/2$ and $0 < r \leqslant r_{\text{near}}$. Assume $\eta(x_i) = a$, $\eta'(x_i) = 0$ and $\left\|a\,\eta''(x) - K^{(02)}(x_i, x)\right\| \leqslant \delta$ for $\mathfrak{d}_{\mathfrak{g}}(x, x_i) < r$. Then $|\eta(x)| \leqslant 1 - \frac{\varepsilon_2 - 2\delta}{2}\,\mathfrak{d}_{\mathfrak{g}}(x, x_i)^2$ on this ball.*

Proof. Let x with $L \overset{\text{def.}}{=} \mathfrak{d}_{\mathfrak{g}}(x, x_i) < r$. Write $x = x_i + v$ with $v \in \mathbb{R}$ and $|v| = L/\sqrt{\mathfrak{g}}$. On $\mathbb{T}$, v is the representative of $x - x_i$ of smallest absolute value, and we see η as a 1-periodic function on $\mathbb{R}$. For $t \in [0, 1]$, the point $z_t \overset{\text{def.}}{=} x_i + tv$ satisfies $\mathfrak{d}_{\mathfrak{g}}(z_t, x_i) \leqslant tL < r$. On the ball, the hypothesis on η'' reads $|a\eta''(z) - \kappa''(x_i - z)| \leqslant \delta\mathfrak{g}$, and, since $r \leqslant r_{\text{near}}$, the definitions of ε_2 and B_{02} give $-B_{02}\mathfrak{g} \leqslant \kappa''(x_i - z) \leqslant -\varepsilon_2\mathfrak{g}$. Hence

$$-(B_{02} + \delta)\mathfrak{g} \leqslant a\eta''(z_t) \leqslant -(\varepsilon_2 - \delta)\mathfrak{g}, \qquad t \in [0, 1]. \tag{3.8}$$

The Taylor formula with integral remainder, with $a\eta(x_i) = a^2 = 1$ and $\eta'(x_i) = 0$, gives

$$a\eta(x) = 1 + v^2 \int_0^1 (1 - t)\, a\eta''(z_t)\, \mathrm{d}t.$$

Since $\int_0^1 (1 - t)\,\mathrm{d}t = \frac{1}{2}$ and $\mathfrak{g}v^2 = L^2$, (3.8) gives

$$1 - \tfrac{B_{02}+\delta}{2}L^2 \leqslant a\eta(x) \leqslant 1 - \tfrac{\varepsilon_2 - \delta}{2}L^2.$$

By Remark 3.6, $B_{02} = 1$ and $\varepsilon_2 \leqslant 1$, so $\delta < \frac{1}{2}$ and $L^2 < {r_{\text{near}}}^2 < 1$. So $\frac{B_{02}+\delta}{2}L^2 < \frac{3}{4}$ and $a\eta(x) > 0$. Hence, with $\delta \geqslant 0$, $|\eta(x)| = a\eta(x) \leqslant 1 - \frac{\varepsilon_2 - \delta}{2}L^2 \leqslant 1 - \frac{\varepsilon_2 - 2\delta}{2}L^2$. □

For real certificates in dimension one the proof gives the better constant $(\varepsilon_2 - \delta)/2$. The lower bound on $a\eta$ is where the condition $r_{\text{near}} < B_{02}^{-1/2}$ enters: without it, η could cross $-a$.

Now we build the certificate. We look for η_V in the form (3.6) and solve the interpolation system directly.

Definition 3.7 (Certificate Ansatz)**.** For $\alpha_1, \alpha_2 \in \mathbb{R}^s$, $\eta(x) \overset{\text{def.}}{=} \sum_j \alpha_{1,j} K(x_j, x) + \sum_j \alpha_{2,j} K^{(10)}(x_j, x)$. The coefficients solve the linear system $\Upsilon\alpha = (\operatorname{sign}(a), 0_s)$ that expresses $\eta(x_i) = \operatorname{sign}(a_i)$ and $\eta'(x_i) = 0$.

Explicitly, since $\partial_x K(x_j,x) = K^{(01)}(x_j,x)$ and $\partial_x K^{(10)}(x_j,x) = K^{(11)}(x_j,x)$, the matrix of the system is, with i the row index and j the column index,

$$\Upsilon \overset{\text{def.}}{=} \begin{pmatrix} \big(K(x_j,x_i)\big)_{i,j} & \big(K^{(10)}(x_j,x_i)\big)_{i,j} \\ \big(K^{(01)}(x_j,x_i)\big)_{i,j} & \big(K^{(11)}(x_j,x_i)\big)_{i,j} \end{pmatrix} \in \mathbb{R}^{2s\times 2s}. \tag{3.9}$$

When Υ is invertible, η_∞ denotes the Ansatz with $\alpha = \Upsilon^{-1}(\operatorname{sign}(a), 0_s)$. The index ∞ comes from the random measurement model of [16], where η_∞ is the limit of certificates built from m random measurements when $m \to \infty$.

Theorem 3.1 (Existence of a non-degenerate certificate)**.** *Let Assumption 3.1 hold with s the number of spikes of μ^0, and assume $\kappa \geqslant 0$. Let $\Delta' \geqslant \Delta_{\text{adm}}$ with $r_{\text{near}} \leqslant \Delta'/2$, and assume $\mathfrak{d}_{\mathfrak{g}}(x_j,x_k) \geqslant \Delta'$ for $j \neq k$. Then the system $\Upsilon\alpha = (\operatorname{sign}(a), 0_s)$ has a unique solution, with*

$$\|\alpha_1 - \operatorname{sign}(a)\|_\infty \leqslant \frac{2h}{1-2h}, \qquad \|\alpha_1\|_\infty \leqslant \frac{1}{1-2h}, \qquad \sqrt{\mathfrak{g}}\,\|\alpha_2\|_\infty \leqslant \frac{2h}{1-2h}.$$

The Ansatz η_∞ is an $(\varepsilon_0/2, \varepsilon_2/4, r_{\text{near}})$-non-degenerate certificate, and

$$\left\|\operatorname{sign}(a_j)\eta_\infty''(x) - K^{(02)}(x_j,x)\right\| \leqslant \frac{\varepsilon_2}{16}, \qquad x \in \mathcal{X}_j^{\text{near}}, \quad j = 1,\dots,s.$$

For $s \geqslant 2$, the condition $r_{\text{near}} \leqslant \Delta'/2$ follows from Assumption 3.1; we admit this. For $s = 1$ there is no pair of spikes, and it only asks $\Delta' \geqslant 2r_{\text{near}}$. We write $\|M\|_\infty \overset{\text{def.}}{=} \max_i \sum_j |M_{ij}|$ for a square matrix M. It is the operator norm of M on $(\mathbb{R}^n, \|\cdot\|_\infty)$, so it is submultiplicative. If $\|E\|_\infty \leqslant q < 1$, the Neumann series $\sum_{n\geqslant 0}(-E)^n$ converges to $(\mathrm{Id}+E)^{-1}$, and $\|(\mathrm{Id}+E)^{-1}\|_\infty \leqslant 1/(1-q)$.

Proof, Step 3 admitted. Step 1, invertibility. Let $D \overset{\text{def.}}{=} \operatorname{diag}(\mathrm{Id}_s, \mathfrak{g}^{-1/2}\mathrm{Id}_s)$ and $\tilde{\Upsilon} \overset{\text{def.}}{=} D\Upsilon D$. By (3.9), the entry of $\tilde{\Upsilon}$ in row i and column ℓ of the block (p,q), with $p,q \in \{0,1\}$, is $\mathfrak{g}^{-(p+q)/2}K^{(qp)}(x_\ell,x_i)$. Its absolute value is $\left\|K^{(qp)}(x_\ell,x_i)\right\|$. The diagonal entries of the four blocks are $\kappa(0) = 1$, $\pm\mathfrak{g}^{-1/2}\kappa'(0) = 0$ and $-\kappa''(0)/\mathfrak{g} = 1$. So $E \overset{\text{def.}}{=} \tilde{\Upsilon} - \mathrm{Id}_{2s}$ has a zero diagonal. If $s = 1$, $E = 0$. Let $s \geqslant 2$. The separation $\min_{j\neq k}\mathfrak{d}_{\mathfrak{g}}(x_j,x_k) \geqslant \Delta'$ is larger than $\Delta(h,s)$: if $\Delta(h,s) > 0$ because $\Delta' \geqslant 2\Delta(h,s)$, and if $\Delta(h,s) = 0$ because $\Delta' \geqslant 2r_{\text{near}} > 0$. So Definition 3.5 applies to the family (x_ℓ), with x_i listed first. Row i of the block row p of E has absolute sum $\sum_{\ell\neq i}\left(\left\|K^{(0p)}(x_\ell,x_i)\right\| + \left\|K^{(1p)}(x_\ell,x_i)\right\|\right)$. These norms depend only on the order and on $|x_\ell - x_i|$, so they equal $\left\|K^{(p0)}(x_i,x_\ell)\right\|$ and $\left\|K^{(p1)}(x_i,x_\ell)\right\|$, with $(p,0),(p,1) \in \{0,1,2\}\times\{0,1\}$. Each sum is at most h, so $\|E\|_\infty \leqslant 2h \leqslant 1/64$. Hence $\tilde{\Upsilon} = \mathrm{Id} + E$ is invertible, with

$$\left\|\tilde{\Upsilon}^{-1}\right\|_\infty \leqslant \frac{1}{1-2h}, \qquad \left\|\tilde{\Upsilon}^{-1} - \mathrm{Id}\right\|_\infty = \left\|\tilde{\Upsilon}^{-1}E\right\|_\infty \leqslant \frac{2h}{1-2h}.$$

So $\Upsilon = D^{-1}\tilde{\Upsilon}D^{-1}$ is invertible, and the system has a unique solution α.

Step 2, coefficients. Let $u \overset{\text{def.}}{=} (\operatorname{sign}(a), 0_s)$. Since $Du = u$, the vector $\tilde{\alpha} \overset{\text{def.}}{=} D^{-1}\alpha = (\alpha_1, \sqrt{\mathfrak{g}}\alpha_2)$ solves $\tilde{\Upsilon}\tilde{\alpha} = u$. Hence $\tilde{\alpha} - u = (\tilde{\Upsilon}^{-1} - \mathrm{Id})u$ and $\tilde{\alpha} = \tilde{\Upsilon}^{-1}u$. Since $\|u\|_\infty = 1$, this gives the three bounds of the statement.

Step 3, far region (admitted). Let $x \in \mathcal{X}^{\text{far}}$. The separation leaves at most one spike $x_{i^\star}$ with $\mathfrak{d}_{\mathfrak{g}}(x,x_{i^\star}) < \Delta'/2$. Its term in the Ansatz is at most $|\alpha_{1,i^\star}|\,K(x_{i^\star},x) \leqslant (1-\varepsilon_0)/(1-2h)$, by $\kappa \geqslant 0$ and $\mathfrak{d}_{\mathfrak{g}}(x,x_{i^\star}) \geqslant r_{\text{near}}$. The other value terms are bounded with the kernel width applied to the family made of x and of the x_j, $j \neq i^\star$, which is only $\Delta'/2$-separated. The derivative terms are bounded with $B_{10} + h$. With $h \leqslant \varepsilon_0/(64B_0)$ this gives $|\eta_\infty(x)| \leqslant 1 - \varepsilon_0/2$ [16]. This is where the factor 2 in Δ_{adm} is used.

Step 4, near region. Fix i, let $\sigma_i \overset{\text{def.}}{=} \operatorname{sign}(a_i)$ and $x \in \mathcal{X}_i^{\text{near}}$. With $K^{(10)}(x_j, y) = \kappa'(x_j - y) = -\kappa'(y - x_j)$, the Ansatz reads $\eta_\infty(y) = \sum_j \alpha_{1,j}\kappa(y - x_j) - \alpha_{2,j}\kappa'(y - x_j)$. Since $K^{(02)}(x_i, x) = \kappa''(x - x_i)$,

$$\frac{\sigma_i\eta_\infty''(x) - \kappa''(x - x_i)}{\mathfrak{g}} = (\sigma_i\alpha_{1,i} - 1)\frac{\kappa''(x - x_i)}{\mathfrak{g}} + \sigma_i \sum_{j \neq i} \alpha_{1,j}\frac{\kappa''(x - x_j)}{\mathfrak{g}} - \sigma_i \sum_j \sqrt{\mathfrak{g}}\alpha_{2,j}\frac{\kappa'''(x - x_j)}{\mathfrak{g}^{3/2}}.$$

For $j \neq i$, $\mathfrak{d}_\mathfrak{g}(x, x_j) \geqslant \mathfrak{d}_\mathfrak{g}(x_i, x_j) - \mathfrak{d}_\mathfrak{g}(x, x_i) > \Delta' - r_{\text{near}} \geqslant \Delta'/2$. So the family made of x and of the x_j, $j \neq i$, has at most s points and separations larger than $\Delta(h, s)$. Definition 3.5, with x listed first, at the orders 2 and 3, bounds the sums over $j \neq i$ of $\mathfrak{g}^{-1}|\kappa''(x - x_j)|$ and of $\mathfrak{g}^{-3/2}|\kappa'''(x - x_j)|$ by h. The term $j = i$ of the last sum is bounded with B_{12}. With the bounds of Step 2,

$$\left\|\sigma_i\eta_\infty''(x) - K^{(02)}(x_i, x)\right\| \leqslant \frac{2hB_{02} + h + 2h(B_{12} + h)}{1 - 2h} \leqslant \frac{32}{31}2h(B_{02} + B_{12} + 1) \leqslant 4hB_2 \leqslant \frac{\varepsilon_2}{16},$$

where we used $1 + 2h \leqslant 2$, $h \leqslant 1/128$ and $h \leqslant \varepsilon_2/(64B_2)$. By construction $\eta_\infty(x_i) = \sigma_i$ and $\eta_\infty'(x_i) = 0$. Lemma 3.2 with $a = \sigma_i$, $\delta = \varepsilon_2/16$ and $r = r_{\text{near}}$ gives, on $\mathcal{X}_i^{\text{near}}$, $|\eta_\infty(x)| \leqslant 1 - \frac{7}{16}\varepsilon_2\,\mathfrak{d}_\mathfrak{g}(x, x_i)^2 \leqslant 1 - \frac{\varepsilon_2}{4}\mathfrak{d}_\mathfrak{g}(x, x_i)^2$.

Last, $\eta_\infty \in \operatorname{Im}(\Phi^\star)$ since $K(x_j, \cdot) = \Phi^\star\varphi_{x_j}$ and $K^{(10)}(x_j, \cdot) = \Phi^\star\partial_x\varphi_{x_j}$ (Remark 3.4). □

The Ansatz and the vanishing derivative pre-certificate solve the same interpolation problem. The next proposition makes this precise and draws the consequences for Lecture 4.

Proposition 3.4 (The certificate of the theorem is η_V)**.** *Under the assumptions of Theorem 3.1, with K the kernel of Φ:*

1. *$\Upsilon = \Gamma_{x_0}^\star\Gamma_{x_0}$, so $\eta_\infty = \eta_V$ and Γ_{x_0} is injective;*
2. *$\eta_\infty = \eta_0$ and μ^0 satisfies (NDSC);*
3. *$\eta_\infty = \Phi^\star p_\infty$ with $\|p_\infty\|_\mathbb{H}^2 = \langle \alpha_1, \operatorname{sign}(a)\rangle \leqslant s/(1 - 2h)$.*

Proof. Let $\Gamma \overset{\text{def.}}{=} \Gamma_{x_0}$ and $u \overset{\text{def.}}{=} (\operatorname{sign}(a), 0_s)$.

(1) By (3.5) and the symmetry $K^{(ij)}(x, x') = K^{(ji)}(x', x)$ of Remark 2.4, the entries of $\Gamma^\star\Gamma$ are those of (3.9): for instance $K^{(01)}(x_i, x_j) = K^{(10)}(x_j, x_i)$. So $\Upsilon = \Gamma^\star\Gamma$. By Theorem 3.1, Υ is invertible. If $\Gamma c = 0$, then $\Upsilon c = \Gamma^\star\Gamma c = 0$ and $c = 0$, so Γ is injective. Then $\alpha = (\Gamma^\star\Gamma)^{-1}u$ is the vector of Remark 3.4 (a), and (3.6) is the Ansatz. So $\eta_\infty = \eta_V$, and we set $p_\infty \overset{\text{def.}}{=} p_V = \Gamma\alpha$.

(2) Every $x \in \mathcal{X}$ lies in $\mathcal{X}^{\text{far}}$ or in some $\mathcal{X}_i^{\text{near}}$. By Theorem 3.1, $|\eta_\infty| \leqslant 1 - \varepsilon_0/2 < 1$ on $\mathcal{X}^{\text{far}}$, and $|\eta_\infty(x)| \leqslant 1 - \frac{\varepsilon_2}{4}\mathfrak{d}_\mathfrak{g}(x, x_i)^2$ on $\mathcal{X}_i^{\text{near}}$, which is at most 1, and less than 1 for $x \neq x_i$. So $\|\eta_V\|_\infty \leqslant 1$ and $|\eta_V(x)| < 1$ for $x \notin \{x_1, \dots, x_s\}$. By (1) and Proposition 3.3, (SC) holds and $\eta_0 = \eta_V = \eta_\infty$. It remains to check the curvature. Fix i and let $\psi \overset{\text{def.}}{=} \operatorname{sign}(a_i)\eta_0$. Then $\psi(x_i) = 1$, $\psi'(x_i) = 0$ (Remark 3.3), and $\psi(x_i + v) \leqslant 1 - \frac{\varepsilon_2}{4}\mathfrak{g}v^2$ for $\sqrt{\mathfrak{g}}|v| < r_{\text{near}}$. The Taylor-Young formula gives $\psi(x_i + v) = 1 + \frac{1}{2}\psi''(x_i)v^2 + o(v^2)$. Dividing by v^2 and letting $v \to 0$ gives $\frac{1}{2}\psi''(x_i) \leqslant -\frac{\varepsilon_2}{4}\mathfrak{g}$, that is $\operatorname{sign}(a_i)\eta_0''(x_i) \leqslant -\frac{\varepsilon_2}{2}\mathfrak{g} < 0$. So (NDSC) holds.

(3) By Remark 3.4 (b), $\|p_\infty\|_\mathbb{H}^2 = \langle \alpha_1, \operatorname{sign}(a)\rangle$. This is at most $\sum_j |\alpha_{1,j}| \leqslant s\|\alpha_1\|_\infty \leqslant s/(1 - 2h)$ by Theorem 3.1. □

The same computation gives the lower bound $\|p_\infty\|_\mathbb{H}^2 = s + \langle \alpha_1 - \operatorname{sign}(a), \operatorname{sign}(a)\rangle \geqslant s(1 - 4h)/(1 - 2h)$, so $\|p_\infty\|_\mathbb{H}$ is of order $\sqrt{s}$. Item 3 is the constant of Corollary 4.3: $\|p_\infty\|_\mathbb{H} \leqslant \sqrt{s/(1 - 2h)}$. Since $p_\infty = p_0$, every p with $\Phi^\star p \in \partial\|\mu^0\|_{\text{TV}}$ has $\|p\|_\mathbb{H} \geqslant \|p_\infty\|_\mathbb{H}$. The proof of item 2 also gives the curvature $\operatorname{sign}(a_i)\eta_0''(x_i) \leqslant -\varepsilon_2\mathfrak{g}/2$, of order f_c^2 for the Fourier measurements.

Proposition 3.5 (Constants for the two examples). *Gaussian kernel: every* $0 < r_{\text{near}} < 1$ *is allowed,* $\varepsilon_0 \geqslant 1 - e^{-{r_{\text{near}}}^2/2}$, $\varepsilon_2 \geqslant (1 - {r_{\text{near}}}^2)e^{-{r_{\text{near}}}^2/2}$, *and for every* $h > 0$ *there is* $D(h)$ *such that* $\Delta(h, s) \leqslant D(h)$ *for all* s. *Jackson kernel with* $f_c \geqslant 128$ *even:* $r_{\text{near}} = 1/(8\sqrt{2})$, $\varepsilon_0 \geqslant 0.00097$, $\varepsilon_2 \geqslant 0.941$, $\Delta(h, s) = \mathcal{O}(s^{1/4})$, *that is a separation of order at most* $s^{1/4}/f_c$ *on the torus.*

The Gaussian kernel is that of Example 2.2, written in the metric: $K(x, x') = e^{-\mathfrak{d}_{\mathfrak{g}}(x,x')^2/2}$ with $\mathfrak{d}_{\mathfrak{g}}(x, x') = |x - x'| \,/(\sqrt{3}\sigma)$. Its constants are proved in Exercise 3.3; there the two lower bounds are equalities. In dimension one the width does not grow with s: it is enough to separate the means of the mixture by a multiple of σ that does not depend on the number of components. The Jackson kernel is that of Example 2.1, with $\mathfrak{g} = \frac{\pi^2}{3} f_c(f_c + 4)$ (Exercise 2.1), so $\sqrt{\mathfrak{g}} \geqslant 1.81 f_c$, and a separation Δ in the metric is a separation of at most $0.56\Delta/f_c$ on $\mathbb{T}$. Both kernels are nonnegative, as Theorem 3.1 requires.

Remark 3.8 (Unweighted Fourier measurements). For the plain low-pass operator, a certificate exists for real amplitudes under the separation $1.87/f_c$ when $f_c \geqslant 128$ [18]. Nothing in these notes shows that this certificate is η_0, so this result does not give (NDSC) for $s \geqslant 2$.

The plain low-pass operator has $\varphi_x = (e^{2i\pi\omega x})_{|\omega| \leqslant f_c}$ and the Dirichlet kernel, with $K(x, x) = 2f_c + 1$, so Section 3.3 does not apply as it stands. The certificate of [18] is the Ansatz of Definition 3.7 built with the Jackson kernel J_{f_c} instead of the kernel of Φ. It lies in $\operatorname{Im}(\Phi^\star)$ because J_{f_c} is a trigonometric polynomial of degree f_c, and it is strict off the support. The atoms φ_{x_i} are linearly independent: the separation gives $s \leqslant f_c/1.87$, and the coordinates $\omega = 0, \ldots, s - 1$ form a Vandermonde matrix in the distinct numbers $e^{2i\pi x_i}$. So Proposition 3.1 gives exact recovery. For the weighted operator of Example 2.1 with the Jackson weights, (NDSC) under the separation Δ_{adm} follows from Proposition 3.4.

Exercises

Exercise 3.1 (A two-dimensional example). Let $\mathcal{X} = \mathbb{T}$, $\mathbb{H} = \mathbb{R}^2$ with the Euclidean scalar product, and $\varphi_x \overset{\text{def.}}{=} (\cos(2\pi x), \sin(2\pi x))$.

1. Compute K. For $p = \rho(\cos\theta, \sin\theta)$ with $\rho \geqslant 0$, show that $(\Phi^\star p)(x) = \rho\cos(2\pi x - \theta)$. Describe $C = \{p \,;\, \|\Phi^\star p\|_\infty \leqslant 1\}$.

2. Let $\mu^0 = a\delta_0$ with $a \neq 0$. Show that the set P of Section 3.1 is $\{(\operatorname{sign}(a), 0)\}$ and compute η_0. Show that $-a\delta_{1/2}$ is also a solution of $\mathcal{P}_0(\Phi\mu^0)$. Which hypothesis of Proposition 3.1 fails?

3. Let $\mu^0 = \delta_0 - \delta_{1/2}$. Find a certificate η with $|\eta| < 1$ off $\{0, 1/2\}$. Show that $2\delta_0$ is also a solution of $\mathcal{P}_0(\Phi\mu^0)$. Which hypothesis of Proposition 3.1 fails?

4. Let $\mu^0 = \delta_0 + \delta_{1/4}$. Show that (SC) fails. Show that $\sqrt{2}\delta_{1/8}$ is feasible and that μ^0 is not a solution of $\mathcal{P}_0(\Phi\mu^0)$.

5. In question 4, compute $\sup_{p \in C} \langle p, \Phi\mu^0 \rangle$ and $\min \mathcal{P}_0(\Phi\mu^0)$. Is the supremum attained? Compare with Lemma 3.1.

6. Replace φ_x by $(\cos(4\pi x), \sin(4\pi x))$ and let $\mu^0 = \delta_0$. Show that (SC) holds, that every certificate equals 1 at $1/2$, and that $\delta_{1/2}$ is also a solution of $\mathcal{P}_0(\Phi\mu^0)$. Which hypothesis of Proposition 3.1 fails?

Exercise 3.2 (Two spikes of opposite signs). Let Φ be as in Example 2.2 with $\sigma^2 = 1/3$, so that $\kappa(u) = e^{-u^2/2}$ and $\mathfrak{g} = 1$. Let $d > 0$, $x_1 = -d/2$, $x_2 = d/2$, and $a_1 > 0 > a_2$. Admit that Γ_{x_0} is injective.

1. Write the matrix $\Gamma_{x_0}^\star \Gamma_{x_0}$ with κ, κ' and κ'' at 0 and d.

2. Show that $\alpha_1 = (c, -c)$ and $\alpha_2 = (b, b)$, with

$$b = -\frac{c\,\kappa'(d)}{1-\kappa''(d)}, \qquad c = \frac{1}{D(d)}, \qquad D(d) \stackrel{\text{def.}}{=} 1 - \kappa(d) - \frac{\kappa'(d)^2}{1-\kappa''(d)}.$$

   Hint: first show that $1 - \kappa''(d) > 0$ and $D(d) > 0$, with the Cauchy-Schwarz inequality for the vectors $\varphi_{x_1} - \varphi_{x_2}$ and $\partial_x\varphi_{x_1} + \partial_x\varphi_{x_2}$, which are linearly independent. Then check that these values solve the system, and use injectivity.

3. Show that $\|p_V\|_{\mathbb{H}}^2 = 2c$. Deduce that $\|p_V\|_{\mathbb{H}} \to \infty$ when $d \to 0$.

4. Let $\eta = \Phi^\star p$ be any certificate of μ^0. Show that $\|p\|_{\mathbb{H}} \geqslant \|p_V\|_{\mathbb{H}}$ and $\|p\|_{\mathbb{H}} \geqslant 2/d$. Hint: the mean value theorem and $|\eta'| \leqslant \|\partial_x\varphi_x\|_{\mathbb{H}}\,\|p\|_{\mathbb{H}}$.

5. Assume the hypotheses of Theorem 3.1. Deduce from question 4 and from Proposition 3.4 that $d \geqslant \sqrt{2(1-2h)}$.

6. For $d = 1$, one finds $c \approx 39.08$, $b \approx 23.70$ and $\|\eta_V\|_\infty \approx 3.18$ (Figure 3.1). What can be said of (SC) and of (NDSC)?

Exercise 3.3 (The constants of the Gaussian kernel)**.** Let $\mathcal{X} = \mathbb{R}$ and $\kappa(u) = e^{-u^2/2}$, so that $\mathfrak{g} = 1$ and $\mathfrak{d}_{\mathfrak{g}}(x, x') = |x - x'|$.

1. Compute κ', κ'' and κ'''. Show that $B_{00} = B_{02} = 1$ and $B_{10} = e^{-1/2}$. Deduce that Assumption 3.1 allows every $r_{\text{near}} < 1$.

2. Show that $\varepsilon_0(r) = 1 - e^{-r^2/2}$ for $r > 0$ and that $\varepsilon_2(r) = (1 - r^2)e^{-r^2/2}$ for $0 < r \leqslant 1$. Hint: study $u \mapsto (1 - u^2)e^{-u^2/2}$ on $[0, \sqrt{3}]$.

3. Let $\Delta > 0$ and let $x_1, \ldots, x_m$ be points of $\mathbb{R}$ with $|x_k - x_\ell| \geqslant \Delta$ for $k \neq \ell$. Let $g : [0, \infty) \to [0, \infty)$ and $G(t) \stackrel{\text{def.}}{=} \sup_{u \geqslant t} g(u)$. Show that $\sum_{k \geqslant 2} g(|x_1 - x_k|) \leqslant 2 \sum_{j \geqslant 1} G(j\Delta)$. Hint: on each side of x_1, the j-th closest point is at distance at least $j\Delta$.

4. Show that $\left|\kappa^{(n)}(u)\right| \leqslant u^3 e^{-u^2/2}$ for $n \leqslant 3$ and $u \geqslant 2$, and that $u \mapsto u^3 e^{-u^2/2}$ is decreasing on $[\sqrt{3}, \infty)$. Deduce that for every $h > 0$ there is $D(h)$ such that $\Delta(h, s) \leqslant D(h)$ for all s.

5. In Example 2.2, show that $\mathfrak{d}_{\mathfrak{g}}(x, x') = |x - x'|\,/(\sqrt{3}\sigma)$ and that the constants above are those of the kernel of this example. Give a separation of the means, in units of σ and independent of s, under which Theorem 3.1 and Proposition 3.4 apply.

Exercise 3.4 (Secondary certificates)**.** Assume that $\Gamma \stackrel{\text{def.}}{=} \Gamma_{x_0}$ is injective, and let $(e_1, \ldots, e_{2s})$ be the canonical basis of $\mathbb{R}^{2s}$.

1. Show that $\Gamma^\star\Gamma$ is symmetric positive definite, and that $\Gamma^\star$ is onto.

2. For $e \in \mathbb{R}^{2s}$, show that $q_e \stackrel{\text{def.}}{=} \Gamma(\Gamma^\star\Gamma)^{-1}e$ is the element of minimal norm of $\{q \in \mathbb{H}\,;\, \Gamma^\star q = e\}$, and that $\|q_e\|_{\mathbb{H}}^2 = \langle e,\, (\Gamma^\star\Gamma)^{-1}e\rangle$.

3. Let $u_i \stackrel{\text{def.}}{=} q_{e_i}$ and $v_i \stackrel{\text{def.}}{=} q_{e_{s+i}}$, $\xi_i \stackrel{\text{def.}}{=} \Phi^\star u_i$ and $\zeta_i \stackrel{\text{def.}}{=} \Phi^\star v_i$. Show that $\xi_i(x_j) = \delta_{ij}$, $\xi_i'(x_j) = 0$, $\zeta_i(x_j) = 0$ and $\zeta_i'(x_j) = \delta_{ij}$. These are the functions of Appendix A. Show that $\left\|\xi_i^{(k)}\right\|_\infty \leqslant \sup_x \left\|\partial_x^k\varphi_x\right\|_{\mathbb{H}} \|u_i\|_{\mathbb{H}}$ for $k \leqslant 2$.

4. Assume the setting of Theorem 3.1, and let $\tilde{\Upsilon} = D\Upsilon D$ be the matrix of its proof. Show that $\|u_i\|_{\mathbb{H}}^2 = (\tilde{\Upsilon}^{-1})_{ii}$ and $\|v_i\|_{\mathbb{H}}^2 = \mathfrak{g}^{-1}(\tilde{\Upsilon}^{-1})_{s+i,s+i}$. Deduce that $\|u_i\|_{\mathbb{H}}^2 \leqslant 1/(1-2h)$ and $\|v_i\|_{\mathbb{H}}^2 \leqslant 1/(\mathfrak{g}(1-2h))$.

5. Show that $p_V = \sum_i \operatorname{sign}(a_i)u_i$. Compare with item 3 of Proposition 3.4.

Chapter 4

Noise: prediction, localisation, support stability

Throughout, $\mu^0 = \sum_{i=1}^s a_i \delta_{x_i}$, $y = \Phi\mu^0 + w$, and $\hat{\mu}$ is any solution of $\mathcal{P}_\lambda(y)$.

As in Lecture 3, $s \geqslant 1$, the points x_i are pairwise distinct, $a_i \neq 0$, $\lambda > 0$, and the assumption of Proposition 2.1 holds. So $\mathcal{P}_\lambda(y)$ has a solution (Proposition 2.2), $\Phi^\star p \in C_0(\mathcal{X})$ for every $p \in \mathbb{H}$, and (2.8) holds. We write $J_\lambda(\mu) \overset{\text{def.}}{=} \frac{1}{2}\|\Phi\mu - y\|_{\mathbb{H}}^2 + \lambda\|\mu\|_{\mathrm{TV}}$ for the objective of $\mathcal{P}_\lambda(y)$.

The lecture has two parts. Sections 4.1 and 4.2 hold at any noise level: a certificate of μ^0 bounds the prediction error and the mass of $\hat{\mu}$ away from the spikes. Section 4.3 holds at low noise only: under (NDSC), $\hat{\mu}$ has exactly s spikes, close to those of μ^0.

4.1 Prediction and localisation

The first result is Proposition 1.3 on measures, with the same proof. Only the last step changes. The margin of the certificate is no longer a constant off the support: it is quadratic near each spike and constant far from the spikes (Definition 3.6).

Proposition 4.1 (Stability away from the support)**.** *Let $\eta = \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$ and $\rho \overset{\text{def.}}{=} \|w\|_{\mathbb{H}} + \lambda\|p\|_{\mathbb{H}}$.*

1. *$\|\Phi(\hat{\mu} - \mu^0)\|_{\mathbb{H}} \leqslant 2\rho$ and $\|\hat{\mu}\|_{\mathrm{TV}} - \int \eta \,\mathrm{d}\hat{\mu} \leqslant \rho^2/(2\lambda)$.*
2. *Assume the setting of Section 3.3. If η is $(\varepsilon_0, \varepsilon_2, r_{\mathrm{near}})$-non-degenerate and the near regions $\mathcal{X}_i^{\mathrm{near}}$ are pairwise disjoint, then*

$$\varepsilon_0 \,|\hat{\mu}|\,(\mathcal{X}^{\mathrm{far}}) + \varepsilon_2 \sum_{i=1}^s \int_{\mathcal{X}_i^{\mathrm{near}}} \mathfrak{d}_{\mathfrak{g}}(x, x_i)^2 \,\mathrm{d}\,|\hat{\mu}|\,(x) \leqslant \frac{\rho^2}{2\lambda}. \tag{4.1}$$

Corollary 4.1 (Choice of the regularisation parameter)**.** *If $\|w\|_{\mathbb{H}} \leqslant \delta$ with $\delta > 0$ and $\lambda = \delta/\|p\|_{\mathbb{H}}$, then $\|\Phi(\hat{\mu} - \mu^0)\|_{\mathbb{H}} \leqslant 4\delta$ and the left-hand side of (4.1) is at most $2\delta\|p\|_{\mathbb{H}}$.*

Proof. All norms and scalar products are those of $\mathbb{H}$. Let $h \overset{\text{def.}}{=} \Phi(\hat{\mu} - \mu^0)$ and $D \overset{\text{def.}}{=} \|\hat{\mu}\|_{\mathrm{TV}} - \langle \eta, \hat{\mu}\rangle_{\mathcal{M}}$. By Proposition 2.3, $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \mu^0\rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}$. So $D \geqslant 0$ by (2.6).

Compare with μ^0. Since $\Phi\hat{\mu} - y = h - w$ and $\Phi\mu^0 - y = -w$, the inequality $J_\lambda(\hat{\mu}) \leqslant J_\lambda(\mu^0)$ reads

$$\tfrac{1}{2}\|h - w\|^2 + \lambda\|\hat{\mu}\|_{\mathrm{TV}} \leqslant \tfrac{1}{2}\|w\|^2 + \lambda\|\mu^0\|_{\mathrm{TV}}, \quad \text{that is} \quad \tfrac{1}{2}\|h\|^2 - \langle h, w\rangle + \lambda(\|\hat{\mu}\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}}) \leqslant 0.$$

Use the certificate. By (2.8), $\langle \eta, \hat{\mu} - \mu^0\rangle_{\mathcal{M}} = \langle p, \Phi(\hat{\mu} - \mu^0)\rangle = \langle p, h\rangle$. Hence

$$\|\hat{\mu}\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}} = D + \langle \eta, \hat{\mu} - \mu^0\rangle_{\mathcal{M}} = D + \langle p, h\rangle.$$

With the Cauchy-Schwarz inequality,

$$\tfrac{1}{2}\|h\|^2 + \lambda D \leqslant \langle h,\, w - \lambda p\rangle \leqslant \|h\|\,(\,\|w\| + \lambda\,\|p\|\,) = \rho\,\|h\|\,. \tag{4.2}$$

Item 1. Since $\lambda D \geqslant 0$, (4.2) gives $\frac{1}{2}\|h\|^2 \leqslant \rho\,\|h\|$, hence $\|h\| \leqslant 2\rho$ (divide by $\|h\|$ if $h \neq 0$). Then $\lambda D \leqslant \rho\,\|h\| - \frac{1}{2}\|h\|^2 \leqslant \max_{u\geqslant 0}\,(\rho u - \frac{1}{2}u^2) = \frac{1}{2}\rho^2$.

Item 2. The sets $\mathcal{X}_i^{\mathrm{near}}$ are open and $\mathcal{X}^{\mathrm{far}}$ is closed. Since the $\mathcal{X}_i^{\mathrm{near}}$ are pairwise disjoint, $\mathcal{X}^{\mathrm{far}}, \mathcal{X}_1^{\mathrm{near}}, \dots, \mathcal{X}_s^{\mathrm{near}}$ is a partition of $\mathcal{X}$ into Borel sets. By (2.6),

$$D \geqslant \int_{\mathcal{X}} (1 - |\eta|)\,\mathrm{d}\,|\hat{\mu}| = \int_{\mathcal{X}^{\mathrm{far}}} (1 - |\eta|)\,\mathrm{d}\,|\hat{\mu}| + \sum_{i=1}^{s} \int_{\mathcal{X}_i^{\mathrm{near}}} (1 - |\eta|)\,\mathrm{d}\,|\hat{\mu}|\,.$$

By Definition 3.6, $1 - |\eta| \geqslant \varepsilon_0$ on $\mathcal{X}^{\mathrm{far}}$ and $1 - |\eta(x)| \geqslant \varepsilon_2\,\mathfrak{d}_{\mathfrak{g}}(x, x_i)^2$ on $\mathcal{X}_i^{\mathrm{near}}$. So D is at least the left-hand side of (4.1), and item 1 concludes. □

Proof of Corollary 4.1. Since $s \geqslant 1$, $|\eta(x_1)| = 1$, so $\eta \neq 0$ and $p \neq 0$ (Remark 1.6 (b)). So $\lambda > 0$, $\lambda\,\|p\|_{\mathbb{H}} = \delta$ and $\rho \leqslant 2\delta$. Item 1 gives $\|\Phi(\hat{\mu} - \mu^0)\|_{\mathbb{H}} \leqslant 4\delta$, and $\rho^2/(2\lambda) \leqslant 4\delta^2\,\|p\|_{\mathbb{H}}\,/(2\delta) = 2\delta\,\|p\|_{\mathbb{H}}$. □

Remark 4.1 (On the proof). (a) No optimality condition is used, only $J_\lambda(\hat{\mu}) \leqslant J_\lambda(\mu^0)$. So the bounds hold for every measure with this property, and for every certificate p, not only for p_0. No uniqueness, no separation and no bound on the noise are used, and the dimension of $\mathbb{H}$ plays no role. (b) The quantity $D = \|\hat{\mu}\|_{\mathrm{TV}} - \int \eta\,\mathrm{d}\hat{\mu}$ is the Bregman divergence of the total variation at μ^0 for the subgradient η. Item 2 bounds it from below. Lecture 5 recovers (4.2) from Fenchel-Young losses (Theorem 5.3), and Appendix A uses the bound $D \geqslant \int (1 - |\eta_0|)\,\mathrm{d}\,|\hat{\mu}|$. (c) If $\mathfrak{d}_{\mathfrak{g}}(x_j, x_k) \geqslant 2r_{\mathrm{near}}$ for $j \neq k$, the near regions are pairwise disjoint (Remark 3.7). Without disjointness the sum in (4.1) could count the same mass twice. (d) Item 1 controls $\Phi\hat{\mu}$, not $\hat{\mu}$. Item 2 controls where the mass of $\hat{\mu}$ lies, but not its sign.

The bounds use $\|p\|_{\mathbb{H}}$ in a monotone way. So an upper bound on the norm is enough to choose λ.

Remark 4.2 (A bound on the norm is enough). Let $B \geqslant \|p\|_{\mathbb{H}}$ with $B > 0$, $\|w\|_{\mathbb{H}} \leqslant \delta$ with $\delta > 0$, and $\lambda \overset{\text{def.}}{=} \delta/B$. Then $\lambda\,\|p\|_{\mathbb{H}} \leqslant \delta$, so $\rho \leqslant 2\delta$, $\|\Phi(\hat{\mu} - \mu^0)\|_{\mathbb{H}} \leqslant 4\delta$ and $\rho^2/(2\lambda) \leqslant 4\delta^2 B/(2\delta) = 2\delta B$. So Corollary 4.1, and Corollary 4.2 below, hold with B in place of $\|p\|_{\mathbb{H}}$.

Exercise 4.1 treats a model error: the data are $y = \Phi(\mu^0 + \tilde{\mu}_0) + w$ with a small measure $\tilde{\mu}_0$.

The left-hand side of (4.1) is a transport cost. To read it as a distance between $|\hat{\mu}|$ and a measure supported on $\{x_1, \dots, x_s\}$, we use a distance that moves mass at a quadratic cost and creates or destroys mass at a linear cost.

Definition 4.1 (Partial optimal transport)**.** For positive measures μ, ν on $\mathcal{X}$,

$$\mathcal{T}^2_{\mathfrak{d}_{\mathfrak{g}}}(\mu, \nu) \overset{\text{def.}}{=} \inf_{\tilde{\mu}, \tilde{\nu}} \{W^2_{\mathfrak{d}_{\mathfrak{g}}}(\tilde{\mu}, \tilde{\nu}) + |\mu - \tilde{\mu}|\,(\mathcal{X}) + |\nu - \tilde{\nu}|\,(\mathcal{X})\},$$

where the infimum is over positive measures of equal mass and $W_{\mathfrak{d}_{\mathfrak{g}}}$ is the Wasserstein-2 distance for the metric $\mathfrak{d}_{\mathfrak{g}}$.

For positive measures $\tilde{\mu}, \tilde{\nu}$ of equal mass, a coupling is a positive measure γ on $\mathcal{X} \times \mathcal{X}$ with first marginal $\tilde{\mu}$ and second marginal $\tilde{\nu}$, and $W^2_{\mathfrak{d}_{\mathfrak{g}}}(\tilde{\mu}, \tilde{\nu})$ is the infimum of $\int \mathfrak{d}_{\mathfrak{g}}(x, x')^2\,\mathrm{d}\gamma(x, x')$ over the couplings. We only use that one coupling gives an upper bound. If both measures are zero, the coupling $\gamma = 0$ gives $W^2_{\mathfrak{d}_{\mathfrak{g}}} = 0$.

Corollary 4.2 (Localisation in transport distance)**.** *Under the assumptions of Proposition 4.1, item 2, with $\varepsilon_0, \varepsilon_2 > 0$, and of Corollary 4.1, with $\hat{A}_i \overset{\text{def.}}{=} |\hat{\mu}|\,(\mathcal{X}_i^{\mathrm{near}})$,*

$$\mathcal{T}^2_{\mathfrak{d}_{\mathfrak{g}}}\Big(\,|\hat{\mu}|\,, \textstyle\sum_{i=1}^{s} \hat{A}_i \delta_{x_i}\Big) \leqslant \frac{2\delta\|p\|_{\mathbb{H}}}{\min(\varepsilon_0, \varepsilon_2)}.$$

Proof. Let L be the left-hand side of (4.1), $N \overset{\text{def.}}{=} \bigcup_i \mathcal{X}_i^{\text{near}}$ and $\nu \overset{\text{def.}}{=} \sum_i \hat{A}_i \delta_{x_i}$. We bound the infimum of Definition 4.1 by its value at one pair. Discard the mass of $|\hat{\mu}|$ on $\mathcal{X}^{\text{far}}$: let $\tilde{\mu} \overset{\text{def.}}{=} |\hat{\mu}|_{|N}$ and $\tilde{\nu} \overset{\text{def.}}{=} \nu$. Since the $\mathcal{X}_i^{\text{near}}$ are pairwise disjoint, $\tilde{\mu}(\mathcal{X}) = \sum_i \hat{A}_i = \tilde{\nu}(\mathcal{X})$. Also $|\hat{\mu}| - \tilde{\mu} = |\hat{\mu}|_{|\mathcal{X}^{\text{far}}}$ and $\nu - \tilde{\nu} = 0$.

Move the mass on $\mathcal{X}_i^{\text{near}}$ to x_i: the positive measure $\gamma \overset{\text{def.}}{=} \sum_i |\hat{\mu}|_{|\mathcal{X}_i^{\text{near}}} \otimes \delta_{x_i}$ has first marginal $\sum_i |\hat{\mu}|_{|\mathcal{X}_i^{\text{near}}} = \tilde{\mu}$ and second marginal $\sum_i \hat{A}_i \delta_{x_i} = \nu$. So it is a coupling, and

$$W_{\mathfrak{d}_{\mathfrak{g}}}^2(\tilde{\mu}, \nu) \leqslant \int \mathfrak{d}_{\mathfrak{g}}(x, x')^2 \, \mathrm{d}\gamma(x, x') = \sum_{i=1}^{s} \int_{\mathcal{X}_i^{\text{near}}} \mathfrak{d}_{\mathfrak{g}}(x, x_i)^2 \, \mathrm{d}\,|\hat{\mu}|\,(x).$$

Adding the cost $|\hat{\mu}|\,(\mathcal{X}^{\text{far}})$ of the discarded mass,

$$\mathcal{T}_{\mathfrak{d}_{\mathfrak{g}}}^2(|\hat{\mu}|\,, \nu) \leqslant |\hat{\mu}|\,(\mathcal{X}^{\text{far}}) + \sum_{i=1}^{s} \int_{\mathcal{X}_i^{\text{near}}} \mathfrak{d}_{\mathfrak{g}}(x, x_i)^2 \, \mathrm{d}\,|\hat{\mu}|\,(x) \leqslant \frac{L}{\min(\varepsilon_0, \varepsilon_2)} \leqslant \frac{2\delta \, \|p\|_{\mathbb{H}}}{\min(\varepsilon_0, \varepsilon_2)},$$

by Corollary 4.1. □

The proof gives $\mathcal{T}_{\mathfrak{d}_{\mathfrak{g}}}^2 \leqslant L/\min(\varepsilon_0, \varepsilon_2) \leqslant \rho^2/(2\lambda \min(\varepsilon_0, \varepsilon_2))$ for every $\lambda > 0$. With Remark 4.2, it gives $2\delta B/\min(\varepsilon_0, \varepsilon_2)$ for $\lambda = \delta/B$.

Remark 4.3 (Mass far from the spikes). Assume the setting of Proposition 4.1, item 2, with $\varepsilon_0, \varepsilon_2 > 0$. Let $t > 0$ and $E_t \overset{\text{def.}}{=} \{x \in \mathcal{X} \,;\, \mathfrak{d}_{\mathfrak{g}}(x, x_i) \geqslant t \text{ for all } i\}$. A point of E_t lies in $\mathcal{X}^{\text{far}}$, or in some $\mathcal{X}_i^{\text{near}}$ with $\mathfrak{d}_{\mathfrak{g}}(x, x_i) \geqslant t$, and there $1 \leqslant \mathfrak{d}_{\mathfrak{g}}(x, x_i)^2/t^2$. This is Markov's inequality. With L the left-hand side of (4.1),

$$|\hat{\mu}|\,(E_t) \leqslant |\hat{\mu}|\,(\mathcal{X}^{\text{far}}) + \frac{1}{t^2} \sum_{i=1}^{s} \int_{\mathcal{X}_i^{\text{near}}} \mathfrak{d}_{\mathfrak{g}}(x, x_i)^2 \, \mathrm{d}\,|\hat{\mu}|\,(x) \leqslant \max\Big(\frac{1}{\varepsilon_0}, \frac{1}{\varepsilon_2 t^2}\Big) L.$$

Under the assumptions of Corollary 4.1, $L \leqslant 2\delta \, \|p\|_{\mathbb{H}}$: for every fixed t, the mass at distance at least t from the spikes is $\mathcal{O}(\delta)$.

Remark 4.4 (Amplitudes, admitted). The bounds above say where the mass of $\hat{\mu}$ lies, not that the near region $\mathcal{X}_j^{\text{near}}$ receives the signed mass a_j. Under the assumptions of Proposition 4.1, item 2, assume that for some j there is $\eta_j = \Phi^\star p_j$ with $\eta_j(x_j) = 1$, $\eta_j(x_\ell) = 0$ for $\ell \neq j$, $|1 - \eta_j(x)| \leqslant b\, \mathfrak{d}_{\mathfrak{g}}(x, x_j)^2$ on $\mathcal{X}_j^{\text{near}}$ for some $b \geqslant 0$, $|\eta_j(x)| \leqslant \varepsilon_2 \, \mathfrak{d}_{\mathfrak{g}}(x, x_\ell)^2$ on $\mathcal{X}_\ell^{\text{near}}$ for $\ell \neq j$, and $|\eta_j| \leqslant 1$ on $\mathcal{X}^{\text{far}}$. If $\varepsilon_0, \varepsilon_2 > 0$, then

$$\Big|\hat{\mu}(\mathcal{X}_j^{\text{near}}) - a_j\Big| \leqslant 2\rho \, \|p_j\|_{\mathbb{H}} + \Big(\max\big(1, \tfrac{b}{\varepsilon_2}\big) + \frac{1}{\varepsilon_0}\Big) \frac{\rho^2}{2\lambda}.$$

The proof is Exercise 4.2.

4.2 The two examples

We apply Section 4.1 with the certificate $\eta_\infty = \Phi^\star p_\infty$ of Proposition 3.4. The only new input is the bound $\|p_\infty\|_{\mathbb{H}} \leqslant \sqrt{s/(1 - 2h)}$. For the Gaussian mixture we also bound the noise w_n, which comes from the sampling.

Corollary 4.3 (Weighted Fourier measurements)**.** *In Example 2.1 with the Jackson weights, let Assumption 3.1 hold for the kernel J_{f_c}, with s the number of spikes of μ^0. Assume that the x_i are Δ_{adm}-separated, that is $\mathfrak{d}_{\mathfrak{g}}(x_j, x_k) \geqslant \Delta_{\text{adm}}$ for $j \neq k$, and that $\|w\|_{\mathbb{H}} \leqslant \delta$ with $\delta > 0$. With the certificate $\eta_\infty = \Phi^\star p_\infty$ of Proposition 3.4 and $\lambda = \delta/\|p_\infty\|_{\mathbb{H}}$, Corollaries 4.1 and 4.2 hold with constants $(\varepsilon_0/2, \varepsilon_2/4)$ and $\|p_\infty\|_{\mathbb{H}} \leqslant \sqrt{s/(1 - 2h)}$.*

Proof. The setting of Section 3.3 holds: $\mathcal{X} = \mathbb{T}$, $x \mapsto \varphi_x$ is $\mathscr{C}^\infty$, and the kernel is $J_{f_c}(x - x')$ with $J_{f_c}(0) = 1$. The kernel is a fourth power, so $\kappa = J_{f_c} \geqslant 0$. If $s \geqslant 2$, let $\Delta' \stackrel{\text{def.}}{=} \Delta_{\text{adm}}$; then $r_{\text{near}} \leqslant \Delta'/2$, as noted after Theorem 3.1. If $s = 1$, let $\Delta' \stackrel{\text{def.}}{=} \max(\Delta_{\text{adm}}, 2r_{\text{near}})$: there is no pair of spikes to separate. In both cases Theorem 3.1 and Proposition 3.4 apply. So $\eta_\infty = \Phi^\star p_\infty$ is a $(\varepsilon_0/2, \varepsilon_2/4, r_{\text{near}})$-non-degenerate certificate, $\|p_\infty\|_{\mathbb{H}} \leqslant \sqrt{s/(1-2h)}$, and $\eta_\infty \in \partial\|\mu^0\|_{\text{TV}}$ (Remark 3.7 (b)). For $j \neq k$, $\mathfrak{d}_{\mathfrak{g}}(x_j, x_k) \geqslant \Delta' \geqslant 2r_{\text{near}}$, so the near regions are pairwise disjoint. So Proposition 4.1 and Corollaries 4.1 and 4.2 apply with $p = p_\infty$ and the constants $(\varepsilon_0/2, \varepsilon_2/4)$. Since $1/\min(\varepsilon_0/2, \varepsilon_2/4) \leqslant 4/\min(\varepsilon_0, \varepsilon_2)$,

$$\mathcal{T}^2_{\mathfrak{d}_{\mathfrak{g}}}\Big(|\hat{\mu}|, \textstyle\sum_{i=1}^s \hat{A}_i \delta_{x_i}\Big) \leqslant \frac{8\delta}{\min(\varepsilon_0, \varepsilon_2)} \sqrt{\frac{s}{1-2h}}. \qquad \square$$

By Proposition 3.5 (admitted), Assumption 3.1 holds for the Jackson kernel when $f_c \geqslant 128$ is even, with $r_{\text{near}} = 1/(8\sqrt{2})$. Since $h \leqslant 1/128$ (Remark 3.6), $\sqrt{s/(1-2h)} \leqslant \sqrt{64s/63}$. In torus units, $\mathfrak{d}_{\mathfrak{g}}(x, x_i)^2 = \mathfrak{g}\,|x - x_i|^2$ with $\mathfrak{g} = \frac{\pi^2}{3} f_c(f_c + 4)$ (Exercise 2.1). So item 2 of Proposition 4.1 gives $\sum_i \int_{\mathcal{X}_i^{\text{near}}} |x - x_i|^2 \,\mathrm{d}\,|\hat{\mu}|\,(x) \leqslant 8\delta\,\|p_\infty\|_{\mathbb{H}}/(\varepsilon_2 \mathfrak{g})$: the spread around the spikes, in squared torus distance, is of order $\delta\sqrt{s}/f_c^2$.

Remark 4.5 (The plain low-pass operator, slide). (a) The factor $\sqrt{s}$ cannot be removed with another certificate. Since $p_\infty = p_0$ has minimal norm among the p with $\Phi^\star p \in \partial\|\mu^0\|_{\text{TV}}$, every such p satisfies $\|p\|^2_{\mathbb{H}} \geqslant \|p_\infty\|^2_{\mathbb{H}} \geqslant s(1-4h)/(1-2h)$, by the lower bound that follows Proposition 3.4. (b) For the plain low-pass operator of Remark 3.8, with $\varphi_x = (e^{2\mathrm{i}\pi\omega x})_{|\omega| \leqslant f_c}$, the characters are orthonormal in $\mathrm{L}^2(\mathbb{T})$. So the certificate of Remark 3.8 satisfies $\|p\|_2 = \|\eta\|_{\mathrm{L}^2(\mathbb{T})} < 1$ for every s. We admit that this gives localisation bounds of order δ, with no factor $\sqrt{s}$, under the separation $1.87/f_c$. The two noise levels are not on the same scale: the plain atoms have norm $\sqrt{2f_c + 1}$, the weighted atoms have norm 1. For the weighted operator, the same computation only gives $\|\eta\|_{\mathrm{L}^2(\mathbb{T})} \leqslant \max_\omega \sqrt{\Lambda(\omega)}\ \|p\|_{\mathbb{H}}$, which bounds η by p and not p by η.

For the Gaussian mixture the noise w_n is random. We avoid the expectation of a random vector of $\mathbb{H}$: we compute expectations frequency by frequency, then integrate in ω.

Lemma 4.1 (Noise of the empirical characteristic function)**.** *In Example 2.2 let* $w_n \stackrel{\text{def.}}{=} y_n - \Phi\mu^0$. *Then* $\mathbb{E}\,\|w_n\|^2_{\mathbb{H}} \leqslant C^2/n$. *For all* $\zeta \in (0,1)$, *with probability at least* $1 - \zeta$, $\|w_n\|_{\mathbb{H}} \leqslant c\sqrt{(1 + \log(1/\zeta))/n}$ *for a numerical constant* c.

Corollary 4.4 (Gaussian mixture)**.** *In Example 2.2, let Assumption 3.1 hold for the kernel* $e^{-(x-x')^2/(6\sigma^2)}$, *with* s *the number of components. Assume that the means* x_i *are* Δ_{adm}*-separated, that is* $|x_j - x_k| \geqslant \sqrt{3}\,\sigma\,\Delta_{\text{adm}}$ *for* $j \neq k$. *Let* $\zeta \in (0,1)$, c *the constant of Lemma 4.1,* $\delta_n \stackrel{\text{def.}}{=} c\sqrt{(1 + \log(1/\zeta))/n}$ *and* $\lambda \stackrel{\text{def.}}{=} \delta_n\sqrt{(1-2h)/s} \asymp \sqrt{(1 + \log(1/\zeta))/(ns)}$. *With probability at least* $1 - \zeta$, *every solution* $\hat{\mu}$ *of* $\mathcal{P}_\lambda(y_n)$ *satisfies* $\|\Phi(\hat{\mu} - \mu^0)\|_{\mathbb{H}} \leqslant 4\delta_n$ *and*

$$\mathcal{T}^2_{\mathfrak{d}_{\mathfrak{g}}}\Big(|\hat{\mu}|, \textstyle\sum_{i=1}^s \hat{A}_i \delta_{x_i}\Big) \leqslant \frac{8c}{\min(\varepsilon_0, \varepsilon_2)} \sqrt{\frac{s\,(1 + \log(1/\zeta))}{(1-2h)\,n}} \lesssim \frac{1}{\min(\varepsilon_0, \varepsilon_2)} \sqrt{\frac{s\,(1 + \log(1/\zeta))}{n}}.$$

Proof of Lemma 4.1. Here $\mathbb{H} = \mathrm{L}^2(\Lambda)$ with the real scalar product, so $\|f\|^2_{\mathbb{H}} = \int_{\mathbb{R}} |f|^2 \,\mathrm{d}\Lambda$, and Λ is a probability.

Variance. Fix $\omega \in \mathbb{R}$ and let $Z_k \stackrel{\text{def.}}{=} Ce^{\mathrm{i}\omega z_k}$. The Z_k are independent complex random variables with the law of Z_1, $|Z_k| = C$, and $\mathbb{E}Z_k = (\Phi\mu^0)(\omega)$ by Example 2.2. So $w_n(\omega) = \frac{1}{n}\sum_k (Z_k - \mathbb{E}Z_k)$. For $k \neq \ell$, independence gives $\mathbb{E}\big[(Z_k - \mathbb{E}Z_k)\overline{(Z_\ell - \mathbb{E}Z_\ell)}\big] = 0$. Hence

$$\mathbb{E}\,|w_n(\omega)|^2 = \frac{1}{n^2}\sum_{k=1}^n \mathbb{E}\,|Z_k - \mathbb{E}Z_k|^2 = \frac{1}{n}\big(\mathbb{E}\,|Z_1|^2 - |\mathbb{E}Z_1|^2\big) = \frac{C^2 - \big|(\Phi\mu^0)(\omega)\big|^2}{n}.$$

The function $(\omega, z_1, \dots, z_n) \mapsto |w_n(\omega)|^2$ is continuous and nonnegative. So Fubini's theorem for nonnegative functions lets us exchange $\mathbb{E}$ and the integral in ω:

$$\mathbb{E}\,\|w_n\|_{\mathbb{H}}^2 = \int_{\mathbb{R}} \mathbb{E}\,|w_n(\omega)|^2 \,\mathrm{d}\Lambda(\omega) = \frac{C^2 - \|\Phi\mu^0\|_{\mathbb{H}}^2}{n} \leqslant \frac{C^2}{n} = \frac{\sqrt{3}}{n}. \tag{4.3}$$

Only expectations of real variables are used.

Bound in probability (admitted inequality). We admit the following form of Pinelis' inequality [19], a Hoeffding inequality in Hilbert spaces. Let $W_1, \dots, W_n$ be independent random vectors of a separable real Hilbert space, with $\|W_k\| \leqslant b$ and $\mathbb{E}\langle W_k, g\rangle = 0$ for every vector g. Then, for $t > 0$,

$$\mathbb{P}\Big(\Big\|\sum_{k=1}^{n} W_k\Big\| \geqslant t\Big) \leqslant 2\exp\Big(-\frac{t^2}{2nb^2}\Big).$$

Let W_k be the function $\omega \mapsto Ce^{\mathrm{i}\omega z_k} - (\Phi\mu^0)(\omega)$, so that $w_n = \frac{1}{n}\sum_k W_k$. The map $z \mapsto Ce^{\mathrm{i}\cdot z}$ is continuous from $\mathbb{R}$ to $\mathbb{H}$ by dominated convergence, so W_k is a random vector of $\mathbb{H}$. Since $\|Ce^{\mathrm{i}\cdot z}\|_{\mathbb{H}} = C$ and $\|\Phi\mu^0\|_{\mathbb{H}} \leqslant \sum_i a_i \|\varphi_{x_i}\|_{\mathbb{H}} = 1 \leqslant C$, we get $\|W_k\|_{\mathbb{H}} \leqslant 2C$. For $g \in \mathbb{H}$, the integrand of $\langle W_k, g\rangle_{\mathbb{H}} = \Re\int W_k\,\overline{g}\,\mathrm{d}\Lambda$ is bounded by $2C\,|g|$, which is integrable for $\mathbb{P}\otimes\Lambda$. So Fubini's theorem and $\mathbb{E}[Ce^{\mathrm{i}\omega z_k}] = (\Phi\mu^0)(\omega)$ give $\mathbb{E}\langle W_k, g\rangle_{\mathbb{H}} = 0$. Apply the inequality with $b = 2C$ and $t = 2C\sqrt{2n(1+\log(1/\zeta))}$. Then $t^2/(2nb^2) = 1 + \log(1/\zeta)$, and the probability is at most $2e^{-1}\zeta < \zeta$. So, with probability at least $1-\zeta$,

$$\|w_n\|_{\mathbb{H}} \leqslant \frac{t}{n} = 2\sqrt{2}\,C\sqrt{\frac{1+\log(1/\zeta)}{n}},$$

and $c = 2\sqrt{2}\,C = 2\sqrt{2}\cdot 3^{1/4} \leqslant 3.73$ works. □

Proof of Corollary 4.4. The setting of Section 3.3 holds for Example 2.2 (Example 3.1), the kernel is positive, and $\mathfrak{d}_{\mathfrak{g}}(x,x') = |x-x'|/(\sqrt{3}\sigma)$ (Section 3.3). So the assumption on the means reads $\mathfrak{d}_{\mathfrak{g}}(x_j,x_k) \geqslant \Delta_{\mathrm{adm}}$. As in the proof of Corollary 4.3, Theorem 3.1 and Proposition 3.4 give a $(\varepsilon_0/2, \varepsilon_2/4, r_{\mathrm{near}})$-non-degenerate certificate $\eta_\infty = \Phi^\star p_\infty \in \partial\|\mu^0\|_{\mathrm{TV}}$, with pairwise disjoint near regions and $\|p_\infty\|_{\mathbb{H}} \leqslant B \overset{\text{def.}}{=} \sqrt{s/(1-2h)}$. By definition $\lambda = \delta_n/B$.

By Lemma 4.1, the event $\{\|w_n\|_{\mathbb{H}} \leqslant \delta_n\}$ has probability at least $1-\zeta$. On this event, apply Proposition 4.1 with $p = p_\infty$ and $w = w_n$, and Remark 4.2. This gives $\|\Phi(\hat\mu - \mu^0)\|_{\mathbb{H}} \leqslant 4\delta_n$ and, as in the proof of Corollary 4.2,

$$\mathcal{T}^2_{\mathfrak{d}_{\mathfrak{g}}}\Big(|\hat\mu|, \textstyle\sum_i \hat A_i \delta_{x_i}\Big) \leqslant \frac{2\delta_n B}{\min(\varepsilon_0/2, \varepsilon_2/4)} \leqslant \frac{8\delta_n B}{\min(\varepsilon_0,\varepsilon_2)}.$$

This is the bound of the statement. The event does not depend on $\hat\mu$, so the bound holds for every solution. By Remark 3.6, $h \leqslant 1/128$, so $63/64 \leqslant 1-2h \leqslant 1$. This gives the last inequality of the statement and $\lambda \asymp \sqrt{(1+\log(1/\zeta))/(ns)}$. □

By Proposition 3.5, for the Gaussian kernel the width $\Delta(h,s)$ is bounded independently of s. So the separation of the means, in units of σ, does not grow with the number of components. The factor $\sqrt{s}$ comes from $\|p_\infty\|_{\mathbb{H}}$. Here all the frequencies are kept, while a sketch keeps only m random frequencies [10].

4.3 Support stability at low noise

In this section $\mathcal{X} = \mathbb{T}$, $\mathbb{H} = \mathrm{L}^2(\mathbb{T})$ and $\Phi\mu = \varphi \star \mu$ with $\varphi \in \mathscr{C}^2(\mathbb{T})$. We write $y_0 = \Phi\mu^0$, $\hat p_\lambda(y)$ for the dual solution of Proposition 2.4, and D_{α,λ_0} for the low-noise regime of (2.1).

This is the convolution of Remark 2.5, with φ real and $\mathbb{H}$ the real space $\mathrm{L}^2(\mathbb{T})$. So $\varphi_x = \varphi(\cdot - x)$, and the assumption of Proposition 2.1 holds (Remark 2.7). The definition of D_{α,λ_0}

in (2.1) does not use the grid: it is the set of (λ, w) with $0 < \lambda \leqslant \lambda_0$ and $\|w\|_{\mathbb{H}} \leqslant \alpha\lambda$. Let $C \overset{\text{def.}}{=} \{p \in \mathbb{H}\,;\, \|\Phi^\star p\|_\infty \leqslant 1\}$. By Remark 2.8, $\hat{p}_\lambda(y)$ is the projection of y/λ onto C, and by Proposition 2.4, $\lambda\hat{p}_\lambda(y) = y - \Phi\hat{\mu}$ for every solution $\hat{\mu}$ of $\mathcal{P}_\lambda(y)$. The section follows the route of Remark 2.1 (c): the dual solution is non-expansive in the data, and it converges to p_0 when the noise vanishes.

Regularity. Let $M_k \overset{\text{def.}}{=} \left\|\varphi^{(k)}\right\|_{\mathrm{L}^2}$ for $k \leqslant 2$. The map $x \mapsto \varphi_x$ is $\mathscr{C}^2$ from $\mathbb{T}$ to $\mathbb{H}$, with $\partial_x^k \varphi_x = (-1)^k \varphi^{(k)}(\cdot - x)$. Indeed, let $k \in \{0, 1\}$ and $u \neq 0$. For each t, the mean value theorem gives $\theta \in (0, 1)$ with

$$\frac{\varphi^{(k)}(t - x - u) - \varphi^{(k)}(t - x)}{u} + \varphi^{(k+1)}(t - x) = \varphi^{(k+1)}(t - x) - \varphi^{(k+1)}(t - x - \theta u).$$

Since $\varphi^{(k+1)}$ is uniformly continuous on $\mathbb{T}$, the right-hand side tends to 0 uniformly in t when $u \to 0$, hence in $\mathrm{L}^2(\mathbb{T})$, since $\mathbb{T}$ has measure one. The continuity of $x \mapsto \varphi^{(k)}(\cdot - x)$ in $\mathbb{H}$ follows in the same way from the uniform continuity of $\varphi^{(k)}$. If $\varphi \in \mathscr{C}^k(\mathbb{T})$, the same argument shows that $x \mapsto \varphi_x$ is $\mathscr{C}^k$. By Remark 2.4, for $p \in \mathbb{H}$, $\Phi^\star p \in \mathscr{C}^2(\mathbb{T})$ and, since the norm of $\mathrm{L}^2(\mathbb{T})$ is invariant by translation,

$$\left\|(\Phi^\star p)^{(k)}\right\|_\infty \leqslant \sup_x \left\|\partial_x^k \varphi_x\right\|_{\mathbb{H}} \|p\|_{\mathbb{H}} = M_k \|p\|_{\mathbb{H}}\,, \qquad k \in \{0, 1, 2\}. \tag{4.4}$$

In particular Γ_x, η_V and (NDSC) of Lecture 3 are defined in this setting. The mean value inequality for the $\mathscr{C}^1$ map $x \mapsto \varphi_x$ also gives $\|\varphi_u - \varphi_v\|_{\mathbb{H}} \leqslant M_1 |u - v|$ for $u, v \in \mathbb{T}$.

The projection inequality. Let $\lambda > 0$ and $y \in \mathbb{H}$. The proof of Proposition 2.4, item 2, gives, for all $q \in C$,

$$\langle y/\lambda - \hat{p}_\lambda(y),\, q - \hat{p}_\lambda(y)\rangle_{\mathbb{H}} \leqslant 0. \tag{4.5}$$

Lemma 4.2 (The dual solution is non-expansive)**.** *For $\lambda > 0$ and $y, y' \in \mathbb{H}$, $\|\hat{p}_\lambda(y) - \hat{p}_\lambda(y')\|_{\mathbb{H}} \leqslant \|y - y'\|_{\mathbb{H}} / \lambda$.*

Theorem 4.1 (Convergence of the certificates in $\mathscr{C}^2$)**.** *Assume (SC). Let $\tilde{\eta}_\lambda \overset{\text{def.}}{=} \Phi^\star \hat{p}_\lambda(y_0 + w)$ and $M_k \overset{\text{def.}}{=} \left\|\varphi^{(k)}\right\|_{\mathrm{L}^2}$. For $k \in \{0, 1, 2\}$,*

$$\left\|\tilde{\eta}_\lambda^{(k)} - \eta_0^{(k)}\right\|_\infty \leqslant M_k(\, \|\hat{p}_\lambda(y_0) - p_0\|_{\mathbb{H}} + \|w\|_{\mathbb{H}} / \lambda).$$

Proof of Lemma 4.2. By Proposition 2.4, item 2, $\hat{p}_\lambda(y)$ and $\hat{p}_\lambda(y')$ are the projections of y/λ and y'/λ onto the closed convex set C. We admit that the projection onto a nonempty closed convex subset of a Hilbert space is 1-Lipschitz [14]. Hence $\|\hat{p}_\lambda(y) - \hat{p}_\lambda(y')\|_{\mathbb{H}} \leqslant \|y/\lambda - y'/\lambda\|_{\mathbb{H}} = \|y - y'\|_{\mathbb{H}} / \lambda$. □

The 1-Lipschitz property also follows from (4.5): write it for y with $q = \hat{p}_\lambda(y')$ and for y' with $q = \hat{p}_\lambda(y)$, add, and apply the Cauchy-Schwarz inequality. This is question 1 of Exercise 1.4.

Proof of Theorem 4.1. By (SC) and Lemma 3.1, p_0 and $\eta_0 = \Phi^\star p_0$ are defined. Let $q \overset{\text{def.}}{=} \hat{p}_\lambda(y_0 + w) - p_0$. By linearity, $\tilde{\eta}_\lambda^{(k)} - \eta_0^{(k)} = (\Phi^\star q)^{(k)}$, so (4.4) gives $\left\|\tilde{\eta}_\lambda^{(k)} - \eta_0^{(k)}\right\|_\infty \leqslant M_k \|q\|_{\mathbb{H}}$. By the triangle inequality and Lemma 4.2 with $y = y_0 + w$ and $y' = y_0$,

$$\|q\|_{\mathbb{H}} \leqslant \|\hat{p}_\lambda(y_0 + w) - \hat{p}_\lambda(y_0)\|_{\mathbb{H}} + \|\hat{p}_\lambda(y_0) - p_0\|_{\mathbb{H}} \leqslant \frac{\|w\|_{\mathbb{H}}}{\lambda} + \|\hat{p}_\lambda(y_0) - p_0\|_{\mathbb{H}}\,. \qquad \square$$

The bound holds for all $\lambda > 0$ and $w \in \mathbb{H}$. The second term is the effect of the noise. On D_{α,λ_0} it is at most α. The first term is a bias at noiseless data. The next proposition shows that it tends to zero.

Proposition 4.2 (The dual bias vanishes)**.** *Assume (SC). For all* $\lambda > 0$, $\|\hat{p}_\lambda(y_0)\|_{\mathbb{H}} \leqslant \|p_0\|_{\mathbb{H}}$, *and* $\hat{p}_\lambda(y_0) \to p_0$ *strongly in* $\mathbb{H}$ *when* $\lambda \to 0$.

Proof. All norms and scalar products are those of $\mathbb{H}$. Write $p_\lambda \overset{\text{def.}}{=} \hat{p}_\lambda(y_0)$. By Lemma 3.1, $p_0 \in C$, p_0 maximises $\langle \cdot, y_0 \rangle$ on C, the set of maximisers is P, and p_0 is the unique element of minimal norm of P.

Norm bound. By (4.5) with $y = y_0$ and $q = p_0$, multiplied by λ,

$$0 \leqslant \langle y_0,\, p_0 - p_\lambda \rangle \leqslant \lambda \langle p_\lambda,\, p_0 - p_\lambda \rangle, \tag{4.6}$$

where the first inequality holds because $p_\lambda \in C$. So $\|p_\lambda\|^2 \leqslant \langle p_\lambda, p_0 \rangle \leqslant \|p_\lambda\| \, \|p_0\|$, and $\|p_\lambda\| \leqslant \|p_0\|$. Also, since $u \, \|p_0\| - u^2 \leqslant \|p_0\|^2 / 4$ for $u \in \mathbb{R}$, (4.6) gives

$$0 \leqslant \langle y_0,\, p_0 - p_\lambda \rangle \leqslant \lambda \big(\, \|p_\lambda\| \, \|p_0\| - \|p_\lambda\|^2 \, \big) \leqslant \frac{\lambda}{4} \, \|p_0\|^2 \, . \tag{4.7}$$

Monotonicity. Let $0 < \lambda' < \lambda$ and $d \overset{\text{def.}}{=} p_{\lambda'} - p_\lambda$. Write (4.5) at λ with $q = p_{\lambda'}$, and at λ' with $q = p_\lambda$, multiplied by λ and λ':

$$\langle y_0 - \lambda p_\lambda,\, d \rangle \leqslant 0, \qquad \langle y_0 - \lambda' p_{\lambda'},\, -d \rangle \leqslant 0.$$

Adding gives $\langle \lambda' p_{\lambda'} - \lambda p_\lambda,\, d \rangle \leqslant 0$. With $p_{\lambda'} = p_\lambda + d$, this reads $\lambda' \, \|d\|^2 \leqslant (\lambda - \lambda') \langle p_\lambda,\, d \rangle$, so $\langle p_\lambda,\, d \rangle \geqslant 0$. Hence

$$\|p_{\lambda'}\|^2 = \|p_\lambda\|^2 + 2 \langle p_\lambda,\, d \rangle + \|d\|^2 \geqslant \|p_\lambda\|^2 + \|p_{\lambda'} - p_\lambda\|^2 \, . \tag{4.8}$$

Convergence. By (4.8), $N(\lambda) \overset{\text{def.}}{=} \|p_\lambda\|^2$ does not increase with λ, and $N \leqslant \|p_0\|^2$. So $N(\lambda)$ has a limit ℓ when $\lambda \to 0$, and $N(\lambda') \leqslant \ell$ for all λ'. For $0 < \lambda' < \lambda$, (4.8) gives $\|p_{\lambda'} - p_\lambda\|^2 \leqslant N(\lambda') - N(\lambda) \leqslant \ell - N(\lambda)$, which tends to 0 when $\lambda \to 0$. So (p_λ) satisfies the Cauchy criterion when $\lambda \to 0$. Since $\mathbb{H}$ is complete, p_λ converges strongly to some $q \in \mathbb{H}$, and $q \in C$ because C is closed.

Identification. Letting $\lambda \to 0$ in (4.7) gives $\langle y_0, q \rangle = \langle y_0, p_0 \rangle$. So q maximises $\langle \cdot, y_0 \rangle$ on C, that is $q \in P$. Also $\|q\| = \lim \|p_\lambda\| \leqslant \|p_0\|$. Since p_0 is the unique element of minimal norm of P, $q = p_0$. □

Remark 4.6 (On the bias). (a) With $\lambda \hat{p}_\lambda(y_0) = y_0 - \Phi \hat{\mu}_\lambda$ for every solution $\hat{\mu}_\lambda$ of $\mathcal{P}_\lambda(y_0)$, the norm bound gives $\|\Phi \hat{\mu}_\lambda - y_0\|_{\mathbb{H}} \leqslant \lambda \, \|p_0\|_{\mathbb{H}}$. (b) Under (SC) alone the proof gives no rate. For the projection onto a closed convex set of $\mathbb{R}^2$ that is not of the form of C, the convergence can be slower than any power of λ given in advance; we admit this. Under (NDSC), with Γ_{x_0} injective and $\varphi \in \mathscr{C}^3(\mathbb{T})$, we admit that $\|\hat{p}_\lambda(y_0) - p_0\|_{\mathbb{H}} = \mathcal{O}(\lambda)$, from the regularity of the path of Theorem 4.2. We do not use a rate. On a grid, under the assumption of Corollary 2.1, the bias is zero for small λ (Remark 2.1 (b)), and Exercise 4.3 gives a case where it is zero on measures. (c) The proof does not use $\mathcal{X} = \mathbb{T}$ or the convolution: it holds in the setting of Lecture 3.

With Theorem 4.1 and Proposition 4.2, the certificate $\tilde{\eta}_\lambda$ of the noisy problem is uniformly close to η_0, with its first two derivatives, on D_{α,λ_0} for α and λ_0 small. The main theorem turns this into a statement on the solution.

Theorem 4.2 (Support stability, Duval and Peyré)**.** *Assume that* Γ_{x_0} *is injective and that* μ^0 *satisfies (NDSC). Then there are* $\alpha, \lambda_0 > 0$ *such that, for all* $(\lambda, w) \in D_{\alpha,\lambda_0}$, *the solution of* $\mathcal{P}_\lambda(y_0 + w)$ *is unique and has exactly* s *spikes,* $\hat{\mu}_\lambda = \sum_{i=1}^s \hat{a}_{\lambda,i} \delta_{\hat{x}_{\lambda,i}}$, *with* $\operatorname{sign}(\hat{a}_{\lambda,i}) = \operatorname{sign}(a_i)$. *With* $\lambda = \|w\|_{\mathbb{H}} / \alpha$,

$$|\hat{x}_{\lambda,i} - x_i| = \mathcal{O}(\|w\|_{\mathbb{H}}) \quad \textit{and} \quad |\hat{a}_{\lambda,i} - a_i| = \mathcal{O}(\|w\|_{\mathbb{H}}).$$

If $\varphi \in \mathscr{C}^k(\mathbb{T})$ *with* $k \geqslant 2$, *the map* $(\lambda, w) \mapsto (\hat{a}_\lambda, \hat{x}_\lambda)$ *is* $\mathscr{C}^{k-1}$.

The spikes are labelled so that $\hat{x}_{\lambda,i}$ is close to x_i, and $|\hat{x}_{\lambda,i} - x_i|$ is the distance on $\mathbb{T}$. The rates hold for $0 < \|w\|_{\mathbb{H}} \leqslant \alpha\lambda_0$, so that $(\|w\|_{\mathbb{H}}/\alpha, w) \in D_{\alpha,\lambda_0}$, and the constants in $\mathcal{O}$ depend on μ^0 and φ. The regularity means that the map is the restriction to D_{α,λ_0} of a $\mathscr{C}^{k-1}$ map defined on a neighbourhood of $(0,0)$ in $\mathbb{R} \times \mathbb{H}$.

Outline, Step 3 admitted. Let $\sigma_i \overset{\text{def.}}{=} \operatorname{sign}(a_i)$ and $\sigma \overset{\text{def.}}{=} \operatorname{sign}(a)$. Under (NDSC), (SC) holds, and p_0 and η_0 are defined. For $(\lambda, w) \in D_{\alpha,\lambda_0}$, let $\hat{p} \overset{\text{def.}}{=} \hat{p}_\lambda(y_0 + w)$ and $\tilde{\eta}_\lambda \overset{\text{def.}}{=} \Phi^\star \hat{p}$. Then $\|\tilde{\eta}_\lambda\|_\infty \leqslant 1$ since $\hat{p} \in C$.

Step 1, convergence of the certificates. Let $b(\lambda_0) \overset{\text{def.}}{=} \sup_{0<\lambda\leqslant\lambda_0} \|\hat{p}_\lambda(y_0) - p_0\|_{\mathbb{H}}$. By Proposition 4.2, $b(\lambda_0) \to 0$ when $\lambda_0 \to 0$. On D_{α,λ_0}, $\|w\|_{\mathbb{H}}/\lambda \leqslant \alpha$, so Theorem 4.1 gives $\left\|\tilde{\eta}_\lambda^{(k)} - \eta_0^{(k)}\right\|_\infty \leqslant M_k(b(\lambda_0) + \alpha)$ for $k \leqslant 2$. So, for every $\tau > 0$, there are $\alpha, \lambda_0 > 0$ with $\left\|\tilde{\eta}_\lambda^{(k)} - \eta_0^{(k)}\right\|_\infty < \tau$ for $k \in \{0, 2\}$ and all $(\lambda, w) \in D_{\alpha,\lambda_0}$. Smaller values of α and λ_0 keep this property.

Step 2, at most one spike near each x_i, none elsewhere. Let $\kappa \overset{\text{def.}}{=} \frac{1}{2}\min_i |\eta_0''(x_i)|$, positive by (NDSC). Since $\sigma_i\eta_0(x_i) = 1$, $\sigma_i\eta_0''(x_i) \leqslant -2\kappa$, and η_0, η_0'' are continuous, there is $\varepsilon \in (0, \frac{1}{4})$, smaller than half the distance between two spikes, such that $\sigma_i\eta_0'' \leqslant -\kappa$ and $\sigma_i\eta_0 \geqslant \frac{1}{2}$ on $I_i \overset{\text{def.}}{=} (x_i - \varepsilon, x_i + \varepsilon)$. Every smaller ε has the same properties. The set $\mathbb{T} \setminus \bigcup_i I_i$ is compact and $|\eta_0| < 1$ there, so $r_0 \overset{\text{def.}}{=} 1 - \max_{\mathbb{T}\setminus\cup_i I_i} |\eta_0| > 0$. Step 1 with $\tau = \min(r_0, \frac{1}{2}, \kappa)$ gives α, λ_0 such that, on D_{α,λ_0}, $|\tilde{\eta}_\lambda| < 1$ off $\bigcup_i I_i$, and $\sigma_i\tilde{\eta}_\lambda > 0$ and $\sigma_i\tilde{\eta}_\lambda'' < 0$ on each I_i. If $|\tilde{\eta}_\lambda(u)| = |\tilde{\eta}_\lambda(v)| = 1$ with $u \neq v$ in I_i, then $\sigma_i\tilde{\eta}_\lambda(u) = \sigma_i\tilde{\eta}_\lambda(v) = 1$, and strict concavity gives $\sigma_i\tilde{\eta}_\lambda(\frac{u+v}{2}) > 1$, which contradicts $\|\tilde{\eta}_\lambda\|_\infty \leqslant 1$. So $\{x\,;\, |\tilde{\eta}_\lambda(x)| = 1\}$ has at most one point in each I_i and no point elsewhere. Let $\hat{\mu}$ be a solution of $\mathcal{P}_\lambda(y_0 + w)$. By Proposition 2.4, item 4, and Remark 2.6 (i), $\hat{\mu} = \sum_i \hat{a}_i\delta_{\hat{x}_i}$ with $\hat{x}_i \in I_i$; we set $\hat{a}_i \overset{\text{def.}}{=} 0$ and $\hat{x}_i \overset{\text{def.}}{=} x_i$ if $\hat{\mu}$ has no mass in I_i. By item 1 of the same proposition and Proposition 2.3, $\operatorname{sign}(\hat{a}_i) = \tilde{\eta}_\lambda(\hat{x}_i) = \sigma_i$ when $\hat{a}_i \neq 0$, and then Remark 3.3 gives $\tilde{\eta}_\lambda'(\hat{x}_i) = 0$. This is Proposition A.1.

Step 3, implicit function theorem (admitted). For $a \in \mathbb{R}^s$, $x \in \mathbb{T}^s$, $\lambda \in \mathbb{R}$ and $w \in \mathbb{H}$, let

$$E(a, x, \lambda, w) \overset{\text{def.}}{=} \Gamma_x^\star(\Phi_x a - y_0 - w) + \lambda \begin{pmatrix} \sigma \\ 0_s \end{pmatrix} \in \mathbb{R}^{2s}.$$

For $\lambda > 0$ and $p \overset{\text{def.}}{=} (y_0 + w - \Phi_x a)/\lambda$, (3.4) shows that $E = 0$ if and only if $(\Phi^\star p)(x_i) = \sigma_i$ and $(\Phi^\star p)'(x_i) = 0$ for all i: these are the optimality conditions at the points of the support. If $\varphi \in \mathscr{C}^k(\mathbb{T})$ with $k \geqslant 2$, the map E is $\mathscr{C}^{k-1}$, since $x \mapsto \varphi_x$ is $\mathscr{C}^k$. Let $a_0 \overset{\text{def.}}{=} a$. Then $E(a_0, x_0, 0, 0) = 0$, since $\Phi_{x_0}a_0 = y_0$. In the partial derivative in (a, x) at this point, the terms where $\Gamma_x^\star$ is differentiated are multiplied by the residual $\Phi_{x_0}a_0 - y_0 = 0$, and $\partial_{x_j}(\Phi_x a) = a_j\,\partial_x\varphi_{x_j}$. So

$$\frac{\partial E}{\partial(a,x)}(a_0, x_0, 0, 0) = \Gamma_{x_0}^\star(\Phi_{x_0}, \Phi_{x_0}'\operatorname{diag}(a_0)) = \Gamma_{x_0}^\star\Gamma_{x_0}\begin{pmatrix} \mathrm{Id}_s & 0 \\ 0 & \operatorname{diag}(a_0) \end{pmatrix}.$$

This matrix is invertible because $\Gamma_{x_0}^\star\Gamma_{x_0}$ is (proof of Proposition 3.2) and $a_i \neq 0$. The implicit function theorem gives open neighbourhoods V of (a_0, x_0) and W of $(0,0)$ in $\mathbb{R} \times \mathbb{H}$, and a $\mathscr{C}^{k-1}$ map $f : W \to V$, such that for $(a, x, \lambda, w) \in V \times W$, $E(a, x, \lambda, w) = 0$ if and only if $(a, x) = f(\lambda, w)$. In particular $f(0,0) = (a_0, x_0)$. We admit the regularity of E and the implicit function theorem in $\mathbb{R}^{2s} \times (\mathbb{R} \times \mathbb{H})$ [9].

Step 4, the solution is on the branch. The sets V and W depend on φ, a_0 and x_0 only. Fix $r_V \in (0, \min_i |a_i|]$ such that V contains the points (a', x') with $\|a' - a_0\|_\infty < r_V$ and $|x_i' - x_i| < r_V$ for all i. Since Φ_{x_0} is injective (Remark 3.4 (d)) on a space of finite dimension, there is $\theta > 0$ with $\theta\|b\|_\infty \leqslant \|\Phi_{x_0}b\|_{\mathbb{H}}$ for $b \in \mathbb{R}^s$. Let $\hat{\mu}$ be as in Step 2, with $\alpha, \lambda_0 \leqslant 1$. By Proposition 4.1, item 1, with $p = p_0$, $\|\Phi\hat{\mu} - y_0\|_{\mathbb{H}} \leqslant 2(\|w\|_{\mathbb{H}} + \lambda\|p_0\|_{\mathbb{H}}) \leqslant 2(1 + \|p_0\|_{\mathbb{H}})\lambda_0$. The inequality $J_\lambda(\hat{\mu}) \leqslant J_\lambda(\mu^0)$ gives $\|\hat{\mu}\|_{\mathrm{TV}} \leqslant \|\mu^0\|_{\mathrm{TV}} + \|w\|_{\mathbb{H}}^2/(2\lambda) \leqslant \|\mu^0\|_{\mathrm{TV}} + 1$. Since $\Phi\hat{\mu} - y_0 = \Phi_{x_0}(\hat{a} - a_0) + \sum_i \hat{a}_i(\varphi_{\hat{x}_i} - \varphi_{x_i})$ and $\|\varphi_{\hat{x}_i} - \varphi_{x_i}\|_{\mathbb{H}} \leqslant M_1\varepsilon$,

$$\theta\|\hat{a} - a_0\|_\infty \leqslant 2(1 + \|p_0\|_{\mathbb{H}})\lambda_0 + M_1\varepsilon\,(\|\mu^0\|_{\mathrm{TV}} + 1).$$

Now choose, in this order: ε as in Step 2 with $\varepsilon < r_V$ and $M_1\varepsilon(\|\mu^0\|_{\mathrm{TV}}+1) < \theta r_V/2$; then $\alpha, \lambda_0 \leqslant 1$ as in Step 2 for this ε; then a smaller λ_0 with $2(1+\|p_0\|_{\mathbb{H}})\lambda_0 < \theta r_V/2$ and $D_{\alpha,\lambda_0} \subset W$. Then $\|\hat{a}-a_0\|_\infty < r_V \leqslant \min_i |a_i|$. So all $\hat{a}_i$ are nonzero, $\hat{\mu}$ has exactly s spikes, one in each I_i, with $\mathrm{sign}(\hat{a}_i) = \sigma_i$, and $(\hat{a},\hat{x}) \in V$. By Step 2, $\tilde{\eta}_\lambda(\hat{x}_i) = \sigma_i$ and $\tilde{\eta}'_\lambda(\hat{x}_i) = 0$. Since $\lambda\hat{p} = y_0 + w - \Phi_{\hat{x}}\hat{a}$, this is $E(\hat{a},\hat{x},\lambda,w) = 0$. Hence $(\hat{a},\hat{x}) = f(\lambda,w)$. This holds for every solution, so the solution is unique, and $(\lambda,w) \mapsto (\hat{a},\hat{x})$ is the restriction of f.

Rates. Since $k \geqslant 2$, f is $\mathscr{C}^1$. Near x_0 we see $\mathbb{T}^s$ as an open set of $\mathbb{R}^s$. Reduce λ_0 so that D_{α,λ_0} lies in a ball of $\mathbb{R}\times\mathbb{H}$ centred at $(0,0)$ on which the differential of f is bounded by some L. The mean value inequality gives $|f(\lambda,w)-f(0,0)| \leqslant L(\lambda+\|w\|_{\mathbb{H}})$, which is at most $L(1+1/\alpha)\|w\|_{\mathbb{H}}$ for $\lambda = \|w\|_{\mathbb{H}}/\alpha$. □

Remark 4.7 (What each assumption does). (a) The curvature $\eta_0''(x_i) \neq 0$ is used in Step 2 only: it gives at most one spike near each x_i. The condition $|\eta_0| < 1$ off the spikes gives no spike elsewhere. The injectivity of Γ_{x_0} is used in Steps 3 and 4. (b) The Jacobian of Step 3 does not involve η_0'', because the residual vanishes at $(a_0,x_0,0,0)$. (c) The order of the choices in Step 4 matters. The neighbourhoods V and W come first, then ε, then α and λ_0. (d) Here $\lambda > 0$ in D_{α,λ_0}, since $\mathcal{P}_\lambda(y)$ is not defined for $\lambda = 0$; the map f is still defined at $(0,0)$.

Remark 4.8 (A proof without the implicit function theorem). Appendix A gives a complete proof of the first two conclusions and of the rates, without Step 3. It does not give the regularity of the path.

Remark 4.9 (Where (NDSC) comes from). For the weighted Fourier measurements of Example 2.1, (NDSC) follows from Proposition 3.4. For the Gaussian mixture, which lives on $\mathbb{R}$, the same conclusion is a special case of Theorem 5.4 with a quadratic fidelity. We admit it.

For the weighted Fourier measurements, $\mathbb{H} = \mathbb{C}^{2f_c+1}$ and not $\mathrm{L}^2(\mathbb{T})$. Exercise 4.4 shows that this model is a convolution on $\mathbb{T}$ with the same kernel. Proposition 3.4 gives (NDSC) and the injectivity of Γ_{x_0} under the separation Δ_{adm}, so Theorem 4.2 applies under the assumptions of Corollary 4.3. For the Gaussian mixture, Theorem 5.4 asks $\|w_n\|_{\mathbb{H}} \leqslant \alpha\lambda$ and $\lambda \leqslant \lambda_0$. By Lemma 4.1, both hold with probability at least $1-\zeta$ for $\lambda = \delta_n/\alpha$, as soon as $\delta_n \leqslant \alpha\lambda_0$, that is for n large (Exercise 4.5).

Under $\varphi \in \mathscr{C}^3(\mathbb{T})$, the constants of the proof can be tracked, with $\alpha = \min\left(r/(2M_0), \min_i |\eta_0''(x_i)|/(4M_2)\right)$, where r is the margin of η_0 off the intervals I_i. The constants of the rates involve $\|(\Gamma_{x_0}^\star\Gamma_{x_0})^{-1}\|$, and also $1/\min_i|a_i|$ for the positions. We admit these values. A converse holds [9]: if Γ_{x_0} is injective and, for some $\lambda_0 > 0$, the solutions of $\mathcal{P}_\lambda(y_0)$ for $0 < \lambda < \lambda_0$, and of $\mathcal{P}_0(y_0)$ for $\lambda = 0$, follow a $\mathscr{C}^1$ path of s spikes defined on $[0,\lambda_0)$, then $\|\eta_V\|_\infty \leqslant 1$ and $\eta_V = \eta_0$. So the test $\|\eta_V\|_\infty \leqslant 1$ of Remark 3.5, in its non-strict form, is necessary for such a path.

Exercises

Exercise 4.1 (Model error)**.** Let $y = \Phi(\mu^0+\tilde{\mu}_0)+w$ with $\|\tilde{\mu}_0\|_{\mathrm{TV}} \leqslant \tilde{\delta}$. Show that the bounds of Proposition 4.1 hold with an additional term controlled by $\tilde{\delta}$.

Let $\eta = \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$, $\rho \overset{\text{def.}}{=} \|w\|_{\mathbb{H}} + \lambda\|p\|_{\mathbb{H}}$, $\bar{\mu} \overset{\text{def.}}{=} \mu^0+\tilde{\mu}_0$, $\bar{h} \overset{\text{def.}}{=} \Phi(\hat{\mu}-\bar{\mu})$ and $D \overset{\text{def.}}{=} \|\hat{\mu}\|_{\mathrm{TV}} - \int \eta\,\mathrm{d}\hat{\mu}$.

1. Show that $\|\bar{\mu}\|_{\mathrm{TV}} - \int\eta\,\mathrm{d}\bar{\mu} \leqslant 2\tilde{\delta}$.
2. Compare $J_\lambda(\hat{\mu})$ with $J_\lambda(\bar{\mu})$, and show that
$$\tfrac{1}{2}\left\|\bar{h}-(w-\lambda p)\right\|_{\mathbb{H}}^2 + \lambda D \leqslant \tfrac{1}{2}\|w-\lambda p\|_{\mathbb{H}}^2 + 2\lambda\tilde{\delta}.$$
3. Deduce that $D \leqslant \rho^2/(2\lambda) + 2\tilde{\delta}$ and that $\|\Phi\hat{\mu}-y\|_{\mathbb{H}} \leqslant \|w\|_{\mathbb{H}} + 2\lambda\|p\|_{\mathbb{H}} + 2\sqrt{\lambda\tilde{\delta}}$. Bound $\|\Phi(\hat{\mu}-\mu^0)\|_{\mathbb{H}}$ with $B \overset{\text{def.}}{=} \sup_x \|\varphi_x\|_{\mathbb{H}}$.

4. Under the assumptions of Corollary 4.2, show that

$$\mathcal{T}^2_{\mathfrak{d}_\mathfrak{g}}\left(|\hat{\mu}|\,, \sum_{i=1}^s \hat{A}_i \delta_{x_i}\right) \leqslant \frac{2\delta\|p\|_{\mathbb{H}}+2\tilde{\delta}}{\min(\varepsilon_0,\varepsilon_2)}.$$

Compare with Corollary 4.2. Can the term in $\tilde{\delta}$ be made small by the choice of λ?

Exercise 4.2 (Amplitudes and positions)**.** Assume the setting of Remark 4.4. Let $\nu \overset{\text{def.}}{=} \hat{\mu} - \mu^0$ and let L be the left-hand side of (4.1).

1. Show that $x_\ell \notin \mathcal{X}_j^{\text{near}}$ for $\ell \neq j$, and that $\mu^0(\mathcal{X}_j^{\text{near}}) = a_j$.

2. Show that

$$\hat{\mu}(\mathcal{X}_j^{\text{near}}) - a_j = \int \eta_j \,\mathrm{d}\nu + \int_{\mathcal{X}_j^{\text{near}}} (1-\eta_j)\,\mathrm{d}\hat{\mu} - \sum_{\ell\neq j} \int_{\mathcal{X}_\ell^{\text{near}}} \eta_j \,\mathrm{d}\hat{\mu} - \int_{\mathcal{X}}^{\text{far}} \eta_j \,\mathrm{d}\hat{\mu}.$$

3. Show that $\left|\int \eta_j \,\mathrm{d}\nu\right| \leqslant 2\rho\,\|p_j\|_{\mathbb{H}}$. Bound the three other terms with L, and prove the inequality of Remark 4.4.

4. Let $\mathcal{X} = \mathbb{R}$, assume $\hat{A}_j \overset{\text{def.}}{=} |\hat{\mu}|\,(\mathcal{X}_j^{\text{near}}) > 0$, and let $\bar{x}_j \overset{\text{def.}}{=} \hat{A}_j^{-1} \int_{\mathcal{X}_j^{\text{near}}} x \,\mathrm{d}\,|\hat{\mu}|\,(x)$. Show that $\mathfrak{d}_\mathfrak{g}(\bar{x}_j, x_j)^2 \leqslant \hat{A}_j^{-1} \int_{\mathcal{X}_j^{\text{near}}} \mathfrak{d}_\mathfrak{g}(x, x_j)^2 \,\mathrm{d}\,|\hat{\mu}|\,(x)$.

5. In question 4, assume $\left|\hat{\mu}(\mathcal{X}_j^{\text{near}}) - a_j\right| \leqslant |a_j|/2$. Show that $\mathfrak{d}_\mathfrak{g}(\bar{x}_j, x_j)^2 \leqslant 2L/(\varepsilon_2\,|a_j|)$. With Corollary 4.1, what is the order of $|\bar{x}_j - x_j|$ in terms of δ? Compare with the rate of Theorem 4.2. Which result needs a small noise?

Exercise 4.3 (One spike at low noise)**.** Assume the setting of Section 4.3 with $\|\varphi\|_{\mathrm{L}^2} = 1$, and let $\kappa(u) \overset{\text{def.}}{=} \int_{\mathbb{T}} \varphi(t)\varphi(t+u)\,\mathrm{d}t$. Assume $|\kappa(u)| < 1$ for $u \neq 0$ and $\kappa''(0) < 0$. Let $\mu^0 = a\delta_{x_1}$ with $a \neq 0$, and $\sigma \overset{\text{def.}}{=} \operatorname{sign}(a)$.

1. Check that $K(x,x') = \kappa(x-x')$, that κ is even and that $\kappa(0) = 1$. Deduce from Example 3.1 that (NDSC) holds and that $p_0 = \sigma\varphi_{x_1}$.

2. Let $0 < \lambda < |a|$. With Exercise 2.5, show that $\hat{p}_\lambda(y_0) = p_0$. What is the dual bias of Proposition 4.2?

3. Let $t \in \mathbb{R}$ and $w \overset{\text{def.}}{=} t\varphi_{x_1}$. For $0 < \lambda < |a+t|$ and $\operatorname{sign}(a+t) = \sigma$, compute the solution of $\mathcal{P}_\lambda(y_0 + w)$.

4. Take $t = -\sigma\,|t|$ with $|t|$ small, and $\lambda = \|w\|_{\mathbb{H}}/\alpha$. Show that $|\hat{a}_\lambda - a| = (1 + 1/\alpha)\,\|w\|_{\mathbb{H}}$. What does this say about the constant in the rate of Theorem 4.2?

Exercise 4.4 (Fourier measurements as a convolution)**.** Let Φ, $\mathbb{H} = \mathbb{C}^{2f_c+1}$ and Λ be as in Example 2.1, and let $\varphi(t) \overset{\text{def.}}{=} \sum_{|\omega|\leqslant f_c} \sqrt{\Lambda(\omega)}\, e^{2\mathrm{i}\pi\omega t}$.

1. Show that φ is real, even and $\mathscr{C}^\infty$ on $\mathbb{T}$.

2. For $u \in \mathbb{H}$, let $(Uu)(t) \overset{\text{def.}}{=} \sum_\omega u_{-\omega} e^{2\mathrm{i}\pi\omega t}$, a complex function on $\mathbb{T}$. Show that U is a linear isometry from $\mathbb{H}$ onto the trigonometric polynomials of degree at most f_c, with the scalar product $\Re\int_{\mathbb{T}} f\,\bar{g}\,\mathrm{d}t$. Show that $U\Phi\mu = \varphi \star \mu$ for every real measure μ.

3. Let $y \in \mathbb{H}$. Show that $\|\Phi\mu - y\|^2_{\mathbb{H}} = \|\varphi\star\mu - \Re(Uy)\|^2_{\mathrm{L}^2} + \|\operatorname{Im}(Uy)\|^2_{\mathrm{L}^2}$ for every real measure μ. Deduce that $\mathcal{P}_\lambda(y)$ and the BLASSO with the convolution by φ and the data $\Re(Uy)$ have the same solutions, and that $\|\Re(Uw)\|_{\mathrm{L}^2} \leqslant \|w\|_{\mathbb{H}}$.

4. Show that the kernel of the convolution by φ is the kernel K of Example 2.1. Deduce that $\Gamma^\star_{x_0}\Gamma_{x_0}$, η_V and the test of Remark 3.5 are the same for both operators. Conclude that Theorem 4.2 applies to the weighted Fourier measurements under the assumptions of Corollary 4.3, with the noise $\Re(Uw)$.

Exercise 4.5 (Noise level and regimes for the Gaussian mixture)**.** Let w_n be as in Lemma 4.1.

1. Show that $\|\Phi\mu^0\|^2_{\mathbb{H}} = \sum_{i,j} a_i a_j e^{-(x_i-x_j)^2/(6\sigma^2)}$ and that $1/s \leqslant \|\Phi\mu^0\|^2_{\mathbb{H}} \leqslant 1$. Deduce from (4.3) that $(\sqrt{3}-1)/n \leqslant \mathbb{E}\,\|w_n\|^2_{\mathbb{H}} \leqslant (\sqrt{3}-1/s)/n$.

2. With Markov's inequality applied to $\|w_n\|^2_{\mathbb{H}}$, show that $\|w_n\|_{\mathbb{H}} \leqslant 3^{1/4}/\sqrt{n\zeta}$ with probability at least $1-\zeta$. Compare the dependence on ζ with that of Lemma 4.1.

3. In Corollary 4.4, show that $\|w_n\|_{\mathbb{H}} \leqslant \alpha\lambda$ on the event of the proof, with $\alpha = \sqrt{s/(1-2h)}$. Why does this not place (λ, w_n) in a low-noise regime D_{α,λ_0} with α small?

4. Let $\eta = \Phi^\star p$ be a certificate, $\|w\|_{\mathbb{H}} \leqslant \delta$ and $\lambda = \delta/\alpha$ with $\alpha > 0$. Compute the bound $\rho^2/(2\lambda)$ of Proposition 4.1. Compare it with $2\delta\,\|p\|_{\mathbb{H}}$ when α is small, and explain the price of the low-noise regime for the localisation bound.

Chapter 5

The generalised BLASSO and Fenchel-Young losses

Lectures 2 to 4 use the quadratic fidelity $\frac{1}{2}\|\Phi\mu - y\|_{\mathbb{H}}^2$. In this lecture we keep the measures, the total variation and the operator Φ, and we change the fidelity. The certificates keep their form $\eta = \Phi^\star p$. What changes is the link between the dual vector and the prediction $\Phi\mu$: it becomes the gradient of the fidelity. The main tool is an identity. The duality gap is the sum of two nonnegative losses, one for the fit and one for the regulariser. The bounds of Lectures 2 and 4 are special cases of this identity.

Throughout, $\mathcal{X}$, $\mathbb{H}$ and Φ are as in Definition 2.2, and the assumption of Proposition 2.1 holds. So $\Phi^\star h \in C_0(\mathcal{X})$ for every $h \in \mathbb{H}$, and $\langle \Phi\mu,\, h\rangle_{\mathbb{H}} = \langle \Phi^\star h,\, \mu\rangle_{\mathcal{M}}$ for all $\mu \in \mathcal{M}(\mathcal{X})$ by (2.8). The letter h now denotes a vector of $\mathbb{H}$ that plays the role of a dual variable. It is not the error $\Phi(\hat{\mu} - \mu^0)$ of Lecture 4.

5.1 Beyond the quadratic fidelity

The quadratic fidelity is the negative log-likelihood of a Gaussian noise. Other data call for other fidelities. Take $\mathbb{H} = \mathbb{R}^m$ and data $y \in \mathbb{R}^m$.

- Photon counts, $y_i \in \mathbb{N}$ with a Poisson law of mean $e^{(\Phi\mu^0)_i}$: $F(h) = \sum_i (e^{h_i} - y_i h_i)$.
- Outliers: $F(h) = \sum_i \rho_\delta(h_i - y_i)$ with the Huber function $\rho_\delta(t) = t^2/2$ for $|t| \leqslant \delta$ and $\rho_\delta(t) = \delta\,|t| - \delta^2/2$ otherwise.
- Labels $y_i \in \{-1, 1\}$: $F(h) = \sum_i \log(1 + e^{-y_i h_i})$, the logistic loss.

In each case the estimator minimises $F(\Phi\mu) + \lambda\|\mu\|_{\mathrm{TV}}$.

Definition 5.1 (Generalised BLASSO)**.** Let $F : \mathbb{H} \to \mathbb{R}$ be continuously differentiable, strictly convex and bounded from below. Let $R : \mathcal{M}(\mathcal{X}) \to \mathbb{R} \cup \{+\infty\}$ be proper, convex, weak-$\star$ lower semi-continuous, with sublevel sets bounded in total variation. For $\lambda > 0$ the generalised BLASSO is

$$\hat{\mu}_\lambda \in \operatorname*{argmin}_{\mu \in \mathcal{M}(\mathcal{X})} J(\mu), \qquad J(\mu) \stackrel{\text{def.}}{=} F(\Phi\mu) + \lambda R(\mu). \tag{5.1}$$

In this course $R = \|\cdot\|_{\mathrm{TV}}$. The BLASSO is the case $F(h) = \frac{1}{2}\|h - y\|_{\mathbb{H}}^2$.

Proper means that R is finite at some measure. We write $\operatorname{dom} R \stackrel{\text{def.}}{=} \{\mu\ ;\ R(\mu) < +\infty\}$ and b for a lower bound of F. The total variation satisfies the assumptions on R: it is finite and convex, it is weak-$\star$ lower semi-continuous (first step of the proof of Proposition 2.2), and its sublevel sets are balls. The quadratic fidelity is $\mathscr{C}^1$ with $\nabla F(h) = h - y$, nonnegative, and

strictly convex. The Poisson and logistic fidelities satisfy the assumptions on F (Section 5.5). The Huber fidelity is not strictly convex (Exercise 5.4).

In general R is bounded from below. Let $\nu_0 \in \operatorname{dom} R$ and $S \overset{\text{def.}}{=} \{\mu \,;\, R(\mu) \leqslant R(\nu_0)\}$. The set S lies in a ball of $\mathcal{M}(\mathcal{X})$, which is a compact metric space for the weak-$\star$ topology (Theorem 2.2). Let (μ_n) be a sequence of S with $R(\mu_n) \to \inf_S R$. A subsequence converges weak-$\star$ to some $\bar{\mu}$, and lower semi-continuity gives $R(\bar{\mu}) \leqslant \inf_S R$. Since $R(\bar{\mu}) > -\infty$, $\inf_S R$ is finite. Outside S, $R > R(\nu_0)$. Hence $\inf R = \inf_S R > -\infty$, and $J \geqslant b + \lambda \inf R$. So $\inf J$ is finite.

Remark 5.1 (Existence of a solution). Problem (5.1) has a solution. The proof is that of Proposition 2.2, with two changes. First, if $J(\mu) \leqslant c$, then $R(\mu) \leqslant (c-b)/\lambda$, so the sublevel sets of J are bounded in total variation. Second, the fidelity term is weak-$\star$ lower semi-continuous. Let $\mu_n \overset{\star}{\rightharpoonup} \mu$. By Proposition 2.1, $\Phi\mu_n \rightharpoonup \Phi\mu$ weakly in $\mathbb{H}$. The gradient inequality (5.2) below gives $F(\Phi\mu_n) \geqslant F(\Phi\mu) + \langle \nabla F(\Phi\mu), \Phi\mu_n - \Phi\mu \rangle_{\mathbb{H}}$, and the last term tends to 0. So $F(\Phi\mu) \leqslant \liminf_n F(\Phi\mu_n)$. Strict convexity is not used.

5.2 Fenchel-Young losses

The conjugate of a convex function describes it by its affine minorants. The Fenchel-Young loss measures how far a primal point and a dual point are from being paired by the gradient. In machine learning these losses are used as prediction losses [20]. Here they are the two halves of a duality gap.

Definition 5.2 (Conjugates and Fenchel-Young losses)**.** For $h, h' \in \mathbb{H}$, $f \in C_0(\mathcal{X})$ and $\mu \in \mathcal{M}(\mathcal{X})$,

$$F^\star(h) \overset{\text{def.}}{=} \sup_{h'} \langle h, h' \rangle_{\mathbb{H}} - F(h'), \quad L_F(h' \,\|\, h) \overset{\text{def.}}{=} F(h') + F^\star(h) - \langle h, h' \rangle_{\mathbb{H}},$$
$$R^\star(f) \overset{\text{def.}}{=} \sup_{\mu} \langle f, \mu \rangle_{\mathcal{M}} - R(\mu), \quad L_R(\mu \,\|\, f) \overset{\text{def.}}{=} R(\mu) + R^\star(f) - \langle f, \mu \rangle_{\mathcal{M}}.$$

Both losses take their values in $\mathbb{R}_+ \cup \{+\infty\}$.

The conjugates never take the value $-\infty$: $F^\star(h) \geqslant \langle h, 0 \rangle_{\mathbb{H}} - F(0)$ and $R^\star(f) \geqslant \langle f, \nu_0 \rangle_{\mathcal{M}} - R(\nu_0)$ for $\nu_0 \in \operatorname{dom} R$. So L_F and L_R are well defined in $\mathbb{R} \cup \{+\infty\}$. We have $L_F(h' \,\|\, h) = +\infty$ if and only if $F^\star(h) = +\infty$, and $L_R(\mu \,\|\, f) = +\infty$ if and only if $R(\mu) = +\infty$ or $R^\star(f) = +\infty$. For every h', the definition of the supremum gives $F^\star(h) \geqslant \langle h, h' \rangle_{\mathbb{H}} - F(h')$, that is $L_F(h' \,\|\, h) \geqslant 0$. This is the Fenchel-Young inequality. The same argument gives $L_R \geqslant 0$.

Subdifferentials are defined as in Proposition 2.3. For $h' \in \mathbb{H}$, $h \in \partial F(h')$ means $F(u) \geqslant F(h') + \langle h, u - h' \rangle_{\mathbb{H}}$ for all $u \in \mathbb{H}$. For $\mu \in \mathcal{M}(\mathcal{X})$, $f \in \partial R(\mu)$ means $f \in C_0(\mathcal{X})$, $R(\mu) < \infty$ and $R(\nu) \geqslant R(\mu) + \langle f, \nu - \mu \rangle_{\mathcal{M}}$ for all $\nu \in \mathcal{M}(\mathcal{X})$. For $F^\star : \mathbb{H} \to \mathbb{R} \cup \{+\infty\}$, $h' \in \partial F^\star(h)$ means $F^\star(h) < \infty$ and $F^\star(k) \geqslant F^\star(h) + \langle h', k - h \rangle_{\mathbb{H}}$ for all $k \in \mathbb{H}$.

Lemma 5.1 (Zero loss)**.** *$L_F(h' \,\|\, h) = 0$ if and only if $h \in \partial F(h')$, if and only if $h' \in \partial F^\star(h)$. In the same way $L_R(\mu \,\|\, f) = 0$ if and only if $f \in \partial R(\mu)$.*

Since F is differentiable, $\partial F(h') = \{\nabla F(h')\}$ (see the proof). So $L_F(h' \,\|\, h) = 0$ if and only if $h = \nabla F(h')$. The map ∇F is called the mirror map.

Example 5.1 (Least squares and total variation). For $F(h) = \frac{1}{2}\|h - y\|^2$, $F^\star(h) = \frac{1}{2}\|h\|^2 + \langle h, y \rangle$ and $L_F(h' \,\|\, h) = \frac{1}{2}\|h' - y - h\|^2$. For $R = \|\cdot\|_{\mathrm{TV}}$, $R^\star$ is zero on the unit ball of $C_0(\mathcal{X})$ and $+\infty$ outside, so $L_R(\mu \,\|\, f) = \|\mu\|_{\mathrm{TV}} - \langle f, \mu \rangle_{\mathcal{M}}$ when $\|f\|_\infty \leqslant 1$.

Remark 5.2 (Bregman divergence). If $h = \nabla F(h'')$ then $L_F(h' \,\|\, h) = F(h') - F(h'') - \langle \nabla F(h''), h' - h'' \rangle_{\mathbb{H}}$.

Proof of Lemma 5.1. All scalar products are those of $\mathbb{H}$.

Zero loss and subgradient. Assume $L_F(h' \,\|\, h) = 0$. Then $F^\star(h)$ is finite and equals $\langle h, h'\rangle - F(h')$. By definition of $F^\star$, $\langle h, u\rangle - F(u) \leqslant \langle h, h'\rangle - F(h')$ for all $u \in \mathbb{H}$, that is $F(u) \geqslant F(h') + \langle h, u - h'\rangle$: this is $h \in \partial F(h')$. Conversely, if $h \in \partial F(h')$, the same inequality says that h' attains the supremum that defines $F^\star(h)$. So $F^\star(h) = \langle h, h'\rangle - F(h')$ and $L_F(h' \,\|\, h) = 0$.

The subgradient is the gradient. Let $u \in \mathbb{H}$ and $t \in (0, 1]$. Convexity gives $F(h' + t(u - h')) \leqslant (1-t)F(h') + tF(u)$, so $(F(h' + t(u - h')) - F(h'))/t \leqslant F(u) - F(h')$. When $t \to 0$ the left-hand side tends to $\langle \nabla F(h'), u - h'\rangle$. So

$$F(u) \geqslant F(h') + \langle \nabla F(h'), u - h'\rangle \qquad \text{for all } u, h' \in \mathbb{H}, \tag{5.2}$$

that is $\nabla F(h') \in \partial F(h')$. Conversely, let $h \in \partial F(h')$ and $v \in \mathbb{H}$. For $t > 0$, $F(h' + tv) - F(h') \geqslant t\langle h, v\rangle$. Divide by t and let $t \to 0$: $\langle \nabla F(h'), v\rangle \geqslant \langle h, v\rangle$. With $-v$ in place of v we get equality for every v, so $h = \nabla F(h')$.

The conjugate. Assume $L_F(h' \,\|\, h) = 0$. For $k \in \mathbb{H}$, the definition of $F^\star$ gives $F^\star(k) \geqslant \langle k, h'\rangle - F(h') = F^\star(h) + \langle h', k - h\rangle$, so $h' \in \partial F^\star(h)$. Conversely, let $h' \in \partial F^\star(h)$ and $k \overset{\text{def.}}{=} \nabla F(h')$. By the first two points, $L_F(h' \,\|\, k) = 0$, that is $F^\star(k) = \langle k, h'\rangle - F(h')$. The subgradient inequality at this k reads $\langle k, h'\rangle - F(h') \geqslant F^\star(h) + \langle h', k - h\rangle$, that is $F^\star(h) \leqslant \langle h, h'\rangle - F(h')$. So $L_F(h' \,\|\, h) \leqslant 0$, and $L_F(h' \,\|\, h) = 0$.

The regulariser. If $R(\mu) = +\infty$, then $L_R(\mu \,\|\, f) = +\infty$ and $\partial R(\mu)$ is empty. If $R(\mu) < \infty$, the first point holds word for word with R, $\langle \cdot, \cdot\rangle_{\mathcal{M}}$ and $\mathcal{M}(\mathcal{X})$ in place of F, the scalar product and $\mathbb{H}$. □

The proof of the equivalence with $h' \in \partial F^\star(h)$ uses the differentiability of F, and no biconjugation theorem. For R, the equivalence $L_R(\mu \,\|\, f) = 0 \Leftrightarrow \mu \in \partial R^\star(f)$ also holds, where $\partial R^\star(f)$ is a set of measures. Its proof uses the biconjugation theorem $R^{\star\star} = R$ on $\mathcal{M}(\mathcal{X})$, which we admit [15]. We do not use it.

Computations of Example 5.1. Conjugate. With $u \overset{\text{def.}}{=} h' - y$, $\langle h, h'\rangle - \frac{1}{2}\|h' - y\|^2 = \langle h, y\rangle + \langle h, u\rangle - \frac{1}{2}\|u\|^2 = \langle h, y\rangle + \frac{1}{2}\|h\|^2 - \frac{1}{2}\|u - h\|^2$. The supremum over u is attained at $u = h$, so $F^\star(h) = \frac{1}{2}\|h\|^2 + \langle h, y\rangle$.

Loss. Expanding, $L_F(h' \,\|\, h) = \frac{1}{2}\|h' - y\|^2 + \frac{1}{2}\|h\|^2 - \langle h, h' - y\rangle = \frac{1}{2}\|h' - y - h\|^2$.

Total variation. The conjugate of the total variation is computed after the proof of Proposition 2.3: $R^\star(f) = 0$ if $\|f\|_\infty \leqslant 1$ and $+\infty$ otherwise. So $L_R(\mu \,\|\, f) = \|\mu\|_{\mathrm{TV}} - \langle f, \mu\rangle_{\mathcal{M}}$ if $\|f\|_\infty \leqslant 1$, and $+\infty$ otherwise. □

The example agrees with Lemma 5.1. The quadratic loss vanishes if and only if $h = h' - y = \nabla F(h')$, the residual. The total variation loss vanishes if and only if $\|f\|_\infty \leqslant 1$ and $\langle f, \mu\rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$, that is $f \in \partial\|\mu\|_{\mathrm{TV}}$ by Proposition 2.3. By (2.6), for $\|f\|_\infty \leqslant 1$, $L_R(\mu \,\|\, f) \geqslant \int (1 - |f|)\,\mathrm{d}\,|\mu|$.

Proof of Remark 5.2. By Lemma 5.1, $L_F(h'' \,\|\, h) = 0$, so $F^\star(h) = \langle h, h''\rangle_{\mathbb{H}} - F(h'')$. Substituting in the definition, $L_F(h' \,\|\, h) = F(h') + \langle h, h''\rangle_{\mathbb{H}} - F(h'') - \langle h, h'\rangle_{\mathbb{H}} = F(h') - F(h'') - \langle \nabla F(h''), h' - h''\rangle_{\mathbb{H}}$. □

The right-hand side is the Bregman divergence of F between h' and h''. Strict convexity makes the mirror map injective. If $\nabla F(u) = \nabla F(v) = h$, then u and v both maximise the strictly concave function $\langle h, \cdot\rangle_{\mathbb{H}} - F$ (first point of the proof), so $u = v$. Hence h'' is determined by h. The loss $L_F(h' \,\|\, h)$ is also defined when h is not a gradient, for instance for the dual vectors $h = -\lambda p$ of Section 5.4. This is why we work with L_F and not with the Bregman divergence.

5.3 The duality gap

The dual problem below is the Fenchel-Rockafellar dual of (5.1) (Remark 2.9). We do not use the Fenchel-Rockafellar theorem. We only use the next identity, whose proof is algebra, and the weak duality that follows from it. For the BLASSO the dual vector is $h = -\lambda p$, where p is the dual variable of Proposition 2.4.

Theorem 5.1 (Duality gap identity)**.** *The dual problem of* (5.1) *is*

$$\sup_{h \in \mathbb{H}} J_{\mathbb{H}}(h), \qquad J_{\mathbb{H}}(h) \overset{\text{def.}}{=} -F^\star(h) - \lambda R^\star(-\Phi^\star h/\lambda). \tag{5.3}$$

For all μ with $R(\mu) < \infty$ and all h with $F^\star(h) < \infty$ and $R^\star(\eta) < \infty$, where $\eta \overset{\text{def.}}{=} -\Phi^\star h/\lambda$,

$$J(\mu) - J_{\mathbb{H}}(h) = L_F(\Phi\mu \,\|\, h) + \lambda L_R(\mu \,\|\, \eta) \geqslant 0. \tag{5.4}$$

Corollary 5.1 (Stopping criterion)**.** $0 \leqslant J(\mu) - \inf J \leqslant L_F(\Phi\mu \,\|\, h) + \lambda L_R(\mu \,\|\, \eta)$. *If the right-hand side is at most ε at an iterate (μ_k, h_k) of an algorithm, then μ_k is ε-optimal.*

In the corollary, $\mu \in \mathcal{M}(\mathcal{X})$ and $h \in \mathbb{H}$ are arbitrary, $\eta = -\Phi^\star h/\lambda$, and ε-optimal means $J(\mu_k) \leqslant \inf J + \varepsilon$. If (μ, h) is not as in the theorem, the right-hand side is $+\infty$ and the bound says nothing. The right-hand side involves μ and h only, not the unknown value $\inf J$: it can be computed.

Proof of Theorem 5.1. All the terms are finite. Indeed $F(\Phi\mu) \in \mathbb{R}$, $R(\mu) \in \mathbb{R}$ by assumption, and the conjugates $F^\star(h)$ and $R^\star(\eta)$ are finite by assumption and larger than $-\infty$ (Section 5.2). So $J(\mu)$ and $J_{\mathbb{H}}(h)$ are real numbers, and

$$J(\mu) - J_{\mathbb{H}}(h) = F(\Phi\mu) + F^\star(h) + \lambda R(\mu) + \lambda R^\star(\eta).$$

By (2.8) and the definition of η, $\langle h, \Phi\mu\rangle_{\mathbb{H}} = \langle \Phi^\star h, \mu\rangle_{\mathcal{M}} = -\lambda\langle \eta, \mu\rangle_{\mathcal{M}}$. So we can subtract $\langle h, \Phi\mu\rangle_{\mathbb{H}}$ from the first two terms and add it, in the form $-\lambda\langle \eta, \mu\rangle_{\mathcal{M}}$, to the last two:

$$J(\mu) - J_{\mathbb{H}}(h) = \big[F(\Phi\mu) + F^\star(h) - \langle h, \Phi\mu\rangle_{\mathbb{H}}\big] + \lambda\big[R(\mu) + R^\star(\eta) - \langle \eta, \mu\rangle_{\mathcal{M}}\big].$$

The two brackets are $L_F(\Phi\mu \,\|\, h)$ and $L_R(\mu \,\|\, \eta)$. Both are nonnegative by the Fenchel-Young inequality, and $\lambda > 0$. □

Proof of Corollary 5.1. Since $\inf J$ is finite (Section 5.1), $J(\mu) - \inf J \geqslant 0$ is well defined. If the right-hand side is $+\infty$, there is nothing more to prove. Otherwise $R(\mu) < \infty$, $F^\star(h) < \infty$ and $R^\star(\eta) < \infty$. Let $\nu \in \mathcal{M}(\mathcal{X})$. If $R(\nu) = +\infty$, then $J(\nu) = +\infty \geqslant J_{\mathbb{H}}(h)$. Otherwise Theorem 5.1 applied to (ν, h) gives $J(\nu) \geqslant J_{\mathbb{H}}(h)$. This is weak duality: $\inf J \geqslant J_{\mathbb{H}}(h)$. Hence

$$0 \leqslant J(\mu) - \inf J \leqslant J(\mu) - J_{\mathbb{H}}(h) = L_F(\Phi\mu \,\|\, h) + \lambda L_R(\mu \,\|\, \eta).$$

If the right-hand side is at most ε at (μ_k, h_k), then $J(\mu_k) \leqslant \inf J + \varepsilon$. □

The proof uses weak duality only. Strong duality, $\inf J = \sup J_{\mathbb{H}}$, says in addition that the bound can be made as small as we wish by the choice of h. Remark 5.4 below proves it in our setting.

Remark 5.3 (The BLASSO gap). Let $F(h) = \frac{1}{2}\|h - y\|_{\mathbb{H}}^2$, $R = \|\cdot\|_{\mathrm{TV}}$ and $h = -\lambda p$ with $p \in \mathbb{H}$. Then $\eta = -\Phi^\star h/\lambda = \Phi^\star p$. By Example 5.1, if $\|\Phi^\star p\|_\infty \leqslant 1$,

$$J_{\mathbb{H}}(-\lambda p) = \lambda\langle p, y\rangle_{\mathbb{H}} - \tfrac{\lambda^2}{2}\|p\|_{\mathbb{H}}^2, \qquad L_F(\Phi\mu \,\|\, -\lambda p) = \tfrac{1}{2}\|\Phi\mu - y + \lambda p\|_{\mathbb{H}}^2,$$

and $L_R(\mu \,\|\, \Phi^\star p) = \|\mu\|_{\mathrm{TV}} - \langle \Phi^\star p, \mu\rangle_{\mathcal{M}}$. So $J_{\mathbb{H}}(-\lambda p)$ is the dual function $D(p)$ of Proposition 2.4, and (5.4) is the identity (2.10). If $\|\Phi^\star p\|_\infty > 1$, then $J_{\mathbb{H}}(-\lambda p) = -\infty$: the constraint $p \in C$ of Proposition 2.4 is the domain of $R^\star$. This is the gap used as a stopping test in Section 6.1.

Definition 5.3 (Source condition for the generalised BLASSO)**.** A function $\eta \in C_0(\mathcal{X})$ is a dual certificate of μ if there is $h \in \mathbb{H}$ with $\eta = -\Phi^\star h/\lambda \in \partial R(\mu)$.

For $R = \|\cdot\|_{\mathrm{TV}}$ and $h = -\lambda p$, the condition reads $\Phi^\star p \in \partial\|\mu\|_{\mathrm{TV}}$. So η is a dual certificate in the sense of Definition 2.5, and for $\mu = \mu^0$ this is the source condition (SC) of Definition 3.1. The condition does not involve F: the certificates built in Lecture 3 serve for every fidelity. The fidelity enters through the second condition of the next theorem, which links h to the prediction $\Phi\mu$.

Theorem 5.2 (Certificate and mirror alignment)**.** *Let μ have a dual certificate $\eta = -\Phi^\star h/\lambda$. Then $0 \leqslant J(\mu) - \inf J \leqslant L_F(\Phi\mu \,\|\, h)$. If in addition $h = \nabla F(\Phi\mu)$, then μ is a solution of (5.1) and h is a solution of (5.3). Conversely every solution $\hat\mu$ has a certificate with $\hat h = \nabla F(\Phi\hat\mu)$.*

The condition $h = \nabla F(\Phi\mu)$ is called mirror alignment. By Lemma 5.1, it is equivalent to $L_F(\Phi\mu \,\|\, h) = 0$ and to $\Phi\mu \in \partial F^\star(h)$.

Proof. Bound. Since $\eta \in \partial R(\mu)$, Lemma 5.1 gives $L_R(\mu \,\|\, \eta) = 0$. Corollary 5.1 gives $0 \leqslant J(\mu) - \inf J \leqslant L_F(\Phi\mu \,\|\, h)$.

Mirror alignment. Let $h = \nabla F(\Phi\mu)$. By Lemma 5.1, $L_F(\Phi\mu \,\|\, h) = 0$, so $F^\star(h) < \infty$, and the bound gives $J(\mu) = \inf J$: μ is a solution. Also $R(\mu) < \infty$ and $R^\star(\eta) = \langle \eta,\, \mu\rangle_{\mathcal{M}} - R(\mu) < \infty$, so Theorem 5.1 applies to (μ, h) and gives $J(\mu) = J_{\mathbb{H}}(h)$. For every $k \in \mathbb{H}$, weak duality (proof of Corollary 5.1) gives $J_{\mathbb{H}}(k) \leqslant \inf J = J(\mu) = J_{\mathbb{H}}(h)$; this is clear if $J_{\mathbb{H}}(k) = -\infty$. So h is a solution of (5.3).

Converse. Let $\hat\mu$ be a solution, $\hat h \overset{\text{def.}}{=} \nabla F(\Phi\hat\mu)$ and $\hat\eta \overset{\text{def.}}{=} -\Phi^\star\hat h/\lambda \in C_0(\mathcal{X})$. Since $J(\hat\mu) \leqslant J(\nu_0)$ for $\nu_0 \in \operatorname{dom} R$, $R(\hat\mu) < \infty$. Let $\nu \in \mathcal{M}(\mathcal{X})$ with $R(\nu) < \infty$, and $\tau \in (0, 1]$. By convexity of R, $R(\hat\mu + \tau(\nu - \hat\mu)) - R(\hat\mu) \leqslant \tau\big(R(\nu) - R(\hat\mu)\big)$. Since F is differentiable at $\Phi\hat\mu$, $F(\Phi\hat\mu + \tau\Phi(\nu - \hat\mu)) - F(\Phi\hat\mu) = \tau\langle\hat h,\, \Phi(\nu - \hat\mu)\rangle_{\mathbb{H}} + \tau\varepsilon(\tau)$ with $\varepsilon(\tau) \to 0$ when $\tau \to 0$. The optimality $J(\hat\mu + \tau(\nu - \hat\mu)) \geqslant J(\hat\mu)$ gives, after division by τ,

$$0 \leqslant \langle\hat h,\, \Phi(\nu - \hat\mu)\rangle_{\mathbb{H}} + \varepsilon(\tau) + \lambda\big(R(\nu) - R(\hat\mu)\big).$$

Let $\tau \to 0$ and use (2.8): $R(\nu) \geqslant R(\hat\mu) - \langle\Phi^\star\hat h,\, \nu - \hat\mu\rangle_{\mathcal{M}}/\lambda = R(\hat\mu) + \langle\hat\eta,\, \nu - \hat\mu\rangle_{\mathcal{M}}$. This also holds when $R(\nu) = +\infty$. So $\hat\eta \in \partial R(\hat\mu)$: it is a certificate of $\hat\mu$, with $\hat h = \nabla F(\Phi\hat\mu)$. □

For the BLASSO, $\hat h = \Phi\hat\mu - y = -\lambda\hat p_\lambda$ and $\hat\eta = \Phi^\star\hat p_\lambda$ with $\hat p_\lambda = (y - \Phi\hat\mu)/\lambda$: the converse is item 1 of Proposition 2.4. The theorem splits the first order condition $0 \in \Phi^\star\nabla F(\Phi\hat\mu) + \lambda\partial R(\hat\mu)$ into a half for the regulariser, the source condition, and a half for the loss, mirror alignment.

Remark 5.4 (Consequences). Let $\hat\mu$ be a solution, which exists by Remark 5.1, and $\hat h \overset{\text{def.}}{=} \nabla F(\Phi\hat\mu)$. (a) Strong duality. By Theorem 5.2, $\hat h$ is a solution of (5.3) and $\min J = J(\hat\mu) = J_{\mathbb{H}}(\hat h) = \max J_{\mathbb{H}}$. (b) The dual solution is unique. Let h be a solution of (5.3). Then $J_{\mathbb{H}}(h) = J(\hat\mu)$ is finite, so Theorem 5.1 applies at $(\hat\mu, h)$, and the right-hand side of (5.4) is a sum of two nonnegative terms equal to zero. So $L_F(\Phi\hat\mu \,\|\, h) = 0$, and $h = \hat h$ by Lemma 5.1. (c) The prediction is unique. Let $\hat\mu'$ be another solution. By (b) applied to $\hat\mu'$, $\nabla F(\Phi\hat\mu') = \hat h = \nabla F(\Phi\hat\mu)$. Since ∇F is injective, $\Phi\hat\mu' = \Phi\hat\mu$. This is the only use of the strict convexity of F. (d) For $R = \|\cdot\|_{\mathrm{TV}}$, Remark 2.6 (ii) applied to $\hat\eta$ gives $\operatorname{supp}(\hat\mu) \subset \{x \in \mathcal{X}\ ;\ |\hat\eta(x)| = 1\}$. For the BLASSO, (a) to (d) are items 2 to 4 of Proposition 2.4.

5.4 Error bounds

In this section we fix a measure $\mu^\star$ with a certificate, in practice the target μ^0, and we compare every candidate μ with it. The regulariser loss becomes a Bregman divergence, and the duality gap becomes an error bound.

Proposition 5.1 (Source condition and Bregman divergence)**.** *Let* $\eta^\star = -\Phi^\star h^\star/\lambda \in \partial R(\mu^\star)$. *For all* μ, $L_R(\mu \,\|\, \eta^\star) = R(\mu) - R(\mu^\star) - \langle \eta^\star, \mu - \mu^\star \rangle_{\mathcal{M}}$.

Theorem 5.3 (Fenchel-Young error bound)**.** *Let* $\eta^\star = -\Phi^\star h^\star/\lambda \in \partial R(\mu^\star)$ *with* $F^\star(h^\star) < \infty$, *and let* μ *satisfy* $J(\mu) \leqslant J(\mu^\star)$. *Then*

$$L_F(\Phi\mu \,\|\, h^\star) + \lambda(R(\mu) - R(\mu^\star) - \langle \eta^\star, \mu - \mu^\star \rangle_{\mathcal{M}}) \leqslant L_F(\Phi\mu^\star \,\|\, h^\star). \tag{5.5}$$

Proof of Proposition 5.1. By Lemma 5.1, $L_R(\mu^\star \,\|\, \eta^\star) = 0$. So $R(\mu^\star)$ is finite and $R^\star(\eta^\star) = \langle \eta^\star, \mu^\star \rangle_{\mathcal{M}} - R(\mu^\star)$ is finite. Substituting in the definition of L_R,

$$L_R(\mu \,\|\, \eta^\star) = R(\mu) + \langle \eta^\star, \mu^\star \rangle_{\mathcal{M}} - R(\mu^\star) - \langle \eta^\star, \mu \rangle_{\mathcal{M}} = R(\mu) - R(\mu^\star) - \langle \eta^\star, \mu - \mu^\star \rangle_{\mathcal{M}}.$$

Both sides are $+\infty$ when $R(\mu) = +\infty$. □

The right-hand side is the Bregman divergence of R between μ and $\mu^\star$, for the subgradient $\eta^\star$. It is nonnegative by the subgradient inequality. For $R = \|\cdot\|_{\mathrm{TV}}$, $\langle \eta^\star, \mu^\star \rangle_{\mathcal{M}} = \|\mu^\star\|_{\mathrm{TV}}$, so it equals $\|\mu\|_{\mathrm{TV}} - \langle \eta^\star, \mu \rangle_{\mathcal{M}}$. This is the quantity D of Proposition 4.1 and Remark 4.1 (b). It vanishes exactly on the measures that $\eta^\star$ certifies, and by (2.6) it is at least $\int (1 - |\eta^\star|) \,\mathrm{d}\,|\mu|$.

Proof of Theorem 5.3. By the proof of Proposition 5.1, $R(\mu^\star)$ and $R^\star(\eta^\star)$ are finite. So $J(\mu^\star) < \infty$, and $J(\mu) \leqslant J(\mu^\star)$ gives $R(\mu) < \infty$. Theorem 5.1 applies to $(\mu, h^\star)$ and to $(\mu^\star, h^\star)$:

$$J(\mu) - J_{\mathbb{H}}(h^\star) = L_F(\Phi\mu \,\|\, h^\star) + \lambda L_R(\mu \,\|\, \eta^\star), \qquad J(\mu^\star) - J_{\mathbb{H}}(h^\star) = L_F(\Phi\mu^\star \,\|\, h^\star),$$

where $L_R(\mu^\star \,\|\, \eta^\star) = 0$ by Lemma 5.1. All terms are finite. Subtracting, the dual value cancels:

$$J(\mu) - J(\mu^\star) = L_F(\Phi\mu \,\|\, h^\star) + \lambda L_R(\mu \,\|\, \eta^\star) - L_F(\Phi\mu^\star \,\|\, h^\star). \tag{5.6}$$

The left-hand side is nonpositive. Proposition 5.1 replaces $L_R(\mu \,\|\, \eta^\star)$ by the Bregman divergence, and (5.5) follows. □

The identity (5.6) holds for every μ with $R(\mu) < \infty$. Both terms on the left of (5.5) are nonnegative, so each of them is at most $L_F(\Phi\mu^\star \,\|\, h^\star)$. This is the fit loss of $\mu^\star$ against the dual vector $h^\star$ of its certificate. It is zero if and only if $h^\star = \nabla F(\Phi\mu^\star)$. Every solution $\hat{\mu}$ satisfies $J(\hat{\mu}) \leqslant J(\mu^\star)$, whatever $\mu^\star$. The theorem is also in [21].

Remark 5.5 (Least squares: Proposition 4.1 again). Let $F(h) = \frac{1}{2}\|h - y\|_{\mathbb{H}}^2$ with $y = \Phi\mu^0 + w$, $R = \|\cdot\|_{\mathrm{TV}}$, $\mu^\star = \mu^0$, $\eta^\star = \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$ and $h^\star = -\lambda p$. Let $\hat{\mu}$ be a solution, $u \overset{\text{def.}}{=} \Phi(\hat{\mu} - \mu^0)$ and $D \overset{\text{def.}}{=} \|\hat{\mu}\|_{\mathrm{TV}} - \langle \eta^\star, \hat{\mu} \rangle_{\mathcal{M}}$. By Example 5.1, $L_F(\Phi\hat{\mu} \,\|\, -\lambda p) = \frac{1}{2}\|u - (w - \lambda p)\|_{\mathbb{H}}^2$ and $L_F(\Phi\mu^0 \,\|\, -\lambda p) = \frac{1}{2}\|w - \lambda p\|_{\mathbb{H}}^2$. So (5.5) reads

$$\tfrac{1}{2}\|u - (w - \lambda p)\|_{\mathbb{H}}^2 + \lambda D \leqslant \tfrac{1}{2}\|w - \lambda p\|_{\mathbb{H}}^2, \quad \text{that is} \quad \tfrac{1}{2}\|u\|_{\mathbb{H}}^2 + \lambda D \leqslant \langle u, w - \lambda p \rangle_{\mathbb{H}},$$

the first inequality of (4.2). More directly, the two nonnegative terms give $\|u - (w - \lambda p)\|_{\mathbb{H}} \leqslant \|w - \lambda p\|_{\mathbb{H}}$, so $\|u\|_{\mathbb{H}} \leqslant 2\|w - \lambda p\|_{\mathbb{H}} \leqslant 2\rho$, and $\lambda D \leqslant \frac{1}{2}\|w - \lambda p\|_{\mathbb{H}}^2 \leqslant \rho^2/2$, with $\rho = \|w\|_{\mathbb{H}} + \lambda\|p\|_{\mathbb{H}}$. These are the bounds of item 1 of Proposition 4.1.

The choice $h^\star = \nabla F(\Phi\mu^0)$ would make the right-hand side zero. For least squares it is $h^\star = -w$, and $\Phi^\star w/\lambda$ has no reason to be a certificate of μ^0. The theorem separates the two roles: the certificate is chosen for R, through p, and $h^\star = -\lambda p$ is in general not mirror aligned. The price is the fit loss of μ^0, which measures the distance between $-\lambda p$ and $\nabla F(\Phi\mu^0)$. For a general fidelity it is not a norm of $\Phi(\hat{\mu} - \mu^0)$. Strong convexity gives such a norm.

A function F is σ-strongly convex, for some $\sigma > 0$, if

$$F(u) \geqslant F(v) + \langle \nabla F(v), u - v \rangle_{\mathbb{H}} + \tfrac{\sigma}{2}\|u - v\|_{\mathbb{H}}^2 \qquad \text{for all } u, v \in \mathbb{H}. \tag{5.7}$$

The quadratic fidelity is 1-strongly convex, with equality in (5.7). A strongly convex function is strictly convex: for $u \neq v$, $t \in (0,1)$ and $z = tu + (1-t)v$, apply (5.7) at z to u and to v, and add t times the first inequality to $1-t$ times the second; the gradient terms cancel and $tF(u) + (1-t)F(v) \geqslant F(z) + \frac{\sigma}{2}(t\|u-z\|_{\mathbb{H}}^2 + (1-t)\|v-z\|_{\mathbb{H}}^2) > F(z)$. It is bounded from below: by (5.7) with $v = 0$ and the Cauchy-Schwarz inequality, $F(u) \geqslant F(0) - \|\nabla F(0)\|_{\mathbb{H}}^2/(2\sigma)$. So a $\mathscr{C}^1$ and σ-strongly convex F satisfies Definition 5.1.

Proposition 5.2 (Strongly convex fidelity)**.** *Let $R = \|\cdot\|_{\mathrm{TV}}$ and let F be σ-strongly convex. Let $\eta = \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$ and $g \overset{\text{def.}}{=} \nabla F(\Phi\mu^0)$. Every solution $\hat{\mu}$ of (5.1) satisfies*

$$\left\|\Phi(\hat{\mu} - \mu^0)\right\|_{\mathbb{H}} \leqslant \frac{2\|g + \lambda p\|_{\mathbb{H}}}{\sigma} \quad \textit{and} \quad \lambda\Big(\|\hat{\mu}\|_{\mathrm{TV}} - \int \eta \,\mathrm{d}\hat{\mu}\Big) \leqslant \frac{\|g+\lambda p\|_{\mathbb{H}}^2}{2\sigma}.$$

Assume in addition the setting of item 2 of Proposition 4.1: η is $(\varepsilon_0, \varepsilon_2, r_{\mathrm{near}})$-non-degenerate and the near regions $\mathcal{X}_i^{\mathrm{near}}$ are pairwise disjoint. Then the localisation bound (4.1) holds with $\rho^2/(2\lambda)$ replaced by $\|g+\lambda p\|_{\mathbb{H}}^2/(2\sigma\lambda)$.

The sign is that of Definition 5.3: the certificate $\Phi^\star p$ has dual vector $h^\star = -\lambda p$, and $g + \lambda p = g - h^\star$ is the gap between the mirror image of the target and this dual vector. If it vanishes, μ^0 is mirror aligned with $h^\star$, so μ^0 is a solution (Theorem 5.2), and both bounds are zero.

Proof. All norms and scalar products are those of $\mathbb{H}$. Let $u \overset{\text{def.}}{=} \Phi(\hat{\mu} - \mu^0)$ and $D \overset{\text{def.}}{=} \|\hat{\mu}\|_{\mathrm{TV}} - \langle \eta, \hat{\mu}\rangle_{\mathcal{M}}$. By Proposition 2.3, $\|\eta\|_\infty \leqslant 1$ and $\langle \eta, \mu^0\rangle_{\mathcal{M}} = \|\mu^0\|_{\mathrm{TV}}$, so $D \geqslant 0$ by (2.6).

Compare with μ^0. Since $\hat{\mu}$ is a solution, $J(\hat{\mu}) \leqslant J(\mu^0)$, that is

$$F(\Phi\hat{\mu}) - F(\Phi\mu^0) + \lambda(\|\hat{\mu}\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}}) \leqslant 0.$$

By (5.7) with $\Phi\hat{\mu}$ and $\Phi\mu^0$ in place of u and v, $F(\Phi\hat{\mu}) - F(\Phi\mu^0) \geqslant \langle g, u\rangle + \frac{\sigma}{2}\|u\|^2$.

Use the certificate. By (2.8), $\|\hat{\mu}\|_{\mathrm{TV}} - \|\mu^0\|_{\mathrm{TV}} = D + \langle \eta, \hat{\mu} - \mu^0\rangle_{\mathcal{M}} = D + \langle p, u\rangle$. Hence, with the Cauchy-Schwarz inequality,

$$\tfrac{\sigma}{2}\|u\|^2 + \lambda D \leqslant -\langle g + \lambda p, u\rangle \leqslant \|g + \lambda p\| \, \|u\|. \tag{5.8}$$

Bounds. Let $c \overset{\text{def.}}{=} \|g + \lambda p\|$. Since $\lambda D \geqslant 0$, $\frac{\sigma}{2}\|u\|^2 \leqslant c\|u\|$, so $\|u\| \leqslant 2c/\sigma$ (divide by $\|u\|$ if $u \neq 0$). Then $\lambda D \leqslant c\|u\| - \frac{\sigma}{2}\|u\|^2 \leqslant \max_{t\geqslant 0}(ct - \frac{\sigma}{2}t^2) = c^2/(2\sigma)$, the maximum being attained at $t = c/\sigma$.

Localisation. As in the proof of item 2 of Proposition 4.1, the non-degeneracy of η and the disjointness of the near regions show that D is at least the left-hand side of (4.1). Divide the bound on λD by λ. □

Remark 5.6 (On the proposition). (a) For least squares, $\sigma = 1$ and $g = \Phi\mu^0 - y = -w$. Then $\|g + \lambda p\|_{\mathbb{H}} = \|w - \lambda p\|_{\mathbb{H}} \leqslant \rho$, and the proposition gives Proposition 4.1, with the slightly better quantity $\|w - \lambda p\|_{\mathbb{H}}$ in place of ρ. Inequality (5.8) is then (4.2). (b) As in Remark 4.1 (a), the proof uses only $J(\hat{\mu}) \leqslant J(\mu^0)$. It uses (5.7) for the pair $(\Phi\hat{\mu}, \Phi\mu^0)$ only. If F is strongly convex on a convex set $K \subset \mathbb{H}$ only, the conclusions hold with the modulus on K as soon as $\Phi\hat{\mu}$ and $\Phi\mu^0$ lie in K (Exercise 5.3). (c) The vector g plays the role of the noise. For a fidelity $F = F_0 + \langle w, \cdot\rangle_{\mathbb{H}}$ with $\nabla F_0(\Phi\mu^0) = 0$, as in Section 5.6, $g = w$. With $\|w\|_{\mathbb{H}} \leqslant \delta$, $p \neq 0$ and $\lambda = \delta/\|p\|_{\mathbb{H}}$, the bounds read $\|\Phi(\hat{\mu} - \mu^0)\|_{\mathbb{H}} \leqslant 4\delta/\sigma$ and $\lambda D \leqslant 2\delta^2/\sigma$, as in Corollary 4.1 with an extra factor $1/\sigma$. (d) The second bound also follows from Theorem 5.3 with $h^\star = -\lambda p$: for a σ-strongly convex F, $L_F(\Phi\mu^0 \,\|\, -\lambda p) \leqslant \|g + \lambda p\|_{\mathbb{H}}^2/(2\sigma)$ (Exercise 5.2).

5.5 Examples

In the three examples $\mathbb{H} = \mathbb{R}^m$ with the Euclidean scalar product, and the fidelity is a sum $F(h) = \sum_{i=1}^m f_i(h_i)$ of functions of one variable. Then $F^\star(h) = \sum_i f_i^\star(h_i)$, since the supremum over $h' \in \mathbb{R}^m$ separates, and $L_F(h' \,\|\, h) = \sum_i L_{f_i}(h'_i \,\|\, h_i)$. We use the convention $0 \log 0 = 0$. The certificate $\eta = \Phi^\star p$ of μ^0 is the same for all fidelities. What changes is the dual vector given by mirror alignment, the loss L_F, and the vector $g = \nabla F(\Phi\mu^0)$ of Proposition 5.2.

Example 5.2 (Poisson counts). Let $y \in \mathbb{N}^m$ and $f_i(t) \overset{\text{def.}}{=} e^t - y_i t$, the negative log-likelihood of a Poisson count of mean e^t, up to a constant. Then $\nabla F(h) = (e^{h_i} - y_i)_i$ and

$$f_i^\star(v) = (v + y_i)\log(v + y_i) - (v + y_i) \text{ if } v + y_i \geqslant 0, \qquad f_i^\star(v) = +\infty \text{ otherwise.}$$

With $d(a, b) \overset{\text{def.}}{=} a \log(a/b) - a + b$ for $a \geqslant 0$ and $b > 0$, for $h + y \geqslant 0$,

$$L_F(h' \,\|\, h) = \sum_{i=1}^m d(h_i + y_i, e^{h'_i}),$$

the generalised Kullback-Leibler divergence between the intensities $h_i + y_i$ and $e^{h'_i}$. Mirror alignment reads $h_i + y_i = e^{(\Phi\mu)_i}$: the dual vector is the residual on the scale of the intensities.

Computations. The function f_i is $\mathscr{C}^\infty$ with $f_i'' = e^t > 0$, so F is strictly convex. It is bounded from below: by 0 if $y_i = 0$, and by its value $y_i - y_i \log y_i$ at $t = \log y_i$ otherwise. For $v \in \mathbb{R}$, the function $t \mapsto vt - f_i(t) = (v + y_i)t - e^t$ tends to $-\infty$ at $+\infty$. If $v + y_i < 0$, it tends to $+\infty$ at $-\infty$, so $f_i^\star(v) = +\infty$. If $v + y_i = 0$, the supremum of $-e^t$ is 0. If $v + y_i > 0$, it is attained where $e^t = v + y_i$, with value $(v + y_i)\log(v + y_i) - (v + y_i)$. For the loss, let $r_i \overset{\text{def.}}{=} h_i + y_i \geqslant 0$. Then $L_{f_i}(h'_i \,\|\, h_i) = e^{h'_i} - y_i h'_i + r_i \log r_i - r_i - h_i h'_i = r_i \log r_i - r_i h'_i - r_i + e^{h'_i} = d(r_i, e^{h'_i})$. □

The gradient $e^{h_i} - y_i$ is not Lipschitz, and $f_i'' = e^t$ tends to 0 at $-\infty$. So F is not strongly convex on $\mathbb{R}^m$. On the cube $\{h \,;\, \|h\|_\infty \leqslant M\}$, $f_i'' \geqslant e^{-M}$, and the Taylor formula along segments gives (5.7) with $\sigma = e^{-M}$ for u, v in the cube. If the y_i are independent with Poisson laws of means $e^{(\Phi\mu^0)_i}$, then $g = e^{\Phi\mu^0} - y$ is centred and $\mathbb{E}\,\|g\|_2^2 = \sum_i e^{(\Phi\mu^0)_i}$. Exercise 5.3 turns this into a bound.

Example 5.3 (Logistic labels). Let $y \in \{-1, 1\}^m$, $\varsigma(t) \overset{\text{def.}}{=} 1/(1 + e^{-t})$, $\psi(s) \overset{\text{def.}}{=} \log(1 + e^s)$ and $f_i(t) \overset{\text{def.}}{=} \log(1 + e^{-y_i t}) = \psi(-y_i t)$. Then $\nabla F(h) = \big(-y_i \varsigma(-y_i h_i)\big)_i$, and $F^\star(h) = \sum_{i=1}^m \psi^\star(v_i)$ with $v_i \overset{\text{def.}}{=} -y_i h_i$, where

$$\psi^\star(v) = v \log v + (1 - v)\log(1 - v) \text{ if } v \in [0, 1], \qquad \psi^\star(v) = +\infty \text{ otherwise.}$$

With $z_i \overset{\text{def.}}{=} -y_i h'_i$ and $\mathrm{KL}(a \,\|\, c) \overset{\text{def.}}{=} a \log \frac{a}{c} + (1 - a) \log \frac{1-a}{1-c}$ for $a \in [0, 1]$ and $c \in (0, 1)$,

$$L_F(h' \,\|\, h) = \sum_{i=1}^m \mathrm{KL}(v_i \,\|\, \varsigma(z_i)),$$

a sum of Kullback-Leibler divergences between Bernoulli laws. Mirror alignment reads $v_i = \varsigma(-y_i(\Phi\mu)_i)$, the probability that the model gives to the label $-y_i$.

Computations. We have $\psi' = \varsigma$ and $\psi'' = \varsigma(1 - \varsigma) \in (0, \frac{1}{4}]$, an even function. So F is $\mathscr{C}^\infty$, nonnegative and strictly convex, with $\nabla F(h)_i = -y_i \varsigma(-y_i h_i)$ and $\nabla^2 F(h) = \mathrm{diag}(\varsigma'(h_i))$.

Conjugate of ψ. For $v < 0$, $vs - \psi(s) \to +\infty$ when $s \to -\infty$. For $v > 1$, $vs - \psi(s) \geqslant (v - 1)s - \log 2$ for $s \geqslant 0$, which tends to $+\infty$. For $v = 0$ the supremum of $-\psi$ is 0, and for $v = 1$ the supremum of $s - \psi(s) = -\log(1 + e^{-s})$ is 0. For $v \in (0, 1)$ the supremum is attained where $\varsigma(s) = v$, that is $s = \log \frac{v}{1-v}$, and its value is $v \log \frac{v}{1-v} + \log(1 - v) = \psi^\star(v)$. Since $y_i^2 = 1$, $ut = (-y_i u)(-y_i t)$, so $f_i^\star(u) = \psi^\star(-y_i u)$, and $h_i h'_i = v_i z_i$.

Loss. So $L_{f_i}(h_i' \,\|\, h_i) = \psi(z_i) + \psi^\star(v_i) - v_i z_i$. Since $\log \varsigma(z) = z - \psi(z)$ and $\log(1-\varsigma(z)) = -\psi(z)$,

$$\mathrm{KL}(v \,\|\, \varsigma(z)) = \psi^\star(v) - v(z - \psi(z)) + (1-v)\psi(z) = \psi(z) + \psi^\star(v) - vz. \qquad \square$$

The gradient of F is $\frac{1}{4}$-Lipschitz. Since $\varsigma'(s)$ decreases with $|s|$, F is $\varsigma'(M)$-strongly convex on the cube $\{h \,;\, \|h\|_\infty \leqslant M\}$, and not on $\mathbb{R}^m$. The range of ∇F is the open cube $\{h \,;\, -y_i h_i \in (0,1)\}$, while $\operatorname{dom} F^\star$ is the closed cube.

Remark 5.7 (The naive certificate vanishes). Mirror alignment at the target gives the candidate $-\Phi^\star \nabla F(\Phi\mu^0)/\lambda$. In general it is not a certificate of μ^0. Two cases make it vanish. (a) Least squares without noise: $\nabla F(\Phi\mu^0) = \Phi\mu^0 - y_0 = 0$. (b) Logistic regression in mean. Let the labels be independent with $\mathbb{P}(y_i = 1) = q_i \overset{\text{def.}}{=} \varsigma((\Phi\mu^0)_i)$. Since $\log(1 + e^{-yt}) = \psi(t) - \mathbf{1}_{y=1} t$, the mean fidelity is $\bar{F}(h) \overset{\text{def.}}{=} \mathbb{E}F(h) = \sum_i (\psi(h_i) - q_i h_i)$, and $\nabla \bar{F}(h)_i = \varsigma(h_i) - q_i$. So $\nabla \bar{F}(\Phi\mu^0) = 0$. In both cases the candidate is $\eta = 0$, and $\langle 0, \mu^0 \rangle_{\mathcal{M}} = 0 < \|\mu^0\|_{\mathrm{TV}}$ when $\mu^0 \neq 0$, so $0 \notin \partial \|\mu^0\|_{\mathrm{TV}}$. This is the noiseless mirror alignment $\nabla F_0(\Phi\mu^0) = 0$ of Section 5.6. The certificate must be chosen for R, as $\Phi^\star p$ with $h^\star = -\lambda p$. For least squares, the rescaled dual solutions $\hat{p}_\lambda(y_0)$ converge to p_0 (Proposition 4.2). For a general fidelity, [22] describes the certificate $\eta_0^\star$ of Theorem 5.4 as the limit of $-\Phi^\star h_\lambda/\lambda$ along the noiseless path. For the empirical fidelity, $F = \bar{F} + \langle w, \cdot \rangle$ with $w_i \overset{\text{def.}}{=} q_i - \mathbf{1}_{y_i=1}$, so $g = \nabla F(\Phi\mu^0) = w$ is centred, with $|w_i| < 1$: this is the form of Remark 5.6 (c).

Example 5.4 (Kullback-Leibler with a linear link). Let $y \in (0,\infty)^m$ and $f_i(t) \overset{\text{def.}}{=} t - y_i \log t + y_i \log y_i - y_i$ for $t > 0$, $f_i(t) \overset{\text{def.}}{=} +\infty$ for $t \leqslant 0$. Then $F(h) = \mathrm{KL}(y \,\|\, h)$ is the generalised Kullback-Leibler divergence, $\nabla F(h) = (1 - y_i/h_i)_i$ and $\nabla^2 F(h) = \operatorname{diag}(y_i/h_i^2)$ for $h > 0$, and $f_i^\star(v) = -y_i \log(1-v)$ for $v < 1$, $+\infty$ for $v \geqslant 1$. For $h < 1$ and $h' > 0$,

$$L_F(h' \,\|\, h) = \sum_{i=1}^m y_i \, \psi_{\mathrm{IS}}\Big(\frac{h_i'(1-h_i)}{y_i}\Big), \qquad \psi_{\mathrm{IS}}(t) \overset{\text{def.}}{=} t - 1 - \log t \geqslant 0.$$

Mirror alignment reads $1 - h_i = y_i/(\Phi\mu)_i$: the dual vector is a ratio residual.

Computations. For $v \geqslant 1$, $vt - f_i(t) \geqslant y_i \log t - y_i \log y_i + y_i$ for $t > 0$, so the supremum is $+\infty$. For $v < 1$, the concave function $t \mapsto (v-1)t + y_i \log t$ is maximal on $(0,\infty)$ at $t = y_i/(1-v)$, and the supremum is $-y_i + y_i \log \frac{y_i}{1-v} - y_i \log y_i + y_i = -y_i \log(1-v)$. For the loss, with $t \overset{\text{def.}}{=} h_i'(1-h_i)/y_i$, $L_{f_i}(h_i' \,\|\, h_i) = h_i' - y_i \log h_i' + y_i \log y_i - y_i - y_i \log(1-h_i) - h_i h_i' = y_i t - y_i - y_i \log t$. $\square$

This fidelity is finite only on the open orthant, so it is outside Definition 5.1. The proofs of Theorems 5.1 and 5.3 are algebra, and they hold unchanged when $F(\Phi\mu)$ and $F(\Phi\mu^\star)$ are finite. Existence and strong duality need a qualification condition, a measure $\nu_0 \in \operatorname{dom} R$ with $\Phi\nu_0 > 0$ coordinatewise [15]. The curvature y_i/h_i^2 tends to 0 when $h_i \to \infty$. So F is strongly convex on $(0, M]^m$, with $\sigma = \min_i y_i/M^2$, and not on the orthant.

5.6 Support stability for the generalised BLASSO

Theorem 4.2 says that, at low noise, the BLASSO recovers exactly s spikes. The same holds for the generalised BLASSO. The noise now enters through the fidelity: we consider a family $(F_w)_{w \in \mathbb{H}}$, where F_0 is the fidelity without noise. For least squares, $F_w(h) = \frac{1}{2}\|h - \Phi\mu^0 - w\|_{\mathbb{H}}^2$. For the logistic model of Remark 5.7, $F_w = \bar{F} + \langle w, \cdot \rangle_{\mathbb{H}}$. The low-noise regime D_{α,λ_0} is that of (2.1). In this section $R = \|\cdot\|_{\mathrm{TV}}$, $\mathcal{X} = \mathbb{T}$ or $\mathcal{X}$ is a closed subset of $\mathbb{R}$, and Γ_x is the operator of Section 3.2.

Theorem 5.4 (Support stability and rate). *Let $\mu^0 = \sum_{j=1}^s a_j \delta_{x_j}$ with $a_j \neq 0$ and pairwise distinct x_j in the interior of $\mathcal{X}$, and let $(F_w)_{w \in \mathbb{H}}$ be a family of fidelities. Assume:*

1. *F_0 is $\mathscr{C}^3$, coercive on $\mathbb{H}$, σ_F-strongly convex on a neighbourhood of $\Phi\mu^0$ for some $\sigma_F > 0$, and $\nabla F_0(\Phi\mu^0) = 0$. The map $(w,h) \mapsto \nabla F_w(h)$ is $\mathscr{C}^2$ on a neighbourhood of $(0, \Phi\mu^0)$, and so is $(w,k) \mapsto \nabla F_w^\star(k)$ on a neighbourhood of $(0,0)$. There are neighbourhoods $\mathcal{U}$ of $\Phi\mu^0$ and $\mathcal{V}$ of 0 in $\mathbb{H}$ and $L > 0$ such that, for every w in a neighbourhood of 0,*

$$\sup_{h\in\mathcal{U}} \|\nabla F_w(h) - \nabla F_0(h)\|_{\mathbb{H}} \leqslant L\,\|w\|_{\mathbb{H}} \quad \textit{and} \quad \sup_{k\in\mathcal{V}} \|\nabla F_w^\star(k) - \nabla F_0^\star(k)\|_{\mathbb{H}} \leqslant L\,\|w\|_{\mathbb{H}}\,;$$

2. *$x \mapsto \varphi_x$ is $\mathscr{C}^3$ and Γ_{x_0} is injective;*

3. *with $H_0 \overset{\text{def.}}{=} \nabla^2 F_0(\Phi\mu^0)$, the certificate $\eta_0^\star = \Phi^\star p_0^\star$, where $p_0^\star$ is the unique minimiser of $\langle p,\, H_0^{-1} p\rangle_{\mathbb{H}}$ among the $p \in \mathbb{H}$ with $\Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$, exists, $|\eta_0^\star(x)| < 1$ for $x \notin \{x_1,\dots,x_s\}$, and $\operatorname{sign}(a_j)\,(\eta_0^\star)''(x_j) < 0$ for all j.*

Then there are $\alpha, \lambda_0, c > 0$ such that, for $0 < \lambda \leqslant \lambda_0$ and $\|w\|_{\mathbb{H}} \leqslant \alpha\lambda$, the generalised BLASSO with fidelity F_w has a unique solution. It has s spikes $\hat{a}_j \delta_{\hat{x}_j}$ with $\operatorname{sign}(\hat{a}_j) = \operatorname{sign}(a_j)$ and

$$\sum_{j=1}^s |\hat{a}_j - a_j| + \sum_{j=1}^s |\hat{x}_j - x_j| \leqslant c\,(\lambda + \|w\|_{\mathbb{H}}).$$

The conditions on the family (F_w) in assumption 1 are a regularity condition taken from [22]. The condition $\nabla F_0(\Phi\mu^0) = 0$ is the noiseless mirror alignment: $\Phi\mu^0$ minimises F_0. The same preprint also states that $(\lambda, w) \mapsto (\hat{a}, \hat{x})$ extends to a $\mathscr{C}^2$ map on the closure of D_{α,λ_0}, with value (a, x_0) at $(0,0)$. The point $p_0^\star$ is the certificate of minimal norm for the norm $\|p\|_{H_0^{-1}} \overset{\text{def.}}{=} \langle p,\, H_0^{-1} p\rangle_{\mathbb{H}}^{1/2}$ (Remark 5.9).

Remark 5.8 (The mechanism and the quadratic case). (a) Let $\sum_j b_j \delta_{x'_j}$ be a solution with $\operatorname{sign}(b) = \sigma \overset{\text{def.}}{=} \operatorname{sign}(a)$ and pairwise distinct points x'_j in the interior of $\mathcal{X}$. By Theorem 5.2, $\eta \overset{\text{def.}}{=} -\Phi^\star \nabla F_w(\Phi_{x'} b)/\lambda$ lies in $\partial\|\sum_j b_j \delta_{x'_j}\|_{\mathrm{TV}}$. So $\eta(x'_j) = \sigma_j$, and the argument of Remark 3.3 gives $\eta'(x'_j) = 0$. By (3.4) these conditions read $E(b, x', \lambda, w) = 0$ with

$$E(b, x', \lambda, w) \overset{\text{def.}}{=} \Gamma_{x'}^\star \nabla F_w(\Phi_{x'} b) + \lambda \begin{pmatrix} \sigma \\ 0_s \end{pmatrix}.$$

For $F_w(h) = \frac{1}{2}\|h - y_0 - w\|_{\mathbb{H}}^2$, $\nabla F_w(\Phi_{x'} b) = \Phi_{x'} b - y_0 - w$, and E is the map of Step 3 of Theorem 4.2. Since $\nabla F_0(\Phi\mu^0) = 0$, the terms where $\Gamma_{x'}^\star$ is differentiated vanish at $(a, x_0, 0, 0)$, as in Lecture 4, and the chain rule gives

$$\frac{\partial E}{\partial(b, x')}(a, x_0, 0, 0) = \Gamma_{x_0}^\star H_0 \left(\Phi_{x_0}, \Phi'_{x_0} \operatorname{diag}(a)\right) = \Gamma_{x_0}^\star H_0 \Gamma_{x_0} \begin{pmatrix} \mathrm{Id}_s & 0 \\ 0 & \operatorname{diag}(a) \end{pmatrix}.$$

Since $H_0 \succeq \sigma_F \mathrm{Id}$, $\langle z,\, \Gamma_{x_0}^\star H_0 \Gamma_{x_0} z\rangle \geqslant \sigma_F \|\Gamma_{x_0} z\|_{\mathbb{H}}^2$ for $z \in \mathbb{R}^{2s}$, which is positive for $z \neq 0$ by injectivity. With $a_j \neq 0$, the Jacobian is invertible, and the implicit function theorem gives a branch of candidate solutions. The kernel K is replaced by the reweighted kernel $\langle \varphi_x,\, H_0 \varphi_{x'}\rangle_{\mathbb{H}}$, and assumption 3 is (NDSC) for this kernel. As in Step 2 of the proof of Theorem 4.2, the two conditions on $\eta_0^\star$ in assumption 3 exclude spurious spikes. (b) For $F_w(h) = \frac{1}{2}\|h - \Phi\mu^0 - w\|_{\mathbb{H}}^2$, F_0 is $\mathscr{C}^\infty$, coercive and 1-strongly convex, $\nabla F_0(\Phi\mu^0) = 0$, $\nabla F_w(h) - \nabla F_0(h) = -w$ and, by Example 5.1, $\nabla F_w^\star(k) = k + \Phi\mu^0 + w$. So assumption 1 holds with $L = 1$ and $\mathcal{U} = \mathcal{V} = \mathbb{H}$. Here $H_0 = \mathrm{Id}$, so $p_0^\star = p_0$ and $\eta_0^\star = \eta_0$ (Lemma 3.1), and assumption 3 is (NDSC) of Definition 3.3. The theorem gives the first conclusions of Theorem 4.2, without the convolution structure, on $\mathcal{X} = \mathbb{T}$ or $\mathcal{X} \subset \mathbb{R}$, for $x \mapsto \varphi_x$ of class $\mathscr{C}^3$. (c) For the Gaussian mixture of Example 2.2, $\mathcal{X} = \mathbb{R}$ has no boundary. If $x \mapsto \varphi_x$ is $\mathscr{C}^3$ (one adapts the computation of Example 2.2; we admit it), Γ_{x_0} is injective and (NDSC) holds, then (b) applies with $w = w_n$. This is the use made in

Remark 4.9. (d) The population logistic fidelity on L^2 is not covered: it is neither coercive nor strongly convex on any ball of L^2, and the domain of its conjugate has empty interior. A version of the theorem on bounded sets of predictions is given in [22].

As in Lecture 3, assumption 3 can be tested on a pre-certificate with a closed form.

Remark 5.9 (The weighted pre-certificate). Let H_0 be a bounded self-adjoint operator on $\mathbb{H}$ with $\langle H_0 v,\, v\rangle_{\mathbb{H}} \geqslant \sigma_F \|v\|_{\mathbb{H}}^2$ for all v, let $\Gamma \overset{\text{def.}}{=} \Gamma_{x_0}$ be injective, and $u \overset{\text{def.}}{=} (\operatorname{sign}(a), 0_s)$. (a) H_0 is invertible, $\left\|H_0^{-1}\right\| \leqslant 1/\sigma_F$, and $\langle p,\, H_0^{-1} q\rangle_{\mathbb{H}}$ is a scalar product whose norm is equivalent to $\|\cdot\|_{\mathbb{H}}$. If (SC) holds, $p_0^\star$ exists and is unique. (b) The vector $p_V^H \overset{\text{def.}}{=} H_0\Gamma(\Gamma^\star H_0 \Gamma)^{-1} u$ is the unique minimiser of $\langle p,\, H_0^{-1} p\rangle_{\mathbb{H}}$ among the p with $\Gamma^\star p = u$. (c) If $\left\|\Phi^\star p_V^H\right\|_\infty \leqslant 1$, then $p_V^H = p_0^\star$. For $H_0 = \mathrm{Id}$ these are Lemma 3.1, Proposition 3.2 and one direction of Proposition 3.3.

Proof. All norms and scalar products are those of $\mathbb{H}$. (a) For $v \in \mathbb{H}$, $\sigma_F \|v\|^2 \leqslant \langle H_0 v,\, v\rangle \leqslant \|H_0 v\|\,\|v\|$, so $\|H_0 v\| \geqslant \sigma_F \|v\|$. Hence H_0 is injective and its range is closed. The orthogonal of the range is the kernel of $H_0^\star = H_0$, which is $\{0\}$. So H_0 is onto, and $\left\|H_0^{-1}\right\| \leqslant 1/\sigma_F$. With $v = H_0^{-1} p$, $\langle p,\, H_0^{-1} p\rangle = \langle H_0 v,\, v\rangle \geqslant \sigma_F \|v\|^2 \geqslant \sigma_F \|p\|^2 / \|H_0\|^2$, and $\langle p,\, H_0^{-1} p\rangle \leqslant \|p\|^2 / \sigma_F$. So the norms are equivalent. The set $P \overset{\text{def.}}{=} \{p\,;\, \Phi^\star p \in \partial \|\mu^0\|_{\mathrm{TV}}\}$ is closed and convex, and nonempty under (SC). The minimal norm step of the proof of Lemma 3.1, for this scalar product, gives a unique element of minimal norm. (b) For $c \in \mathbb{R}^{2s}$, $\langle c,\, \Gamma^\star H_0 \Gamma c\rangle \geqslant \sigma_F \|\Gamma c\|^2$, so $\Gamma^\star H_0 \Gamma$ is invertible. Let $c_0 \overset{\text{def.}}{=} (\Gamma^\star H_0 \Gamma)^{-1} u$ and $q \overset{\text{def.}}{=} p_V^H = H_0 \Gamma c_0$. Then $\Gamma^\star q = u$. Let p with $\Gamma^\star p = u$ and $d \overset{\text{def.}}{=} p - q$. Then $\Gamma^\star d = 0$ and $\langle d,\, H_0^{-1} q\rangle = \langle d,\, \Gamma c_0\rangle = \langle \Gamma^\star d,\, c_0\rangle = 0$. Hence $\langle p,\, H_0^{-1} p\rangle = \langle q,\, H_0^{-1} q\rangle + \langle d,\, H_0^{-1} d\rangle$, and the last term is positive if $d \neq 0$ by (a). (c) Since $\Gamma^\star p_V^H = u$, (3.4) gives $(\Phi^\star p_V^H)(x_i) = \operatorname{sign}(a_i)$. With $\left\|\Phi^\star p_V^H\right\|_\infty \leqslant 1$, Proposition 2.3 gives $p_V^H \in P$. Every $p \in P$ satisfies $\Gamma^\star p = u$ by Remark 3.3, since the x_i are interior points. So p_V^H minimises the weighted norm on a set that contains P, and it lies in P. Hence it is the minimiser on P. □

Exercises

Exercise 5.1 (One atom)**.** Let $\mathcal{X}$ be compact, $\mathbb{H} = \mathbb{R}$, and $\varphi \in C(\mathcal{X})$ with $\varphi(x_0) = 1$ for some $x_0 \in \mathcal{X}$ and $|\varphi(x)| < 1$ for $x \neq x_0$. Let $\Phi\mu \overset{\text{def.}}{=} \int \varphi \,\mathrm{d}\mu$, $R = \|\cdot\|_{\mathrm{TV}}$, and let $F : \mathbb{R} \to \mathbb{R}$ be as in Definition 5.1.

1. Show that $\Phi^\star h = h\varphi$ and that the assumption of Proposition 2.1 holds. Show that 0 is a solution of (5.1) if and only if $|F'(0)| \leqslant \lambda$.
2. Assume $F'(0) < -\lambda$. Show that F' is strictly increasing and that $\lim_{t \to +\infty} F'(t) \geqslant 0$. Deduce that there is a unique $\hat{t}$ with $F'(\hat{t}) = -\lambda$, and that $\hat{t} > 0$.
3. Compute $\hat{h}$ and $\hat{\eta}$ for $\hat{\mu} \overset{\text{def.}}{=} \hat{t}\delta_{x_0}$, and show with Theorem 5.2 that $\hat{\mu}$ is a solution.
4. Show that $\hat{\mu}$ is the unique solution. Hint: Remark 5.4 and Remark 2.6.
5. Logistic loss with one label $y = 1$, $F(h) = \log(1 + e^{-h})$: show that 0 is a solution if and only if $\lambda \geqslant 1/2$, and compute $\hat{t}$ for $\lambda < 1/2$. Poisson loss with a count $y \in \mathbb{N}$, $F(h) = e^h - yh$: when is 0 a solution? Compute $\hat{t}$ when $y > 1 + \lambda$.

Exercise 5.2 (Strong convexity and the conjugate)**.** Let F be $\mathscr{C}^1$ and satisfy (5.7) with $\sigma > 0$.

1. Let $v, h \in \mathbb{H}$ and $g \overset{\text{def.}}{=} \nabla F(v)$. Show that $\langle h,\, u\rangle - F(u) \leqslant \langle h,\, v\rangle - F(v) + \langle h - g,\, u - v\rangle - \frac{\sigma}{2}\|u - v\|^2$ for all u. Deduce that $F^\star(h) < \infty$ and $L_F(v \,\|\, h) \leqslant \|h - g\|^2 / (2\sigma)$.
2. Let $\eta = \Phi^\star p \in \partial \|\mu^0\|_{\mathrm{TV}}$ and $h^\star = -\lambda p$. With Theorem 5.3, recover the second bound of Proposition 5.2.

3. Let $u, v \in \mathbb{H}$. With Remark 5.2, show that $L_F(u \,\|\, \nabla F(v)) \geqslant \frac{\sigma}{2} \|u - v\|^2$.

4. Assume that $-\lambda p = \nabla F(v^\star)$ for some $v^\star \in \mathbb{H}$. Deduce from Theorem 5.3 a bound on $\|\Phi\hat{\mu} - v^\star\|$ for every solution $\hat{\mu}$. For least squares, what is $v^\star$, and what bound do you get?

5. For $F(h) = \frac{1}{2}\|h - y\|^2$, show that the inequalities of questions 1 and 3 are equalities.

Exercise 5.3 (Poisson counts)**.** Let F be as in Example 5.2, $R = \|\cdot\|_{\mathrm{TV}}$, $\mu^0 \in \mathcal{M}(\mathcal{X})$, $\eta = \Phi^\star p \in \partial\|\mu^0\|_{\mathrm{TV}}$, $g \overset{\text{def.}}{=} \nabla F(\Phi\mu^0)$, $B \overset{\text{def.}}{=} \sup_x \|\varphi_x\|_2$, and let $\hat{\mu}$ be a solution of (5.1).

1. Show that F is not σ-strongly convex on $\mathbb{R}^m$ for any $\sigma > 0$.

2. Let $K \subset \mathbb{R}^m$ be convex with $\langle \nabla^2 F(u)v,\, v\rangle \geqslant \sigma\|v\|^2$ for all $u \in K$ and $v \in \mathbb{R}^m$. Show that (5.7) holds for $u, v \in K$. Deduce that the two bounds of Proposition 5.2 hold if $\Phi\mu^0$ and $\Phi\hat{\mu}$ lie in K.

3. Let $L_0 \overset{\text{def.}}{=} F(\Phi\mu^0) - \inf F$. Show that $L_0 = \sum_i d(y_i, e^{(\Phi\mu^0)_i})$ with d as in Example 5.2, and that $L_0 = L_F(\Phi\mu^0 \,\|\, 0)$. Show that $\|\hat{\mu}\|_{\mathrm{TV}} \leqslant \|\mu^0\|_{\mathrm{TV}} + L_0/\lambda$.

4. Let $M \overset{\text{def.}}{=} B(\|\mu^0\|_{\mathrm{TV}} + L_0/\lambda)$. Show that $\|\Phi\mu\|_\infty \leqslant B\|\mu\|_{\mathrm{TV}}$ for every μ. Deduce that $\|\Phi(\hat{\mu} - \mu^0)\|_2 \leqslant 2e^M \|g + \lambda p\|_2$ and $\lambda(\|\hat{\mu}\|_{\mathrm{TV}} - \int \eta \,\mathrm{d}\hat{\mu}) \leqslant e^M \|g + \lambda p\|_2^2/2$.

5. Assume that the y_i are independent with Poisson laws of means $e^{(\Phi\mu^0)_i}$. Compute $\mathbb{E}g$ and $\mathbb{E}\|g\|_2^2$. How does M behave when $\lambda \to 0$? Why does this call for a local result such as Theorem 5.4?

Exercise 5.4 (Huber fidelity)**.** Let $\delta > 0$, ρ_δ be the Huber function of Section 5.1, $\mathbb{H} = \mathbb{R}^m$ and $F(h) = \sum_i \rho_\delta(h_i - y_i)$.

1. Show that ρ_δ is $\mathscr{C}^1$ and convex, with $\rho'_\delta(t) = \max(-\delta, \min(\delta, t))$, and that it is not strictly convex. Which assumption of Definition 5.1 fails?

2. Show that $\rho_\delta^\star(u) = u^2/2$ for $|u| \leqslant \delta$ and $+\infty$ otherwise. Deduce $F^\star(h) = \frac{1}{2}\|h\|^2 + \langle h,\, y\rangle$ on $\{h \,;\, \|h\|_\infty \leqslant \delta\}$, and $+\infty$ outside.

3. Show that $L_F(h' \,\|\, h) = \frac{1}{2}\|h' - y - h\|^2$ when $\|h' - y\|_\infty \leqslant \delta$ and $\|h\|_\infty \leqslant \delta$.

4. Show that Lemma 5.1 and Theorems 5.1, 5.2 and 5.3 hold for this F. Which statement of Remark 5.4 uses strict convexity?

5. Let $\hat{\mu}$ be a solution and $\hat{h} = \nabla F(\Phi\hat{\mu})$. Show that $\left|\hat{h}_i\right| = \delta$ when $|(\Phi\hat{\mu})_i - y_i| \geqslant \delta$. Interpret: how much can one outlier move the certificate $-\Phi^\star\hat{h}/\lambda$?

Exercise 5.5 (A computable stopping test)**.** Let F be as in Definition 5.1 and $R = \|\cdot\|_{\mathrm{TV}}$. For $\mu \in \mathcal{M}(\mathcal{X})$ let $u_\mu \overset{\text{def.}}{=} \nabla F(\Phi\mu)$, $c_\mu \overset{\text{def.}}{=} \min\left(1, \lambda/\left\|\Phi^\star u_\mu\right\|_\infty\right)$, with $c_\mu \overset{\text{def.}}{=} 1$ if $\Phi^\star u_\mu = 0$, $h_\mu \overset{\text{def.}}{=} c_\mu u_\mu$ and $\eta_\mu \overset{\text{def.}}{=} -\Phi^\star h_\mu/\lambda$.

1. Show that $F^\star$ is convex, that $F^\star(0) = -\inf F$ is finite and that $F^\star(u_\mu) = \langle u_\mu,\, \Phi\mu\rangle_{\mathbb{H}} - F(\Phi\mu)$. Deduce that $F^\star(h_\mu) < \infty$.

2. Let $\Delta(\mu) \overset{\text{def.}}{=} L_F(\Phi\mu \,\|\, h_\mu) + \lambda(\|\mu\|_{\mathrm{TV}} - \int \eta_\mu \,\mathrm{d}\mu)$. Show that $\Delta(\mu)$ is finite and that $J(\mu) - \min J \leqslant \Delta(\mu)$.

3. Show that $\Delta(\hat{\mu}) = 0$ for every solution $\hat{\mu}$.

4. For $F = \frac{1}{2}\|\cdot - y\|_{\mathbb{H}}^2$, show that $h_\mu = -\lambda p_\mu/\max(1, \|\Phi^\star p_\mu\|_\infty)$ with $p_\mu = (y - \Phi\mu)/\lambda$. Compare with Remark 6.1.

5. Let $\hat{\mu}$ be a solution and (μ_k) a sequence with $\Phi\mu_k \to \Phi\hat{\mu}$ in $\mathbb{H}$ and $\|\mu_k\|_{\mathrm{TV}} \to \|\hat{\mu}\|_{\mathrm{TV}}$. Show that $c_{\mu_k} \to 1$ and that $\Delta(\mu_k) \to 0$. Hint: $F^\star(cu) \leqslant cF^\star(u) + (1 - c)F^\star(0)$ for $c \in [0, 1]$.

Chapter 6

Algorithms for the BLASSO

This chapter is a reading guide for the projects. For each method it gives the principle, the iteration in pseudo-code, the convergence statement with its hypotheses, and the source. The convergence results are admitted. The only proof is that of the stopping test, which takes a few lines from Lectures 2 and 5. Throughout, the setting is that of Definition 2.2, $\mathcal{X}$ is compact, and the assumption of Proposition 2.1 holds. So $\Phi^\star p$ is a continuous function on $\mathcal{X}$ for every $p \in \mathbb{H}$, and $|\Phi^\star p|$ attains its maximum.

Each method computes the certificate of its iterate: it says where to put mass, and the duality gap of Lecture 5 says when to stop. For a measure μ, let $p_\mu \overset{\text{def.}}{=} (y - \Phi\mu)/\lambda$ and $\eta_\mu \overset{\text{def.}}{=} \Phi^\star p_\mu$. By item 1 of Proposition 2.4 and Proposition 2.3, μ solves $(\mathcal{P}_\lambda(y))$ if and only if $\|\eta_\mu\|_\infty \leqslant 1$ and $\langle \eta_\mu, \mu \rangle_{\mathcal{M}} = \|\mu\|_{\mathrm{TV}}$. For a discrete measure, the second condition says that η_μ equals the sign of μ on its support. On non-negative measures the first variation of the objective is $J'_\nu = \lambda(1 - \eta_\nu)$ (Exercise 6.3): mass is missing where $J'_\nu < 0$, that is where $\eta_\nu > 1$.

The methods differ in the way they find this missing mass. The global search is the maximisation of $|\eta_\mu|$ over all of $\mathcal{X}$; on a net, its cost grows exponentially with the dimension d of $\mathcal{X}$.

Method	Measures	Global search	Guarantee	Section
Frank-Wolfe	signed	every step	$\mathcal{O}(1/k)$ on the objective	6.2
Sliding Frank-Wolfe	signed	every step	exact after finitely many steps	6.2
CPGD	non-negative	none	linear rate, spread start	6.3
FastPart	non-negative	none	no statement here	6.4
FS&P	non-negative	none	rate in expectation, sparse start	6.5

6.1 The duality gap as a stopping test

None of the methods knows the optimal value. The duality gap bounds the excess objective by a number that depends on the iterate and on a dual vector only. The dual vector must be feasible, and feasibility is a bound on $\Phi^\star p$ over all of $\mathcal{X}$.

Let $\mathcal{L}(\mu) \overset{\text{def.}}{=} \frac{1}{2}\|\Phi\mu - y\|_{\mathbb{H}}^2 + \lambda\|\mu\|_{\mathrm{TV}}$ and $\mathcal{L}^\star \overset{\text{def.}}{=} \min \mathcal{L}$.

Definition 6.1 (BLASSO duality gap)**.** For $\mu \in \mathcal{M}(\mathcal{X})$ and $p \in \mathbb{H}$ with $\|\Phi^\star p\|_\infty \leqslant 1$,

$$\Delta(\mu, p) \overset{\text{def.}}{=} \mathcal{L}(\mu) - \left(\lambda\langle p, y\rangle_{\mathbb{H}} - \tfrac{\lambda^2}{2}\|p\|_{\mathbb{H}}^2\right) = \tfrac{1}{2}\|\Phi\mu - y + \lambda p\|_{\mathbb{H}}^2 + \lambda\left(\|\mu\|_{\mathrm{TV}} - \langle \Phi^\star p, \mu\rangle_{\mathcal{M}}\right). \quad (6.1)$$

The function $\mathcal{L}$ is the objective of Proposition 2.2, so the minimum exists. The term in brackets is the dual function $D(p)$ of Proposition 2.4. The second equality is the identity (2.10), which holds for every $p \in \mathbb{H}$. By Remark 5.3, the two terms on the right are the two Fenchel-Young losses $L_F(\Phi\mu \,\|\, -\lambda p)$ and $\lambda L_R(\mu \,\|\, \Phi^\star p)$.

Corollary 6.1 (Stopping criterion)**.** *Let an algorithm produce $\mu_k \in \mathcal{M}(\mathcal{X})$ and $p_k \in \mathbb{H}$ with $\|\Phi^\star p_k\|_\infty \leqslant 1$. Then $\mathcal{L}(\mu_k) - \mathcal{L}^\star \leqslant \Delta(\mu_k, p_k)$. The test $\Delta(\mu_k, p_k) \leqslant \varepsilon$ certifies ε-optimality.*

Proof. Norms and scalar products are those of $\mathbb{H}$. Let $p \in \mathbb{H}$ with $\|\Phi^\star p\|_\infty \leqslant 1$, and let $\nu \in \mathcal{M}(\mathcal{X})$. By (2.10), $\mathcal{L}(\nu) - D(p) = \frac{1}{2}\|\Phi\nu - y + \lambda p\|^2 + \lambda(\|\nu\|_{\mathrm{TV}} - \langle \Phi^\star p, \nu\rangle_{\mathcal{M}})$. By (2.5), $\langle \Phi^\star p, \nu\rangle_{\mathcal{M}} \leqslant \|\nu\|_{\mathrm{TV}}$, so both terms are nonnegative and $\mathcal{L}(\nu) \geqslant D(p)$. This holds for every ν, so $\mathcal{L}^\star \geqslant D(p)$. With $p = p_k$,

$$0 \leqslant \mathcal{L}(\mu_k) - \mathcal{L}^\star \leqslant \mathcal{L}(\mu_k) - D(p_k) = \Delta(\mu_k, p_k).$$

If $\Delta(\mu_k, p_k) \leqslant \varepsilon$, then $\mathcal{L}(\mu_k) \leqslant \mathcal{L}^\star + \varepsilon$. □

The proof uses weak duality only. Strong duality says more: the gap vanishes exactly at optimal pairs. If $\Delta(\mu, p) = 0$, then $\mathcal{L}(\mu) = D(p) \leqslant \mathcal{L}^\star \leqslant \mathcal{L}(\mu)$, so μ is a solution and p maximises D on C, hence $p = \hat{p}_\lambda$ by item 3 of Proposition 2.4. Conversely, the same item gives $\Delta(\hat{\mu}, \hat{p}_\lambda) = 0$ for every solution $\hat{\mu}$. The gap bounds the objective, not the error on the measure. Error bounds follow all the same. The identity (5.6) holds for every μ with $R(\mu) < \infty$. So, if $J(\mu) \leqslant J(\mu^\star) + \varepsilon$, it gives the bound (5.5) of Theorem 5.3 with ε added to the right-hand side. The test of Corollary 6.1 gives this with $\mu^\star = \mu^0$, since $\mathcal{L}^\star \leqslant \mathcal{L}(\mu^0)$.

Remark 6.1 (Choice of dual variable). Take $p_k \overset{\text{def.}}{=} p_{\mu_k} / \max(1, \|\Phi^\star p_{\mu_k}\|_\infty)$, the sup norm being bounded by its maximum on a fine net of $\mathcal{X}$ plus the covering radius times a Lipschitz constant of $\Phi^\star p_{\mu_k}$. The global search over $\mathcal{X}$ is thus needed at stopping time.

The bound on the sup norm reads as follows. Let $N_h \subset \mathcal{X}$ be a finite net with covering radius h: every $x \in \mathcal{X}$ has a point $z \in N_h$ with $\|x - z\| \leqslant h$. If $x \mapsto \varphi_x$ is L_φ-Lipschitz, then $|\Phi^\star p(x) - \Phi^\star p(z)| \leqslant L_\varphi \|x - z\| \, \|p\|_{\mathbb{H}}$ by the Cauchy-Schwarz inequality, so

$$\|\Phi^\star p\|_\infty \leqslant \max_{z \in N_h} |\Phi^\star p(z)| + L_\varphi h \, \|p\|_{\mathbb{H}} \overset{\text{def.}}{=} s(p).$$

Dividing p_{μ_k} by $\max(1, s(p_{\mu_k}))$ keeps p_k feasible. A maximum on the net alone can be smaller than the supremum, and then Corollary 6.1 does not apply. The net has of order h^{-d} points.

Two facts tell what the test can do. First, with the exact sup norm, the gap tends to zero along every sequence with $\Phi\mu_k \to \Phi\hat{\mu}$ and $\|\mu_k\|_{\mathrm{TV}} \to \|\hat{\mu}\|_{\mathrm{TV}}$ (Exercise 5.5, questions 4 and 5). This holds as soon as $\mathcal{L}(\mu_k) \to \mathcal{L}^\star$: by (2.9), $\Phi\mu_k \to \Phi\hat{\mu}$, and then $\lambda\|\mu_k\|_{\mathrm{TV}} = \mathcal{L}(\mu_k) - \frac{1}{2}\|\Phi\mu_k - y\|_{\mathbb{H}}^2 \to \lambda\|\hat{\mu}\|_{\mathrm{TV}}$. So the test fires after finitely many steps, for every $\varepsilon > 0$. Second, with the bound s, the gap does not vanish at a solution. Let $s \overset{\text{def.}}{=} s(\hat{p}_\lambda) > 1$. Since $p_{\hat{\mu}} = \hat{p}_\lambda$ and $\langle \Phi^\star \hat{p}_\lambda, \hat{\mu}\rangle_{\mathcal{M}} = \|\hat{\mu}\|_{\mathrm{TV}}$,

$$\Delta(\hat{\mu}, \hat{p}_\lambda / s) = \lambda(1 - \tfrac{1}{s})\|\hat{\mu}\|_{\mathrm{TV}} + \tfrac{\lambda^2}{2}(1 - \tfrac{1}{s})^2 \, \|\hat{p}_\lambda\|_{\mathbb{H}}^2,$$

and $1 - 1/s \leqslant s - 1 \leqslant L_\varphi h \, \|\hat{p}_\lambda\|_{\mathbb{H}}$. The net must be fine enough for this floor to be below ε.

Remark 6.2 (Gap on non-negative measures). Let $\mathcal{L}_+^\star \overset{\text{def.}}{=} \inf_{\mathcal{M}_+(\mathcal{X})} \mathcal{L}$, so that $\mathcal{L}_+^\star \geqslant \mathcal{L}^\star$. Let $\nu \in \mathcal{M}_+(\mathcal{X})$ and $q \in \mathbb{H}$ with only $\Phi^\star q \leqslant 1$ on $\mathcal{X}$. Then the right-hand side of (6.1) at (ν, q) is at least $\mathcal{L}(\nu) - \mathcal{L}_+^\star$. Indeed, for $\nu' \in \mathcal{M}_+(\mathcal{X})$, $\|\nu'\|_{\mathrm{TV}} = \nu'(\mathcal{X})$, so $\|\nu'\|_{\mathrm{TV}} - \langle \Phi^\star q, \nu'\rangle_{\mathcal{M}} = \int_{\mathcal{X}} (1 - \Phi^\star q) \, \mathrm{d}\nu' \geqslant 0$. The proof of Corollary 6.1, run over $\nu' \in \mathcal{M}_+(\mathcal{X})$, gives $\mathcal{L}_+^\star \geqslant D(q)$, and $\mathcal{L}(\nu) - \mathcal{L}_+^\star \leqslant \mathcal{L}(\nu) - D(q)$, which is the right-hand side of (6.1) by (2.10). No dual problem on $\mathcal{M}_+(\mathcal{X})$ is needed. This allows the one-sided normalisation $q_k \overset{\text{def.}}{=} p_{\nu_k} / \max(1, \max_{\mathcal{X}} \Phi^\star p_{\nu_k})$, with the same net bound without absolute value. The sup norm normalisation of Remark 6.1 is also valid on $\mathcal{M}_+(\mathcal{X})$, since $\mathcal{L}_+^\star \geqslant \mathcal{L}^\star$, but it compares ν_k with the signed problem: its gap can stay away from zero at the minimiser on $\mathcal{M}_+(\mathcal{X})$, while the one-sided gap vanishes there (Exercise 6.2).

The test, in pseudo-code, for a measure μ given by its spikes $(a_i, x_i)_{i \leqslant N}$:

Duality gap test.
Input: spikes $(a_i, x_i)_{i \leqslant N}$, data y, penalty λ, net N_h of radius h, constant L_φ, tolerance $\varepsilon > 0$.

1. Residual: $r \leftarrow y - \sum_i a_i \varphi_{x_i}$ and $p \leftarrow r/\lambda$.
2. Upper bound: $M \leftarrow \max_{z \in N_h} |\langle \varphi_z, p \rangle_{\mathbb{H}}| + L_\varphi h \, \|p\|_{\mathbb{H}}$. For a measure in $\mathcal{M}_+(\mathcal{X})$, one may drop the absolute value.
3. Feasible point: $p \leftarrow p / \max(1, M)$.
4. Gap: $\Delta \leftarrow \frac{1}{2} \|\lambda p - r\|_{\mathbb{H}}^2 + \lambda \sum_i (\, |a_i| - a_i \langle \varphi_{x_i}, p \rangle_{\mathbb{H}})$.
5. If $\Delta \leqslant \varepsilon$, return "$\varepsilon$-optimal".

Line 4 is (6.1), with $\Phi\mu - y = -r$ and (2.7). Lines 1 and 4 cost $\mathcal{O}(N)$ evaluations of the features; line 2 costs $\mathcal{O}(h^{-d})$.

6.2 Frank-Wolfe and sliding Frank-Wolfe

Frank-Wolfe adds one spike where the optimality test fails most. In optimisation terms it is the conditional gradient method. The gradient of the data fit at μ_k is $\Phi^\star(\Phi\mu_k - y) = -\lambda\eta_k$. On the ball $\mathcal{B}_R$ of Theorem 6.1, the linear form $\mu \mapsto -\lambda\langle \eta_k, \mu \rangle_{\mathcal{M}}$ has minimum $-\lambda R \, \|\eta_k\|_\infty$ by (2.5), attained at the signed Dirac $R\,\mathrm{sign}(\eta_k(x^\star))\delta_{x^\star}$, where $x^\star$ maximises $|\eta_k|$. So the linear oracle is the maximisation of $|\eta_k|$.

Definition 6.2 (Frank-Wolfe iteration)**.** Starting from $\mu_0 = 0$, iterate:

1. Form the dual variable $p_k \overset{\text{def.}}{=} \frac{1}{\lambda}(y - \Phi\mu_k)$ and the certificate $\eta_k \overset{\text{def.}}{=} \Phi^\star p_k$.
2. Locate the worst constraint violator, by a global search over $\mathcal{X}$, $x^\star \overset{\text{def.}}{=} \operatorname*{argmax}_{x \in \mathcal{X}} |\eta_k(x)|$.
3. If $|\eta_k(x^\star)| \leqslant 1$, stop. Otherwise add a Dirac $\mu_{k+1} \overset{\text{def.}}{=} \mu_k + \theta_k \,\mathrm{sign}(\eta_k(x^\star))\, \delta_{x^\star}$ for an optimally chosen $\theta_k > 0$.

Let $\varsigma \overset{\text{def.}}{=} \mathrm{sign}(\eta_k(x^\star))$. If $|\eta_k(x^\star)| > 1$, then $\varphi_{x^\star} \neq 0$. When $\varsigma\, \mu_k(\{x^\star\}) \geqslant 0$, the optimal step is $\theta_k = \lambda(|\eta_k(x^\star)| - 1)/\, \|\varphi_{x^\star}\|_{\mathbb{H}}^2$. In all cases each step decreases $\mathcal{L}$ by at least $\lambda^2(\|\eta_k\|_\infty - 1)^2/(2B^2)$, with $B \overset{\text{def.}}{=} \sup_x \|\varphi_x\|_{\mathbb{H}}$ (Exercise 6.1).

Step 3 tests only the first optimality condition, $\|\eta_k\|_\infty \leqslant 1$. When it fires, p_k is feasible, the first term of (6.1) vanishes, and $\Delta(\mu_k, p_k) = \lambda(\|\mu_k\|_{\mathrm{TV}} - \langle \eta_k, \mu_k \rangle_{\mathcal{M}})$. Since $p_{\hat{\mu}} = \hat{p}_\lambda$ for every solution $\hat{\mu}$, μ_k is a solution if and only if this number is zero. In the plain iteration the old amplitudes are never optimised again, so the test of step 3 alone is not an optimality test; the gap decides, at no extra cost. In the fully corrective form, step 3 is followed by a LASSO on the finite set $S_{k+1} \overset{\text{def.}}{=} \mathrm{supp}(\mu_k) \cup \{x^\star\}$: μ_{k+1} minimises $\mathcal{L}$ among the measures supported in S_{k+1}. Item 1 of Proposition 2.4 and Proposition 2.3, applied with the finite set S_{k+1} in place of $\mathcal{X}$, give $\eta_{k+1}(x) = \mathrm{sign}(\mu_{k+1}(\{x\}))$ at every atom x of μ_{k+1}. So, in this form, when the test of step 3 fires, the iterate is a solution.

Theorem 6.1 (Vanilla FW convergence)**.** *The Frank-Wolfe iteration, run in its fully corrective form on the ball* $\mathcal{B}_R \overset{\text{def.}}{=} \{\mu \,;\, \|\mu\|_{\mathrm{TV}} \leqslant R\}$*, produces measures* μ_k *such that* $\mathcal{L}(\mu_k) - \mathcal{L}^\star \leqslant 8LR^2/(k+2)$*, where* $L = \|\Phi\|_{\mathcal{M} \to \mathbb{H}}^2$ *and* $R = \mathcal{L}(0)/\lambda = \|y\|_{\mathbb{H}}^2/(2\lambda)$.

Three remarks help to read the constants. First, $L = B^2$: $\|\Phi\mu\|_{\mathbb{H}} \leqslant B\|\mu\|_{\mathrm{TV}}$ (Section 2.5), with equality at a Dirac mass where $\|\varphi_x\|_{\mathbb{H}} = B$. Second, the radius comes from the iterates. The objective decreases from $\mu_0 = 0$, so $\lambda\|\mu_k\|_{\mathrm{TV}} \leqslant \mathcal{L}(\mu_k) \leqslant \mathcal{L}(0) = \frac{1}{2}\|y\|_{\mathbb{H}}^2$. Every

solution $\hat{\mu}$ satisfies $\lambda\|\hat{\mu}\|_{\mathrm{TV}} \leqslant \mathcal{L}(0)$ too, so the minimum of $\mathcal{L}$ on $\mathcal{B}_R$ is $\mathcal{L}^\star$. Third, the iteration is the conditional gradient method for $(\mu, t) \mapsto \frac{1}{2}\|\Phi\mu - y\|_{\mathbb{H}}^2 + \lambda t$ on the convex set $\{(\mu, t)\ ;\ \|\mu\|_{\mathrm{TV}} \leqslant t \leqslant R\}$. Its extreme points are $(0, 0)$ and $(\pm R\delta_x, R)$, where the linearised objective takes the values 0 and $\lambda R(1 \mp \eta_k(x))$. So the oracle is step 2. The curvature constant is at most $\sup_{\mu,\mu' \in \mathcal{B}_R} \|\Phi(\mu - \mu')\|_{\mathbb{H}}^2 \leqslant B^2(2R)^2 = 4LR^2$, and Jaggi's rate $2C/(k+2)$ with $C = 4LR^2$ is the bound of the theorem. The fully corrective iterate is at least as good as the convex combination step of that analysis.

The rate is slow in practice: the method adds new spikes near the true ones instead of moving the old ones. The sliding step corrects this. It moves the spikes and refits their amplitudes, with the number of spikes and their signs fixed.

Definition 6.3 (Sliding Frank-Wolfe iteration)**.** After step 3 of Definition 6.2, run a non-convex local solver on

$$\min_{(a_i, x_i)_{i=1}^{N_k}} \frac{1}{2}\Big\| \sum_{i=1}^{N_k} a_i \varphi_{x_i} - y \Big\|_{\mathbb{H}}^2 + \lambda \sum_{i=1}^{N_k} |a_i|\,,$$

seeded at the post-insertion measure (the current support together with the spike just added in step 3, so N_k counts it), until convergence. Replace μ_{k+1} by the resulting discrete measure.

With the signs frozen, $|a_i| = \mathrm{sign}(a_i)a_i$ near the seed, and the objective is a smooth function of $(a, x) \in \mathbb{R}^{N_k} \times \mathcal{X}^{N_k}$ when $x \mapsto \varphi_x$ is smooth. Gradient or quasi-Newton steps apply. A spike whose amplitude reaches zero is removed.

Sliding Frank-Wolfe.
Input: data y, penalty λ, tolerance $\varepsilon > 0$.
Set $\mu_0 \leftarrow 0$. For $k = 0, 1, 2, \ldots$:

1. Certificate: $p_k \leftarrow (y - \Phi\mu_k)/\lambda$ and $\eta_k \leftarrow \Phi^\star p_k$.
2. Global search: $x^\star \leftarrow \operatorname*{argmax}_{x \in \mathcal{X}} |\eta_k(x)|$ (a net, then a local refinement).
3. Test: if $|\eta_k(x^\star)| \leqslant 1$, return μ_k.
4. Insertion: $\mu_{k+1/2} \leftarrow \mu_k + \theta_k \varsigma \delta_{x^\star}$, with θ_k as in Exercise 6.1.
5. Sliding: run the local solver of Definition 6.3 from the spikes of $\mu_{k+1/2}$; remove the spikes of zero amplitude; call the result μ_{k+1}.
6. Optional: run the duality gap test on μ_{k+1} and stop if $\Delta \leqslant \varepsilon$.

Without step 5 this is vanilla Frank-Wolfe. Replacing step 5 by a refit of the amplitudes with the positions fixed gives the fully corrective form of Theorem 6.1.

Theorem 6.2 (Finite-time convergence of SFW)**.** *Let $\mathcal{X}$ be compact and $\varphi \in \mathscr{C}^2(\mathcal{X}, \mathbb{H})$. Let $\hat{\mu} = \sum_{j=1}^{s} \hat{a}_j \delta_{\hat{x}_j}$ be the unique solution of the BLASSO at $\lambda > 0$, with dual certificate $\eta_\lambda \overset{\text{def.}}{=} \frac{1}{\lambda}\Phi^\star(y - \Phi\hat{\mu})$. Assume that η_λ is non-degenerate:*

$$|\eta_\lambda(x)| < 1 \quad \forall\, x \in \mathcal{X} \setminus \{\hat{x}_1, \ldots, \hat{x}_s\}, \qquad \nabla^2 \eta_\lambda(\hat{x}_j) \text{ invertible}, \quad j = 1, \ldots, s. \tag{6.2}$$

Then SFW recovers $\hat{\mu}$ after finitely many outer iterations: there is $k^\star \in \mathbb{N}$ with $\mu_{k^\star} = \hat{\mu}$, and the stopping test of Definition 6.2 then fires.

The mechanism of the proof has four parts. The iterates converge weak-$\star$ to $\hat{\mu}$. The certificates η_k then converge to η_λ in $\mathscr{C}^2(\mathcal{X})$. Non-degeneracy then leaves, for k large, exactly one spike near each $\hat{x}_j$ and none elsewhere. At that point η_k certifies μ_k, so μ_k is the solution.

Remark 6.3 (Hypotheses left implicit). (a) The proof uses that each $\hat{x}_j$ is an interior extremum of η_λ, so that the invertible Hessian is definite. Read the theorem with $\mathcal{X} = \mathbb{T}^d$, or with the $\hat{x}_j$ in the interior of $\mathcal{X} \subset \mathbb{R}^d$. (b) Uniqueness of $\hat{\mu}$ is a hypothesis; it does not follow from (6.2). (c) Each sliding problem is solved exactly. Its output is then a critical point, so $\eta_k(x_i) = \operatorname{sign}(a_i)$ and $\nabla\eta_k(x_i) = 0$ at every spike x_i of μ_k, which the proof uses. With an inexact sliding solve, the theorem does not apply: use Corollary 6.1. (d) The first step of the proof applies the rate of Theorem 6.1 to SFW. The rate needs each SFW iterate to be at least as good as the convex combination step of that analysis. The fully corrective iterate is. A local solver seeded at the post-insertion measure only guarantees $\mathcal{L}(\mu_{k+1}) \leqslant \mathcal{L}(\mu_{k+1/2})$, and $\mu_{k+1/2}$ keeps the old amplitudes. The LASSO refit of the amplitudes before sliding, as in the algorithm of [23], gives the required inequality. (e) No bound on $k^\star$ is known. The frequent claim of s outer iterations is an observation.

To stop SFW: when step 3 fires, p_k is feasible and $\Delta(\mu_k, p_k) = \lambda(\|\mu_k\|_{\mathrm{TV}} - \langle \eta_k,\, \mu_k\rangle_{\mathcal{M}})$ costs nothing more, since step 2 has computed $\|\eta_k\|_\infty$.

6.3 Conic particle gradient descent

Frank-Wolfe needs a global search at every step. Particle methods remove it. They start from many particles, and the first variation moves them and changes their masses. In the particles the problem is non-convex. On measures, the dynamics is a gradient flow of a convex objective for the Wasserstein-Fisher-Rao metric: the Wasserstein part transports mass, the Fisher-Rao part creates and destroys it [24].

Let Θ be a d-dimensional Riemannian manifold, compact and without boundary, such as $\mathbb{T}^d$. We minimise over $\nu \in \mathcal{M}_+(\Theta)$

$$J(\nu) \overset{\text{def.}}{=} F\Big(\int_\Theta \varphi(\theta)\,\mathrm{d}\nu(\theta)\Big) + \lambda\,\nu(\Theta), \qquad J^\star \overset{\text{def.}}{=} \inf_{\mathcal{M}_+(\Theta)} J, \tag{6.3}$$

with $F \geqslant 0$ convex, φ and F twice Fréchet differentiable with locally Lipschitz second derivatives, and ∇F bounded on sublevel sets; the BLASSO is $F(f) = \frac{1}{2}\|f - y\|_{\mathbb{H}}^2$. The first variation is $J'_\nu(\theta) = \langle \varphi(\theta),\, \nabla F(\int \varphi\,\mathrm{d}\nu)\rangle_{\mathbb{H}} + \lambda$. The continuous-time limit of CPGD is the Wasserstein-Fisher-Rao flow

$$\partial_t \nu_t = -4\alpha\,\nu_t\,J'_{\nu_t} + \beta\,\operatorname{div}\left(\nu_t \nabla J'_{\nu_t}\right), \qquad \alpha, \beta > 0. \tag{6.4}$$

Here R is the regulariser of Lecture 5, and the data fit is F. For the BLASSO, $\nabla F(f) = f - y$ and $J'_\nu = \lambda(1 - \eta_\nu)$. Since $F \geqslant 0$, $\nu(\Theta) \leqslant J(\nu)/\lambda$. The first term of (6.4) removes mass where $J'_\nu > 0$ and adds mass where $J'_\nu < 0$, at places where ν_t already has mass. The second term transports mass along $-\nabla J'_\nu$. A solution of (6.4) is understood in the weak sense; the existence of a solution is not asserted here, and the statements below hold for every solution. On a domain with boundary, such as a box of $\mathbb{R}^d$, the results below need in addition that the support of the minimiser lies in the interior. Signed measures enter through the doubled dictionary $\{\varphi(\theta)\} \cup \{-\varphi(\theta)\}$ [24, App. A].

Definition 6.4 (Conic Particle Gradient Descent)**.** Given step sizes $\alpha, \beta > 0$, initial particles $(r_i^{(0)}, \theta_i^{(0)})_{i=1}^m$ and a cone-compatible retraction Ret, iterate

$$(r_i^{(k+1)}, \theta_i^{(k+1)}) \leftarrow \mathrm{Ret}_{(r_i^{(k)}, \theta_i^{(k)})}\big(-2\alpha r_i^{(k)} J'_{\nu_k}(\theta_i^{(k)}), -\beta \nabla J'_{\nu_k}(\theta_i^{(k)})\big), \tag{6.5}$$

where $\nu_k = \frac{1}{m}\sum_j (r_j^{(k)})^2 \delta_{\theta_j^{(k)}}$. With the mirror retraction, $r_i^{(k+1)} = r_i^{(k)} e^{-2\alpha J'_{\nu_k}(\theta_i^{(k)})}$: mass grows where $\eta_{\nu_k} > 1$ and decays where $\eta_{\nu_k} < 1$.

A retraction maps a point and a tangent step to a new point; it is the manifold version of $x + v$. The mirror retraction is $\mathrm{Ret}_{(r,\theta)}(\delta r, \delta\theta) = (re^{\delta r/r}, \widetilde{\mathrm{Ret}}_\theta(\delta\theta))$, with a retraction $\widetilde{\mathrm{Ret}}$ of Θ. On $\mathbb{T}^d$, take $\widetilde{\mathrm{Ret}}_\theta(\delta\theta) = \theta + \delta\theta$ modulo $\mathbb{Z}^d$. The masses $(r_i)^2/m$ are then multiplied by $e^{-4\alpha J'_{\nu_k}(\theta_i)}$, a multiplicative weights step. It never creates mass where there is none. This is why the global result below asks for a spread initialisation.

CPGD on $\mathbb{T}^d$, mirror retraction, BLASSO data fit.
Input: data y, penalty λ, steps $\alpha, \beta > 0$, number of particles m, period n_{gap}.

1. Initialise: $\theta_i \sim$ uniform on $\mathbb{T}^d$ (independent), $r_i \leftarrow r_0 > 0$, for $i = 1, \dots, m$.
2. For $k = 0, 1, 2, \dots$:
   (a) $\nu \leftarrow \frac{1}{m}\sum_j r_j^2 \delta_{\theta_j}$ and residual $u \leftarrow \Phi\nu - y$.
   (b) For each i: $g_i \leftarrow \langle \varphi(\theta_i),\, u\rangle_{\mathbb{H}} + \lambda$ and $G_i \leftarrow \langle \nabla\varphi(\theta_i),\, u\rangle_{\mathbb{H}}$.
   (c) For each i: $r_i \leftarrow r_i\, e^{-2\alpha g_i}$ and $\theta_i \leftarrow \theta_i - \beta G_i$ modulo $\mathbb{Z}^d$.
   (d) Every n_{gap} iterations, run the duality gap test (one-sided form) on ν.

Line (b) computes J'_ν and its gradient at the particles. One iteration costs $\mathcal{O}(m)$ evaluations of φ and $\nabla\varphi$, plus the residual.

Assumption 6.1 (Non-degeneracy of $\nu^\star$)**.** $\nu^\star = \sum_{i=1}^{m^\star} r_i^2 \delta_{\theta_i^\star}$ is the unique global minimiser of J, and, with $f^\star = \int \varphi \,\mathrm{d}\nu^\star$:

(nd1) $\nabla^2 F(f^\star)$ is positive definite;

(nd2) the matrix K with entries $K_{(i,j),(i',j')} = \langle r_i \bar\nabla_j \varphi(\theta_i^\star),\, r_{i'} \bar\nabla_{j'} \varphi(\theta_{i'}^\star)\rangle_{\nabla^2 F(f^\star)}$, where $\bar\nabla\varphi \overset{\text{def.}}{=} (2\varphi, \nabla\varphi)$, satisfies $\sigma_{\min}(K) > 0$;

(nd3) $H_i \overset{\text{def.}}{=} \nabla^2 J'_{\nu^\star}(\theta_i^\star)$ satisfies $\sigma_{\min}(H_i) > 0$ for all i;

(nd4) $J'_{\nu^\star}$ vanishes only at the $\theta_i^\star$, and every other local minimum θ of $J'_{\nu^\star}$ satisfies $J'_{\nu^\star}(\theta) \geqslant v^\star$ for some $v^\star > 0$.

For the BLASSO, (nd1) holds and $H_i = -\lambda\nabla^2\eta_{\nu^\star}(\theta_i^\star)$; we read (nd2) to (nd4) as full rank of Γ_{x_0} and the NDSC (Theorem 4.2) at the working λ.

The scalar product in (nd2) is $\langle u,\, v\rangle_{\nabla^2 F(f^\star)} \overset{\text{def.}}{=} \langle u,\, \nabla^2 F(f^\star) v\rangle_{\mathbb{H}}$, and j runs over the mass coordinate and the d position coordinates, so K has size $m^\star(1+d)$. The weights r_i^2 of $\nu^\star$ are not divided by m. On $\Theta = \mathbb{T}$, (nd2) is the injectivity of the operator Γ of Section 3.2 at the points $\theta_i^\star$, and (nd3) with (nd4) say that $\eta_{\nu^\star} < 1$ outside the spikes and $\eta''_{\nu^\star} < 0$ at the spikes: the one-sided form of (6.2) for the certificate of $\nu^\star$. The bound $\eta_{\nu^\star} > -1$ is not required on non-negative measures.

Theorem 6.3 (Sharpness / PL inequality)**.** *Let* $\|g_\nu\|^2_{L^2(\nu)} \overset{\text{def.}}{=} \int_\Theta \left(4\alpha\,|J'_\nu|^2 + \beta\,\|\nabla J'_\nu\|_\theta^2\right)\mathrm{d}\nu$. *Under Assumption 6.1, there are* $J_0 > J^\star$ *and* $c_{\mathrm{PL}} > 0$ *such that, for every* $\nu \in \mathcal{M}_+(\Theta)$ *with* $J(\nu) \leqslant J_0$ *and every* $\alpha, \beta > 0$, $\frac{1}{2}\|g_\nu\|^2_{L^2(\nu)} \geqslant c_{\mathrm{PL}} \min(\alpha,\beta)(J(\nu) - J^\star)$. c_{PL}^{-1} *and* $(J_0 - J^\star)^{-1}$ *depend polynomially on the problem characteristics.*

The quantity $\|g_\nu\|^2_{L^2(\nu)}$ is the squared norm of the gradient of J for the cone metric. The theorem is a Polyak-Łojasiewicz inequality: near $\nu^\star$, the gradient controls the excess objective without convexity in the particles. The problem characteristics are λ, the smoothness of φ and F, the ratio $\max_i r_i / \min_i r_i$, the inverses of $\sigma_{\min}(\nabla^2 F(f^\star))$, $\sigma_{\min}(K)$ and $\sigma_{\min}(H_i)$, and $1/v^\star$. Since Assumption 6.1 is stated with unit weights, J_0 and c_{PL} do not depend on (α, β).

A descent lemma [24] gives, for every level $J_{\max} \geqslant J^\star$, a threshold $\tau_{\max} > 0$ such that, if $J(\nu_0) \leqslant J_{\max}$ and $\max(\alpha,\beta) \leqslant \tau_{\max}$, the iterates (6.5) are defined for all k and $J(\nu_{k+1}) - J(\nu_k) \leqslant -\frac{1}{2}\|g_{\nu_k}\|^2_{L^2(\nu_k)}$. The lemma gives no value of $\tau_{\max}$.

Corollary 6.2 (Exponential local convergence)**.** *Let J_0, c_{PL} be as in Theorem 6.3 and $\tau_{\max} = \tau_{\max}(J_0)$ the step-size threshold of the descent lemma at $J_{\max} = J_0$. Under Assumption 6.1, if $J(\nu_0) \leqslant J_0$, every solution of (6.4) satisfies $J(\nu_t) - J^\star \leqslant (J(\nu_0) - J^\star)e^{-2c_{\mathrm{PL}} \min(\alpha,\beta)t}$ for all $\alpha, \beta > 0$, and, if $\max(\alpha, \beta) \leqslant \tau_{\max}$, the iterates of (6.5) satisfy $J(\nu_k) - J^\star \leqslant (J(\nu_0) - J^\star)(1 - c_{\mathrm{PL}} \min(\alpha, \beta))^k$.*

Both dynamics decrease J, so they stay in the sublevel set where the PL inequality holds. Along the flow, $\frac{\mathrm{d}}{\mathrm{d}t} J(\nu_t) = -\left\|g_{\nu_t}\right\|^2_{L^2(\nu_t)}$, and Gronwall's lemma gives the first bound. For the iterates, the descent lemma and the PL inequality give $J(\nu_{k+1}) - J^\star \leqslant (1 - c_{\mathrm{PL}} \min(\alpha, \beta))(J(\nu_k) - J^\star)$. If $J(\nu_0) > J^\star$, this forces $1 - c_{\mathrm{PL}} \min(\alpha, \beta) \geqslant 0$.

The local result needs a start in the sublevel set $\{\nu \,;\, J(\nu) \leqslant J_0\}$, which is not known in practice. The global result replaces it by a start close to a reference measure $\nu_\star$ with full support, and a small ratio β/α, so that the Fisher-Rao part dominates. The lower star marks this smooth reference, the upper star the minimiser $\nu^\star$. Let $W_\infty(\mu, \nu) \overset{\text{def.}}{=} \inf_\pi \pi\text{-ess sup}\, \mathrm{dist}(\theta, \theta')$, the infimum being over the couplings π of μ and ν; it is the ∞-Wasserstein distance.

Theorem 6.4 (Global convergence of WFR gradient flow)**.** *Under Assumption 6.1, let $\nu_\star \in \mathcal{M}_+(\Theta)$ have a smooth positive density with $\log \nu_\star$ Lipschitz. There are constants, polynomial in the problem characteristics, such that if $W_\infty(\nu_0, \nu_\star) \leqslant c_W$ and β/α is small enough, then $J(\nu_t) - J^\star \leqslant (J_0 - J^\star)e^{-2c_{\mathrm{PL}} \min(\alpha,\beta)(t-t_0)}$ for $t \geqslant t_0 = 1/\sqrt{\alpha\beta}$. For CPGD: if $W_\infty(\nu_0, \nu_\star) \leqslant (J_0 - J^\star)/C$, $\alpha \leqslant (J_0 - J^\star)^{1+\varepsilon/2}/C$, $\beta \leqslant (J_0 - J^\star)\alpha^2/C'$, for a fixed $\varepsilon \in (0, 1/2]$ and constants C, C' increasing in $1/\varepsilon$, then $J(\nu_k) - J^\star \leqslant (J(\nu_0) - J^\star)(1 - c_{\mathrm{PL}} \min(\alpha, \beta))^{k-k_0}$ for $k \geqslant k_0 = C(J_0 - J^\star)^{-(2+\varepsilon)}$.*

Remark 6.4 (Reading the global theorem). (a) The flow has two phases. Before t_0, an entropic mirror descent estimate brings ν_t into the sublevel set $\{\nu \,;\, J(\nu) \leqslant J_0\}$; then the PL inequality takes over. (b) Couplings of ν_0 and $\nu_\star$ exist only if $\nu_0(\Theta) = \nu_\star(\Theta)$. This equality is implicit in the statement. In practice, draw the $\theta_i^{(0)}$ independently with law $\nu_\star/\nu_\star(\Theta)$, for instance uniform on $\mathbb{T}^d$, and set $(r_i^{(0)})^2 = \nu_\star(\Theta)$. The condition $W_\infty(\nu_0, \nu_\star) \leqslant c_W$ then asks for m of order c_W^{-d} particles, up to logarithmic factors [25]. This number grows exponentially with d. It is the price of dropping the global search. (c) The constants c_W, C, C' and the condition on β/α are not explicit. (d) The bound for CPGD is written with $J(\nu_0) - J^\star$. The contraction starts at step k_0, where $J(\nu_{k_0}) \leqslant J_0$; the factor $J(\nu_0) - J^\star$ uses in addition $J(\nu_{k_0}) \leqslant J(\nu_0)$, the decrease of J along the first k_0 iterates.

To stop CPGD, use the residual p_{ν_k} normalised as in Remark 6.1, or by the one-sided normalisation of Remark 6.2, which bounds $J(\nu_k) - J^\star$ with $J^\star = \mathcal{L}^\star_+$.

6.4 FastPart: stochastic particle descent

CPGD evaluates J'_ν and $\nabla J'_\nu$ exactly at each particle. When the scalar product of $\mathbb{H}$ is an integral, as in $\mathbb{H} = \mathrm{L}^2(\Lambda)$ for the Gaussian mixture of Example 2.2, one replaces these values by unbiased mini-batch estimates, for instance by sampling frequencies $\omega \sim \Lambda$ (random features). This section gives the setting and the stochastic update that FastPart and Fast Spawn and Prune share.

Here $\mathcal{X} \subset \mathbb{R}^d$ is compact and convex, $K(s, t) \overset{\text{def.}}{=} \langle \varphi_s, \varphi_t \rangle_{\mathbb{H}}$, the BLASSO is posed on $\mathcal{M}_+(\mathcal{X})$, and $J'_\nu(t) = \langle \varphi_t, \Phi\nu - y \rangle_{\mathbb{H}} + \lambda$ [26]. The kernel K is the kernel K of Definition 2.2; in this section and the next one, the letter K never denotes the matrix of (nd2). We write J for the restriction of $\mathcal{L}$ to $\mathcal{M}_+(\mathcal{X})$. By [26], $\nu \in \mathcal{M}_+(\mathcal{X})$ minimises J if and only if $J'_\nu \geqslant 0$ on $\mathcal{X}$ and $J'_\nu = 0$ on $\mathrm{supp}(\nu)$.

Assumption 6.2 (Kernel and observation regularity)**.** There are $K_{\min}, C_{\mathrm{P}} > 0$ such that $\|y\|_{\mathbb{H}} \leqslant C_{\mathrm{P}}$ and

1. $\max(\|K\|_\infty, \|\nabla_s K\|_\infty, \|\nabla_s^2 K\|_\infty) \leqslant C_{\mathrm{P}}$ and $\|\nabla_s^2 \nabla_t^2 K\|_\infty \leqslant C_{\mathrm{P}}^2$;

2. $K(s,t) \geqslant K_{\min} > 0$ and $K(t,t) = 1$ for all $s, t \in \mathcal{X}$.

Since $K(t,t) = 1 \leqslant \|K\|_\infty$, the assumption forces $C_{\mathrm{P}} \geqslant 1$. The lower bound holds for the Gaussian kernel of Example 2.2 restricted to a compact $\mathcal{X}$: $K(s,t) = e^{-(s-t)^2/(6\sigma^2)} \geqslant e^{-\mathrm{diam}(\mathcal{X})^2/(6\sigma^2)}$. It fails for Example 2.1: with uniform weights the kernel is a multiple of the Dirichlet kernel and takes negative values; with the Jackson weights it vanishes at $u = 1/(\frac{f_c}{2}+1)$. There the domain $\mathbb{T}$ is not a convex subset of $\mathbb{R}^d$ either. It also fails for signed measures: on the doubled dictionary the kernel between the two copies is $-K$. SFW and CPGD need no sign condition on the kernel.

Assumption 6.3 (Stochastic unbiased gradients)**.** There is a random variable Z with $\widehat{J'_\nu}(t,Z) = J'_\nu(t) + \xi_\nu(t,Z)$ and $\widehat{D_\nu}(t,Z) = \nabla J'_\nu(t) + \zeta_\nu(t,Z)$, where ξ_ν, ζ_ν have zero mean, and constants $\boldsymbol{H}, \boldsymbol{G}, \boldsymbol{M} > 0$ such that almost surely $|\xi_\nu| \vee \|\zeta_\nu\| \leqslant \boldsymbol{M}$ and $\widehat{J'_\nu}(t,Z) \geqslant \boldsymbol{G}\|\nu\|_{\mathrm{TV}} - \boldsymbol{H} + \lambda$. Also, $t \mapsto \widehat{J'_\nu}(t,Z)$ is $\mathfrak{L}$-Lipschitz almost surely.

Remark 6.5 (On the stochastic assumption). (a) The lower bound comes from the positivity of the kernel. For $\nu \in \mathcal{M}_+(\mathcal{X})$ and $t \in \mathcal{X}$, since $\|\varphi_t\|_{\mathbb{H}} = 1$, $J'_\nu(t) = \int_{\mathcal{X}} K(t,s)\,\mathrm{d}\nu(s) - \langle \varphi_t, y\rangle_{\mathbb{H}} + \lambda \geqslant K_{\min}\|\nu\|_{\mathrm{TV}} - \|y\|_{\mathbb{H}} + \lambda$. With noise of amplitude at most $\boldsymbol{M}$, the bound holds with $\boldsymbol{G} = K_{\min}$ and $\boldsymbol{H} = \|y\|_{\mathbb{H}} + \boldsymbol{M}$. So $\boldsymbol{M}$ is chosen first, then $\boldsymbol{H}$. (b) The constants are uniform in ν, which is too strong on all of $\mathcal{M}_+(\mathcal{X})$: the Lipschitz constant of J'_ν is only bounded by $\sqrt{C_{\mathrm{P}}}(C_{\mathrm{P}} + \|\nu\|_{\mathrm{TV}})$ [26]. The analysis uses the assumption along the iterates, which stay in $\{\nu\ ;\ \|\nu\|_{\mathrm{TV}} \leqslant C_{\mathrm{TV}}\}$ by Proposition 6.1. Read the assumption on this set. (c) The bounds are assumed almost surely.

Definition 6.5 (Stochastic Push-Forward update)**.** Let $\hat\nu_k = \sum_{j=1}^{p_k} \omega_j^k \delta_{t_j^k}$, and let $\widehat{J'_{\hat\nu_k}}$ and $\widehat{D_{\hat\nu_k}}$ be the averages of $\widehat{J'_{\hat\nu_k}}(\cdot, Z)$ and $\widehat{D_{\hat\nu_k}}(\cdot, Z)$ over m_k independent copies of Z. Set $\hat\nu_{k^+} \overset{\text{def.}}{=} \sum_{j=1}^{p_k} \omega_j^{k^+} \delta_{t_j^{k^+}}$ with

$$\omega_j^{k^+} = \omega_j^k\, e^{-\alpha \widehat{J'_{\hat\nu_k}}(t_j^k)}, \qquad t_j^{k^+} = \operatorname*{argmin}_{u \in \mathcal{X}} \left\{ \langle u,\, \widehat{D_k}(t_j^k)\rangle + \tfrac{1}{2\beta} \left\| u - t_j^k \right\|^2 \right\}.$$

The position step projects $t_j^k - \beta \widehat{D_k}(t_j^k)$ onto $\mathcal{X}$; a step costs $\mathcal{O}(p_k m_k)$. With exact J' and $\nabla J'$, this update decreases J under (6.6) below while $\|\nu\|_{\mathrm{TV}} \leqslant C_{\mathrm{TV}}$ [26]. FastPart (arXiv:2312.05993) iterates this update on a fixed number of particles, from a spread initialisation, with no birth and no death.

The projection claim follows by completing the square: $\langle u, v\rangle + \frac{1}{2\beta}\|u - t\|^2 = \frac{1}{2\beta}\|u - (t - \beta v)\|^2 + c$, where c does not depend on u, and $\mathcal{X}$ is closed and convex. The weight step is the mirror step of CPGD written on the masses: CPGD multiplies the mass r^2/m by $e^{-4\alpha J'}$, FastPart multiplies ω by $e^{-\alpha J'}$. So the letter α does not have the same meaning in the two sections.

FastPart (stochastic weight and push-forward).
Input: steps $\alpha, \beta > 0$, batch sizes (m_k), initial particles $(\omega_j^0, t_j^0)_{j \leqslant p}$ spread over $\mathcal{X}$, with $\omega_j^0 > 0$.
For $k = 0, 1, 2, \ldots$:

1. Draw m_k independent copies $Z_1, \ldots, Z_{m_k}$ of Z.
2. For each j: $\hat g_j \leftarrow \frac{1}{m_k}\sum_\ell \widehat{J'_{\hat\nu_k}}(t_j^k, Z_\ell)$ and $\hat G_j \leftarrow \frac{1}{m_k}\sum_\ell \widehat{D_{\hat\nu_k}}(t_j^k, Z_\ell)$.
3. For each j: $\omega_j^{k+1} \leftarrow \omega_j^k e^{-\alpha \hat g_j}$ and $t_j^{k+1} \leftarrow \mathrm{proj}_{\mathcal{X}}(t_j^k - \beta \hat G_j)$.
4. From time to time, run the duality gap test (one-sided form) with the exact residual.

As for CPGD, no mass appears away from the initial particles and their trajectories.

To stop FastPart, use the exact residual $p_{\hat\nu_k}$, not its estimate, normalised as in Remark 6.1 or Remark 6.2. The estimate $\widehat{J'}$ has noise of amplitude up to $\boldsymbol{M}$, and a normalisation by an estimated maximum does not guarantee feasibility.

6.5 Fast Spawn and Prune

The weight step never puts mass where there is none, so the stochastic update can stall at a measure whose first variation is negative away from its support. Fast Spawn and Prune (FS&P) adds two random moves to each iteration. A birth samples one point uniformly on $\mathcal{X}$ and puts a small mass there if the estimated first variation is nonpositive, up to the noise level. A death samples one particle and removes it if its weight is small and its estimated first variation is nonnegative. In optimisation terms, FS&P is a stochastic Frank-Wolfe method: the birth replaces the global search by one uniform sample, the death plays the role of a drop step, and the update of Definition 6.5 plays the role of sliding.

Definition 6.6 (FS&P iteration)**.** Fix $\alpha, \beta > 0$, batch sizes m_k, weights $\varepsilon_k > 0$, $a > 0$, and let $c_a \overset{\text{def.}}{=} \boldsymbol{M}\sqrt{2a}$. From $\hat{\nu}_k$:

1. apply Definition 6.5, which gives $\hat{\nu}_{k^+}$, and let $\widehat{J}'$ be the mini-batch estimate of $J'_{\hat{\nu}_{k^+}}$;
2. death: sample V_{k+1} uniformly on $\operatorname{supp}(\hat{\nu}_{k^+})$; remove the particle at V_{k+1} if $\widehat{J}'(V_{k+1}) \geqslant 0$ and its weight is at most $\sqrt{2}\,\varepsilon_k$;
3. birth: sample U_{k+1} uniformly on $\mathcal{X}$; add $\varepsilon_k \delta_{U_{k+1}}$ if $\widehat{J}'(U_{k+1}) \leqslant c_a\sqrt{\log(m_k)/m_k}$.

The threshold $c_a\sqrt{\log(m_k)/m_k}$ is a noise margin. At a point drawn independently of the mini-batch, the average of m_k independent centred terms bounded by $\boldsymbol{M}$ exceeds $c_a\sqrt{\log(m_k)/m_k}$ in absolute value with probability at most $2\exp\big(-m_k c_a^2 \log(m_k)/(2m_k\boldsymbol{M}^2)\big) = 2m_k^{-a}$, by Hoeffding's inequality. So a point where $J'_\nu \leqslant 0$ passes the birth test with probability at least $1 - 2m_k^{-a}$. Births also occur where J'_ν is slightly positive. If $\hat{\nu}_{k^+} = 0$, for instance at the start $\hat{\nu}_0 = 0$, the death step is skipped. One step needs the $\mathcal{O}(p_k m_k)$ evaluations of Definition 6.5 and two more mini-batch evaluations, at V_{k+1} and U_{k+1}.

Fast Spawn and Prune.
Input: steps α, β, horizon T, batch sizes (m_k), birth masses (ε_k), $a > 0$, an initial measure $\hat{\nu}_0$ with few particles (possibly none).
For $k = 0, 1, \ldots, T-1$:

1. Push-forward: compute $\hat{\nu}_{k^+}$ by lines 1 to 3 of FastPart.
2. Fresh estimate: let $\widehat{J}'$ be a mini-batch average of size m_k for $J'_{\hat{\nu}_{k^+}}$.
3. Death: draw a particle V of $\hat{\nu}_{k^+}$ uniformly. If $\widehat{J}'(V) \geqslant 0$ and its weight is at most $\sqrt{2}\,\varepsilon_k$, remove it.
4. Birth: draw U uniformly on $\mathcal{X}$. If $\widehat{J}'(U) \leqslant c_a\sqrt{\log(m_k)/m_k}$, add the particle $\varepsilon_k\delta_U$.
5. Call the result $\hat{\nu}_{k+1}$. From time to time, run the duality gap test (one-sided form).

Output: the iterate with the smallest gap.

Proposition 6.1 (TV-norm boundedness)**.** *Let* $\widehat{\mathfrak{R}} \overset{\text{def.}}{=} e\boldsymbol{H}/\boldsymbol{G} + \sqrt{e^3/\boldsymbol{G}} + 1$. *If* $\sup_k \varepsilon_k \leqslant \alpha \leqslant (1+\widehat{\mathfrak{R}})^{-1}$ *and* $\alpha(\boldsymbol{H} - \lambda)_+ \leqslant 1/10$, *then almost surely, for all* $k \geqslant 0$, $\|\hat{\nu}_k\|_{\mathrm{TV}} \leqslant \widehat{\mathfrak{R}} \vee \|\hat{\nu}_0\|_{\mathrm{TV}} \overset{\text{def.}}{=} C_{\mathrm{TV}}$.

The mechanism is a balance. When $\|\hat{\nu}_k\|_{\mathrm{TV}} > \widehat{\mathfrak{R}}$, the lower bound of Assumption 6.3 makes the estimated first variation positive on all of $\mathcal{X}$, so the weight step shrinks every mass. The position step keeps the total mass, a death removes mass, and a birth adds at most $\varepsilon_k \leqslant \alpha$. Below, $\mu^\star$ is a minimiser of J on $\mathcal{M}_+(\mathcal{X})$. It exists: the sublevel sets of $\mathcal{L}$ are weak-$\star$ compact (Proposition 2.2) and $\mathcal{M}_+(\mathcal{X})$ is weak-$\star$ closed.

Theorem 6.5 (Global convergence of FS&P)**.** *Assume Assumptions 6.2 and 6.3, the step-size conditions*

$$\alpha < \frac{1}{10(1+C_{\mathrm{TV}}+C_{\mathrm{P}}+\lambda)(1\vee C_{\mathrm{TV}})}, \qquad \beta \leqslant \frac{1}{2C_{\mathrm{P}}(C_{\mathrm{P}}+3C_{\mathrm{TV}})e^{1/5}}, \tag{6.6}$$

the four empirical birth and death hypotheses of [26, §4.1], and the hypotheses of Proposition 6.1. Take $\alpha \leqslant \sqrt{8\log 8}/\boldsymbol{M}$, $\beta \leqslant \alpha^{-d/4}T^{-1/2}$, $\varepsilon_k = T^{-1/2}$, $m_k = T$ *and* $a \geqslant d/(2(2+d))$. *Then for every horizon* $T \geqslant 1/\alpha^2$, $\hat{\rho}_T \overset{\text{def.}}{=} \min_{1\leqslant k\leqslant T} J(\hat{\nu}_k) - J(\mu^\star)$ *satisfies*

$$\mathbb{E}[\hat{\rho}_T] \leqslant C\alpha^{-1/2}\Big(\frac{(\log T)^2}{T}\Big)^{1/(2(2+d))}.$$

Remark 6.6 (Reading the theorem). (a) The first of the four empirical hypotheses is a stochastic descent inequality for the update of Definition 6.5; the other three are conditions on the birth and death steps, with J'_ν replaced by its estimate. (b) The schedule fixes $\varepsilon_k = T^{-1/2}$ for the whole run, and $T \geqslant 1/\alpha^2$ gives $\varepsilon_k \leqslant \alpha$, as Proposition 6.1 requires. The constant C depends on the problem constants, not on T. (c) The bound concerns the best iterate in expectation, not the last iterate. The gap test with the exact residual identifies a good iterate, at the cost of a global search on a net. (d) No non-degeneracy and no spread start are assumed, unlike Theorem 6.4: the start can be sparse, for instance $\hat{\nu}_0 = 0$. The price is the exponent $1/(2(2+d))$. If J'_ν has a negative minimum $-v$ and is $\mathfrak{L}$-Lipschitz, the set $\{J'_\nu < 0\}$ contains the intersection of $\mathcal{X}$ with a ball of radius $v/\mathfrak{L}$; a uniform birth must hit this set, whose volume is of order $(v/\mathfrak{L})^d$. (e) Reaching $\mathbb{E}[\hat{\rho}_T] \leqslant \varepsilon$ with this bound takes T of order $\varepsilon^{-2(2+d)}$, up to logarithms and constants. For $d = 2$ and $\varepsilon = 10^{-3}$, the sample complexity bound is of order 10^{48} evaluations. The bound is a worst case.

To stop FS&P, use the exact residual $p_{\hat{\nu}_k}$, with $\Phi^\star p_{\hat{\nu}_k} = 1 - J'_{\hat{\nu}_k}/\lambda$, normalised as in Remark 6.1 or Remark 6.2. The global search that FS&P avoids comes back each time the gap is evaluated.

Exercises

Exercise 6.1 (The Frank-Wolfe step)**.** Let $B \overset{\text{def.}}{=} \sup_x \|\varphi_x\|_{\mathbb{H}} > 0$. Let $\mu \in \mathcal{M}(\mathcal{X})$, $r \overset{\text{def.}}{=} y - \Phi\mu$, $\eta \overset{\text{def.}}{=} \Phi^\star r/\lambda$, and $x \in \mathcal{X}$ with $|\eta(x)| > 1$. Let $\varsigma \overset{\text{def.}}{=} \operatorname{sign}(\eta(x))$ and, for $\theta \geqslant 0$,

$$f(\theta) \overset{\text{def.}}{=} \mathcal{L}(\mu + \theta\varsigma\delta_x), \qquad g(\theta) \overset{\text{def.}}{=} \tfrac{1}{2}\|r - \theta\varsigma\varphi_x\|^2_{\mathbb{H}} + \lambda(\|\mu\|_{\mathrm{TV}} + \theta).$$

1. Show that $f \leqslant g$ on $[0,\infty)$, with $f(0) = g(0)$, and that $f = g$ if $\varsigma\mu(\{x\}) \geqslant 0$.
2. Show that $\varphi_x \neq 0$ and that g is minimal at $\bar{\theta} \overset{\text{def.}}{=} \lambda(|\eta(x)| - 1)/\|\varphi_x\|^2_{\mathbb{H}}$, with $g(0) - g(\bar{\theta}) = \lambda^2(|\eta(x)| - 1)^2/(2\|\varphi_x\|^2_{\mathbb{H}})$.
3. Show that f is convex and has a minimiser $\theta^\star > 0$ on $[0,\infty)$. Deduce that the iteration of Definition 6.2 with $\theta_k \overset{\text{def.}}{=} \theta^\star$ satisfies $\mathcal{L}(\mu_{k+1}) \leqslant \mathcal{L}(\mu_k) - \lambda^2(\|\eta_k\|_\infty - 1)^2/(2B^2)$ as long as it does not stop, and that $\sum_k(\|\eta_k\|_\infty - 1)^2 \leqslant B^2\|y\|^2_{\mathbb{H}}/\lambda^2$.
4. Let $s \overset{\text{def.}}{=} \|\eta\|_\infty > 1$. With question 3 and Corollary 6.1, show that
$$\frac{\lambda^2(s-1)^2}{2B^2} \leqslant \mathcal{L}(\mu) - \mathcal{L}^\star \leqslant \Delta(\mu, r/(\lambda s)).$$
5. One spike. Let $\mathcal{X} = \mathbb{T}$, $K(x,x') = \kappa(x - x')$ with κ continuous, $\kappa(0) = 1$ and $|\kappa(u)| < 1$ for $u \neq 0$. Let $y = a\varphi_{x_0}$ with $|a| > \lambda$. Run Definition 6.2 from $\mu_0 = 0$. Show that $x^\star = x_0$, that $\mu_1 = (a - \lambda\operatorname{sign}(a))\delta_{x_0}$, and that the iteration stops at $k = 1$. Compare with Exercise 2.5.

Exercise 6.2 (Duality gap on the cone)**.** Let $\mathbb{H} = \mathbb{R}$, $\varphi_x = 1$ for all $x \in \mathcal{X}$, $y = -1$ and $0 < \lambda < 1$. So $\Phi\mu = \mu(\mathcal{X})$ and $\Phi^\star p$ is the constant function p.

1. Show that $\mathcal{L}^\star = \lambda - \lambda^2/2$ and that $\mathcal{L}^\star_+ = \frac{1}{2}$, attained at $\nu = 0$ only.

2. At $\nu = 0$, compute p_ν. Show that the sup norm normalisation of Remark 6.1 gives $p = -1$ and $\Delta(0, -1) = \frac{1}{2}(1-\lambda)^2 > 0$. Is the bound $\mathcal{L}(0) - \mathcal{L}^\star \leqslant \Delta(0, -1)$ tight?

3. Show that the one-sided normalisation of Remark 6.2 gives $q = -1/\lambda$, and that the right-hand side of (6.1) at $(0, -1/\lambda)$ is 0. Which conclusion does this give on $\nu = 0$?

4. General case. Let $\nu^\star$ minimise $\mathcal{L}$ on $\mathcal{M}_+(\mathcal{X})$. By [26], $J'_{\nu^\star} = \lambda(1 - \Phi^\star p_{\nu^\star}) \geqslant 0$ on $\mathcal{X}$ and $J'_{\nu^\star} = 0$ on $\mathrm{supp}(\nu^\star)$. Show that the one-sided normalisation leaves $p_{\nu^\star}$ unchanged, and that the right-hand side of (6.1) at $(\nu^\star, p_{\nu^\star})$ is 0.

Exercise 6.3 (Non-degeneracy of the particle problem)**.** Let $\Theta = \mathbb{T}$, $F(f) = \frac{1}{2}\|f - y\|^2_{\mathbb{H}}$, $\theta \mapsto \varphi(\theta)$ of class $\mathscr{C}^2$, and $\nu^\star = \sum_{i=1}^{m^\star} r_i^2 \delta_{\theta_i^\star}$ a minimiser of J on $\mathcal{M}_+(\mathbb{T})$ with $r_i > 0$ and distinct $\theta_i^\star$. Let $\eta^\star \overset{\text{def.}}{=} \eta_{\nu^\star}$.

1. Let $\nu, \nu+\sigma \in \mathcal{M}_+(\mathbb{T})$. Show that $J(\nu+\sigma) - J(\nu) = \langle J'_\nu, \sigma\rangle_{\mathcal{M}} + \frac{1}{2}\|\Phi\sigma\|^2_{\mathbb{H}}$ with $J'_\nu = \lambda(1-\eta_\nu)$. Deduce that $J'_{\nu^\star} \geqslant 0$ on $\mathbb{T}$ and $J'_{\nu^\star}(\theta_i^\star) = 0$ for all i. Hint: take $\sigma = \tau\delta_\theta$ and $\sigma = -\tau r_i^2 \delta_{\theta_i^\star}$, $\tau \in (0, 1]$.

2. Let $\Gamma \overset{\text{def.}}{=} \Gamma_{\theta^\star}$ be the operator of Section 3.2 at the points $\theta_i^\star$. Show that $K = \Lambda\Gamma^\star\Gamma\Lambda$ with $\Lambda \overset{\text{def.}}{=} \mathrm{diag}(2r_1, \dots, 2r_{m^\star}, r_1, \dots, r_{m^\star})$, after a permutation of rows and columns. Deduce that (nd2) holds if and only if Γ is injective.

3. With question 1, show that (nd3) holds if and only if $(\eta^\star)''(\theta_i^\star) < 0$ for all i.

4. Assume (nd3) and $\eta^\star(\theta) < 1$ for $\theta \notin \{\theta_i^\star\}$. Show that there is $\rho > 0$ such that $J'_{\nu^\star}$ is strictly convex on each interval $(\theta_i^\star - \rho, \theta_i^\star + \rho)$. Deduce (nd4).

5. Conclude that, for the BLASSO on $\mathbb{T}$, (nd2) to (nd4) hold if and only if Γ is injective, $\eta^\star < 1$ outside $\{\theta_i^\star\}$ and $(\eta^\star)''(\theta_i^\star) < 0$ for all i. Compare with (6.2): which condition of the signed problem is not needed?

Appendix A

Support stability without the implicit function theorem

The proof of Theorem 4.2 in Lecture 4 admits the implicit function theorem (Step 3) and uses it for uniqueness and for the rates. This appendix proves uniqueness, the number of spikes, the signs and the rates with the tools of Lectures 2 to 4 only. Proposition A.1 is Step 2 of that proof, with all details. Proposition A.2 tests a solution against the minimal norm certificate η_0 and against the secondary certificates of Remark 3.4 (c). It gives explicit rates. It does not give the regularity of the path.

The setting is that of Section 4.3: $\mathcal{X} = \mathbb{T}$, $\mathbb{H} = \mathrm{L}^2(\mathbb{T})$ real, $\Phi\mu = \varphi \star \mu$ with $\varphi \in \mathscr{C}^2(\mathbb{T})$, $y_0 \overset{\text{def.}}{=} \Phi\mu^0$, $\mu^0 = \sum_{i=1}^s a_i \delta_{x_i}$ satisfies (NDSC) and Γ_{x_0} is injective. We write $\sigma_i \overset{\text{def.}}{=} \mathrm{sign}(a_i)$, as in the proof of Theorem 4.2, and $M_k \overset{\text{def.}}{=} \left\|\varphi^{(k)}\right\|_{\mathrm{L}^2}$ for $k \leqslant 2$. All norms and scalar products without index are those of $\mathbb{H}$, and $|u - v|$ is the distance on $\mathbb{T}$. Near a point, we see a function on $\mathbb{T}$ as a 1-periodic function on $\mathbb{R}$, as in Remark 3.3.

By (NDSC), (SC) holds. So p_0 and $\eta_0 = \Phi^\star p_0$ are defined (Lemma 3.1), and $\eta_0 \in \partial\|\mu^0\|_{\mathrm{TV}}$ since $p_0 \in P$. As in Section 4.3, Remark 2.4 gives $\eta_0 \in \mathscr{C}^2(\mathbb{T})$. By Remark 3.3, $\eta_0(x_i) = \sigma_i$ and $\eta_0'(x_i) = 0$. For $(\lambda, w) \in D_{\alpha,\lambda_0}$ we write $\hat{p} \overset{\text{def.}}{=} \hat{p}_\lambda(y_0 + w)$ and $\tilde{\eta}_\lambda \overset{\text{def.}}{=} \Phi^\star\hat{p}$, as in Theorem 4.1. Since $\hat{p} \in C$, $\|\tilde{\eta}_\lambda\|_\infty \leqslant 1$. Let

$$\kappa \overset{\text{def.}}{=} \tfrac{1}{2} \min_i |\eta_0''(x_i)| .$$

By (NDSC), $\sigma_i\eta_0''(x_i) = -|\eta_0''(x_i)| \leqslant -2\kappa < 0$ for all i.

Lemma A.1 (Choice of ε)**.** *There is $\varepsilon \in (0, \frac{1}{4})$ with $2\varepsilon < |x_i - x_j|$ for all $i \neq j$ such that, with $I_i \overset{\text{def.}}{=} (x_i - \varepsilon, x_i + \varepsilon)$,*

$$\sigma_i\eta_0'' \leqslant -\kappa \quad \textit{and} \quad \sigma_i\eta_0 \geqslant \tfrac{1}{2} \quad \textit{on } I_i, \qquad i = 1, \dots, s. \tag{A.1}$$

Every smaller positive number has the same properties. For such an ε:

(a) the intervals I_i are pairwise disjoint;

(b) $1 - |\eta_0(x_i + t)| \geqslant \frac{\kappa}{2} t^2$ for all i and $|t| < \varepsilon$;

(c) $r_0 \overset{\text{def.}}{=} 1 - \max\{|\eta_0(x)| \ ; \ x \in \mathbb{T} \setminus \bigcup_i I_i\} > 0$, with $r_0 \overset{\text{def.}}{=} 1$ if $\mathbb{T} \setminus \bigcup_i I_i$ is empty.

These are the conditions on ε in Step 2 of the proof of Theorem 4.2.

Proof. Existence. The functions η_0 and η_0'' are continuous, $\sigma_i\eta_0(x_i) = 1$ and $\sigma_i\eta_0''(x_i) \leqslant -2\kappa$. So (A.1) holds on a neighbourhood of each x_i, hence for all small ε. The two other conditions also hold for ε small. All of them are kept when ε decreases.

(a) This follows from $2\varepsilon < |x_i - x_j|$ for $i \neq j$. Since $\varepsilon < \frac{1}{2}$, each $x \in I_i$ reads $x = x_i + t$ with a unique $t \in (-\varepsilon, \varepsilon)$, and $|x - x_i| = |t|$.

(b) Let $\psi \overset{\text{def.}}{=} \sigma_i \eta_0$ and $|t| < \varepsilon$. Then $\psi(x_i) = 1$ and $\psi'(x_i) = 0$. The Taylor formula with integral remainder and (A.1) give

$$\psi(x_i + t) = 1 + t^2 \int_0^1 (1 - \theta)\, \psi''(x_i + \theta t)\, \mathrm{d}\theta \leqslant 1 - \tfrac{\kappa}{2} t^2.$$

Since $\psi \geqslant \frac{1}{2}$ on I_i, $|\eta_0| = \psi$ there, and (b) follows.

(c) The set $K \overset{\text{def.}}{=} \mathbb{T} \setminus \bigcup_i I_i$ is closed in $\mathbb{T}$, hence compact, and it does not contain the x_i. By (NDSC), $|\eta_0| < 1$ on K. If K is not empty, the continuous function $|\eta_0|$ attains its maximum on K, and this maximum is less than 1. □

Proposition A.1 (Structure of the solutions)**.** *Let ε and I_i be as in Lemma A.1. There are $\alpha, \lambda_0 > 0$ such that, for all $(\lambda, w) \in D_{\alpha,\lambda_0}$, with $Z \overset{\text{def.}}{=} \{x \in \mathbb{T}\,;\, |\tilde\eta_\lambda(x)| = 1\}$:*

1. *Z has at most one point in each I_i and no point outside $\bigcup_i I_i$;*
2. *every solution of $\mathcal{P}_\lambda(y_0 + w)$ is of the form $\hat\mu = \sum_{i=1}^s \hat a_i \delta_{\hat x_i}$ with $\hat x_i \in I_i \cap Z$ when $\hat a_i \neq 0$, and then $\operatorname{sign}(\hat a_i) = \sigma_i$, $\tilde\eta_\lambda(\hat x_i) = \sigma_i$ and $\tilde\eta'_\lambda(\hat x_i) = 0$.*

If $\hat\mu$ has no mass in I_i, we set $\hat a_i \overset{\text{def.}}{=} 0$ and $\hat x_i \overset{\text{def.}}{=} x_i$, as in Step 2 of the proof of Theorem 4.2. So $\hat x_i \in I_i$ for all i.

Proof. Choice of α and λ_0. Let $\tau \overset{\text{def.}}{=} \min(r_0, \frac{1}{2}, \kappa) > 0$. Since $\eta_0(x_1) = \sigma_1 \neq 0$, $\eta_0 \neq 0$ and $\varphi \neq 0$, so $M \overset{\text{def.}}{=} \max(M_0, M_2) > 0$. As in Step 1 of the proof of Theorem 4.2, let $b(\lambda_0) \overset{\text{def.}}{=} \sup_{0<\lambda\leqslant\lambda_0} \|\hat p_\lambda(y_0) - p_0\|$. By Proposition 4.2, $\hat p_\lambda(y_0) \to p_0$ strongly, so $b(\lambda_0) \to 0$ when $\lambda_0 \to 0$. Let $\alpha \overset{\text{def.}}{=} \tau/(4M)$ and choose $\lambda_0 > 0$ with $b(\lambda_0) \leqslant \tau/(4M)$.

Uniform closeness. Let $(\lambda, w) \in D_{\alpha,\lambda_0}$. Then $\|\hat p_\lambda(y_0) - p_0\| \leqslant b(\lambda_0)$ and $\|w\|/\lambda \leqslant \alpha$. By Theorem 4.1, for $k \in \{0, 2\}$,

$$\left\| \tilde\eta_\lambda^{(k)} - \eta_0^{(k)} \right\|_\infty \leqslant M_k (b(\lambda_0) + \alpha) \leqslant \frac{\tau}{2}. \tag{A.2}$$

Item 1, far from the spikes. Let $x \notin \bigcup_i I_i$. By Lemma A.1 (c) and (A.2), $|\tilde\eta_\lambda(x)| \leqslant 1 - r_0 + \frac{r_0}{2} < 1$. So $x \notin Z$.

Item 1, near the spikes. Fix i and let $f \overset{\text{def.}}{=} \sigma_i \tilde\eta_\lambda$, a 1-periodic $\mathscr{C}^2$ function on $\mathbb{R}$. On the interval $(x_i - \varepsilon, x_i + \varepsilon)$ of $\mathbb{R}$, (A.1) and (A.2) give $f \geqslant \frac{1}{2} - \frac{1}{4} > 0$ and $f'' \leqslant -\kappa + \frac{\kappa}{2} < 0$. Assume that $u < v$ in this interval lie in Z. Since $f > 0$, $f(u) = f(v) = 1$. Let $m \overset{\text{def.}}{=} (u + v)/2$ and $d \overset{\text{def.}}{=} (v - u)/2 > 0$. The Taylor formula with integral remainder gives

$$f(m \pm d) = f(m) \pm f'(m)\, d + d^2 \int_0^1 (1 - \theta) f''(m \pm \theta d)\, \mathrm{d}\theta < f(m) \pm f'(m)\, d,$$

since the integrand is continuous and negative on $[0, 1)$. Adding the two relations gives $2 = f(u) + f(v) < 2f(m)$, so $|\tilde\eta_\lambda(m)| > 1$. This contradicts $\|\tilde\eta_\lambda\|_\infty \leqslant 1$. So Z has at most one point in I_i.

Item 2, form of the solutions. Let $\hat\mu$ be a solution of $\mathcal{P}_\lambda(y_0 + w)$. By Proposition 2.4, item 2, $(y_0 + w - \Phi\hat\mu)/\lambda = \hat p$, so the function $\hat\eta_\lambda$ of item 4 is $\tilde\eta_\lambda$, and item 4 gives $\operatorname{supp}(\hat\mu) \subset Z$. By item 1, Z is finite, and Remark 2.6 (i) gives $\hat\mu = \sum_{z\in Z} \hat\mu(\{z\})\delta_z$. If $Z \cap I_i = \{z\}$ and $\hat\mu(\{z\}) \neq 0$, let $\hat x_i \overset{\text{def.}}{=} z$ and $\hat a_i \overset{\text{def.}}{=} \hat\mu(\{z\})$; otherwise $\hat\mu$ has no mass in I_i, and we use the convention above. Since $Z \subset \bigcup_i I_i$, $\hat\mu = \sum_i \hat a_i \delta_{\hat x_i}$.

Item 2, signs. Let $E \overset{\text{def.}}{=} \{i\,;\, \hat a_i \neq 0\}$. Then $\hat\mu = \sum_{i\in E} \hat a_i \delta_{\hat x_i}$, and these points are pairwise distinct, since the I_i are disjoint. By Proposition 2.4, item 1, $\tilde\eta_\lambda \in \partial\|\hat\mu\|_{\mathrm{TV}}$, and the last sentence of Proposition 2.3 gives $\tilde\eta_\lambda(\hat x_i) = \operatorname{sign}(\hat a_i)$ for $i \in E$. Since $\sigma_i \tilde\eta_\lambda > 0$ on I_i, $\operatorname{sign}(\hat a_i) = \sigma_i$. Last, $\tilde\eta_\lambda$ is $\mathscr{C}^2$ (Remark 2.4), $\|\tilde\eta_\lambda\|_\infty \leqslant 1$ and $\tilde\eta_\lambda(\hat x_i) = \pm 1$, so Remark 3.3 gives $\tilde\eta'_\lambda(\hat x_i) = 0$. □

The curvature of η_0 at the spikes gives at most one point of Z in each I_i. The margin r_0 gives no point elsewhere. The first derivative of $\tilde{\eta}_\lambda$ is not controlled, and the injectivity of Γ_{x_0} is not used. The proposition does not exclude $\hat{a}_i = 0$, nor two different solutions. The amplitudes below settle both points. Smaller values of α and λ_0 keep the conclusions, since D_{α,λ_0} only shrinks.

Proposition A.2 (Rates and uniqueness)**.** *Let ε, α and λ_0 be as in Proposition A.1. After reducing λ_0, there is $c > 0$ such that, for all $(\lambda, w) \in D_{\alpha,\lambda_0}$, the solution of $\mathcal{P}_\lambda(y_0 + w)$ is unique and has exactly s spikes, $\hat{\mu} = \sum_{i=1}^s \hat{a}_i \delta_{\hat{x}_i}$ with $\hat{x}_i \in I_i$ and $\operatorname{sign}(\hat{a}_i) = \sigma_i$, and*

$$|\hat{a}_i - a_i| \leqslant c\,\lambda \quad \text{and} \quad |\hat{x}_i - x_i| \leqslant \frac{2c\,\lambda}{|a_i|}.$$

With $\lambda = \|w\|_{\mathbb{H}} / \alpha$ and $0 < \|w\|_{\mathbb{H}} \leqslant \alpha\lambda_0$, both errors are $\mathcal{O}(\|w\|_{\mathbb{H}})$.

Proof. Secondary certificates. Let $\Gamma \overset{\text{def.}}{=} \Gamma_{x_0}$ and let $(e_1, \dots, e_{2s})$ be the canonical basis of $\mathbb{R}^{2s}$. By Remark 3.4 (c), $u_i \overset{\text{def.}}{=} q_{e_i}$ and $v_i \overset{\text{def.}}{=} q_{e_{s+i}}$ satisfy $\Gamma^\star u_i = e_i$ and $\Gamma^\star v_i = e_{s+i}$. By (3.4), the functions $\xi_i \overset{\text{def.}}{=} \Phi^\star u_i$ and $\zeta_i \overset{\text{def.}}{=} \Phi^\star v_i$ satisfy, for all i, j,

$$\xi_i(x_j) = \delta_{ij}, \qquad \xi_i'(x_j) = 0, \qquad \zeta_i(x_j) = 0, \qquad \zeta_i'(x_j) = \delta_{ij}.$$

By (4.4), $\|\xi_i''\|_\infty \leqslant M_2 \|u_i\|$ and $\|\zeta_i''\|_\infty \leqslant M_2 \|v_i\|$. These are the functions of Exercise 3.4. We use the Taylor formula with Lagrange remainder: for $g \in \mathscr{C}^2(\mathbb{T})$, $x \in \mathbb{T}$ and $|t| < \frac{1}{2}$,

$$|g(x+t) - g(x) - g'(x)\,t| \leqslant \tfrac{1}{2} \|g''\|_\infty\, t^2. \tag{A.3}$$

Injectivity near x_0. The map Φ_{x_0} is injective (Remark 3.4 (d)). The function $b \mapsto \|\Phi_{x_0} b\|$ is continuous and positive on the compact set $\{b \in \mathbb{R}^s\ ;\ \|b\|_\infty = 1\}$, so its minimum θ is positive, and $\|\Phi_{x_0} b\| \geqslant \theta \|b\|_\infty$ for all $b \in \mathbb{R}^s$. Let $z = (z_1, \dots, z_s) \in \mathbb{T}^s$ and $b \in \mathbb{R}^s$. Since $\|\varphi_u - \varphi_v\| \leqslant M_1 |u - v|$ (Section 4.3),

$$\|\Phi_z b\| \geqslant \|\Phi_{x_0} b\| - \sum_i |b_i| \, \|\varphi_{z_i} - \varphi_{x_i}\| \geqslant (\theta - sM_1 \max_i |z_i - x_i|) \, \|b\|_\infty .$$

So Φ_z is injective as soon as $sM_1 \max_i |z_i - x_i| < \theta$.

Constants. Let

$$\beta \overset{\text{def.}}{=} \alpha + \|p_0\|, \qquad Q \overset{\text{def.}}{=} \max_i \max\left(\|u_i\|, \|v_i\|\right), \qquad c \overset{\text{def.}}{=} Q\Big(2\beta + \frac{M_2 \beta^2}{2\kappa}\Big). \tag{A.4}$$

These numbers do not depend on λ_0, and $c > 0$. Reduce λ_0 so that

$$c\lambda_0 < \tfrac{1}{2} \min_i |a_i| \quad \text{and} \quad 2sM_1 c\lambda_0 < \theta \min_i |a_i| .$$

Proposition A.1 still holds. Fix $(\lambda, w) \in D_{\alpha,\lambda_0}$ and a solution $\hat{\mu} = \sum_j \hat{a}_j \delta_{\hat{x}_j}$, written as in Proposition A.1. By Lemma A.1 (a), $\hat{x}_j = x_j + t_j$ with a unique $t_j \in (-\varepsilon, \varepsilon)$, and $t_j = 0$ when $\hat{a}_j = 0$.

Basic inequality. Apply Proposition 4.1, item 1, with $p = p_0$. Here $\rho = \|w\| + \lambda \|p_0\| \leqslant \beta\lambda$. So

$$\left\|\Phi(\hat{\mu} - \mu^0)\right\| \leqslant 2\beta\lambda \quad \text{and} \quad \|\hat{\mu}\|_{\mathrm{TV}} - \int \eta_0 \,\mathrm{d}\hat{\mu} \leqslant \frac{\rho^2}{2\lambda} \leqslant \frac{\beta^2 \lambda}{2}. \tag{A.5}$$

By (2.6), the left-hand side of the second bound is at least $\int (1 - |\eta_0|) \,\mathrm{d}\,|\hat{\mu}|$. By (2.7), applied to the nonzero terms of $\hat{\mu}$, and Lemma A.1 (b),

$$\int (1 - |\eta_0|) \,\mathrm{d}\,|\hat{\mu}| = \sum_{j=1}^s |\hat{a}_j| \, (1 - |\eta_0(x_j + t_j)|) \geqslant \frac{\kappa}{2} \sum_{j=1}^s |\hat{a}_j| \, t_j^2.$$

With (A.5),

$$\sum_{j=1}^{s} |\hat{a}_j|\, t_j^2 \leqslant \frac{\beta^2 \lambda}{\kappa}. \tag{A.6}$$

Amplitudes. Fix i. By (A.3), $\xi_i(x_j + t_j) = \delta_{ij} + R_{ij}$ with $|R_{ij}| \leqslant \frac{1}{2} M_2 \|u_i\| t_j^2$. By (2.7), $\int \xi_i \,\mathrm{d}\mu^0 = \sum_j a_j \xi_i(x_j) = a_i$, and

$$\int \xi_i \,\mathrm{d}(\hat{\mu} - \mu^0) = \sum_{j=1}^{s} \hat{a}_j \xi_i(x_j + t_j) - a_i = \hat{a}_i - a_i + \sum_{j=1}^{s} \hat{a}_j R_{ij}.$$

By (2.8), the left-hand side is $\langle u_i, \Phi(\hat{\mu} - \mu^0)\rangle$, of absolute value at most $2\beta\lambda \|u_i\|$ by (A.5). With (A.6),

$$|\hat{a}_i - a_i| \leqslant 2\beta\lambda \|u_i\| + \tfrac{1}{2} M_2 \|u_i\| \sum_{j=1}^{s} |\hat{a}_j|\, t_j^2 \leqslant \|u_i\| \left(2\beta + \frac{M_2 \beta^2}{2\kappa}\right) \lambda \leqslant c\,\lambda.$$

Number of spikes and signs. For all i, $|\hat{a}_i| \geqslant |a_i| - c\lambda_0 > \frac{1}{2} |a_i| > 0$. So each I_i carries one spike of $\hat{\mu}$, $\hat{\mu}$ has exactly s spikes, and $\operatorname{sign}(\hat{a}_i) = \sigma_i$ by Proposition A.1.

Positions. Fix i. By (A.3), $\zeta_i(x_j + t_j) = \delta_{ij} t_j + R'_{ij}$ with $\left|R'_{ij}\right| \leqslant \frac{1}{2} M_2 \|v_i\| t_j^2$, and $\int \zeta_i \,\mathrm{d}\mu^0 = 0$. So

$$\langle v_i, \Phi(\hat{\mu} - \mu^0)\rangle = \int \zeta_i \,\mathrm{d}(\hat{\mu} - \mu^0) = \hat{a}_i t_i + \sum_{j=1}^{s} \hat{a}_j R'_{ij},$$

and the computation of the amplitudes, with v_i in place of u_i, gives $|\hat{a}_i|\, |t_i| \leqslant c\lambda$. Since $|\hat{a}_i| > \frac{1}{2} |a_i|$, $|\hat{x}_i - x_i| = |t_i| < 2c\lambda / |a_i|$.

Uniqueness. By item 2 of Proposition A.1 and the step above, every solution has exactly one spike in each I_i, at a point of Z. By item 1 of the same proposition, Z has at most one point in I_i. So $Z = \{\hat{x}_1, \dots, \hat{x}_s\}$, and all solutions have the same positions $\hat{x} \overset{\text{def.}}{=} (\hat{x}_1, \dots, \hat{x}_s)$. By the bound on the positions, $sM_1 \max_i |\hat{x}_i - x_i| \leqslant 2sM_1 c\lambda_0 / \min_i |a_i| < \theta$, so $\Phi_{\hat{x}}$ is injective. Write two solutions as $\sum_i \hat{a}_i \delta_{\hat{x}_i}$ and $\sum_i \check{a}_i \delta_{\hat{x}_i}$. By Proposition 2.4, item 2, they have the same image by Φ, that is $\Phi_{\hat{x}}(\hat{a} - \check{a}) = 0$. Hence $\hat{a} = \check{a}$. This is the argument of Remark 2.8 (d).

Rates in the noise. If $0 < \|w\| \leqslant \alpha\lambda_0$ and $\lambda = \|w\| / \alpha$, then $0 < \lambda \leqslant \lambda_0$ and $\|w\| = \alpha\lambda$, so $(\lambda, w) \in D_{\alpha,\lambda_0}$, and $c\lambda = (c/\alpha) \|w\|$. □

The proof never solves the optimality conditions. It uses them to locate the support, in Proposition A.1, and, for uniqueness, through the fact that all solutions have the same image by Φ. The amplitudes and the positions come from the basic inequality of Proposition 4.1, tested against η_0, ξ_i and ζ_i. The curvature of η_0 enters twice: in the structure, and in the quadratic bound (A.6), which makes the Taylor remainders of order λ.

Remark A.1 (Comparison with Theorem 4.2). (a) Hypotheses. They are those of Theorem 4.2 with $k = 2$: $\varphi \in \mathscr{C}^2(\mathbb{T})$, Γ_{x_0} injective and (NDSC). No $\mathscr{C}^3$ assumption and no rate for the dual bias are used. As in Theorem 4.2, $\lambda > 0$ on D_{α,λ_0} (Remark 4.7 (d)).

(b) What is proved. All the conclusions of Theorem 4.2 except the regularity of the path: on D_{α,λ_0} the solution is unique, it has exactly s spikes, one in each I_i, with $\operatorname{sign}(\hat{a}_i) = \operatorname{sign}(a_i)$, and the errors are $\mathcal{O}(\|w\|_{\mathbb{H}})$ when $\lambda = \|w\|_{\mathbb{H}} / \alpha$. The spikes are labelled by the intervals I_i.

(c) What is more precise. The bounds $|\hat{a}_i - a_i| \leqslant c\lambda$ and $|\hat{x}_i - x_i| \leqslant 2c\lambda / |a_i|$ hold on all of D_{α,λ_0}. The order is the same as in Theorem 4.2: its proof gives $L(\lambda + \|w\|_{\mathbb{H}}) \leqslant L(1+\alpha)\lambda$ on D_{α,λ_0}, and Exercise 4.3 gives a case with $|\hat{a}_\lambda - a| = (1+\alpha)\lambda$. The constant c in (A.4) is explicit, and so is $\alpha = \min(r_0, \frac{1}{2}, \kappa) / (4 \max(M_0, M_2))$, of the same type as the value of α admitted in Lecture 4 under $\varphi \in \mathscr{C}^3(\mathbb{T})$. With $G \overset{\text{def.}}{=} \Gamma^\star_{x_0} \Gamma_{x_0}$ and $q_e = \Gamma_{x_0} G^{-1} e$, $\|q_e\|^2 = \langle G^{-1} e, GG^{-1} e\rangle = \langle G^{-1} e, e\rangle$. So $\|u_i\|^2$ and $\|v_i\|^2$ are diagonal entries of G^{-1}, and $Q^2 \leqslant \|G^{-1}\|_{\mathrm{op}}$, the quantity that enters the constants of the rates admitted in Lecture 4. The factor $1/|a_i|$ of the positions is also there. No

neighbourhoods V and W appear, so the order of the choices of Remark 4.7 (c) disappears: ε first, then α and λ_0, and λ_0 is reduced once at the end.

(d) What is weaker. The map $(\lambda, w) \mapsto (\hat{a}, \hat{x})$ is not shown to be $\mathscr{C}^{k-1}$ for $\varphi \in \mathscr{C}^k(\mathbb{T})$, nor even continuous, and it is not extended to a neighbourhood of $(0, 0)$. This regularity gives the rate of the dual bias admitted in Remark 4.6 (b). The number λ_0 is not explicit: it depends on $b(\lambda_0)$, for which these notes prove no rate. Under $\varphi \in \mathscr{C}^3(\mathbb{T})$, with this rate for the dual bias, λ_0 can be made explicit; we admit this.

(e) Where the assumptions enter. The curvature $\sigma_i \eta_0''(x_i) < 0$ gives at most one spike in each I_i and the quadratic bound (A.6). The margin off the spikes gives no spike elsewhere. The injectivity of Γ_{x_0} gives the secondary certificates ξ_i and ζ_i, and the injectivity of $\Phi_{\hat{x}}$ used for uniqueness.

Bibliography

[1] Gaspard Riche de Prony. "Essai expérimental et analytique sur les lois de la dilatabilité des fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alkool, à différentes températures". In: *Journal de l'École Polytechnique* 1.2 (1795), pp. 24–76.

[2] Ankur Moitra. "Super-resolution, Extremal Functions and the Condition Number of Vandermonde Matrices". In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing (STOC)*. 2015, pp. 821–830. DOI: 10.1145/2746539.2746561.

[3] Scott Shaobing Chen, David L. Donoho, and Michael A. Saunders. "Atomic decomposition by basis pursuit". In: *SIAM Review* 43.1 (2001), pp. 129–159. DOI: 10.1137/S003614450037906X.

[4] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288. DOI: 10.1111/j.2517-6161.1996.tb02080.x.

[5] Emmanuel J. Candès and Terence Tao. "Decoding by linear programming". In: *IEEE Transactions on Information Theory* 51.12 (2005), pp. 4203–4215. DOI: 10.1109/TIT.2005.858979.

[6] Emmanuel J. Candès, Justin K. Romberg, and Terence Tao. "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information". In: *IEEE Transactions on Information Theory* 52.2 (2006), pp. 489–509. DOI: 10.1109/TIT.2005.862083.

[7] David L. Donoho. "Compressed Sensing". In: *IEEE Transactions on Information Theory* 52.4 (2006), pp. 1289–1306. DOI: 10.1109/TIT.2006.871582.

[8] Vincent Duval and Gabriel Peyré. "Sparse regularization on thin grids I: the Lasso". In: *Inverse Problems* 33.5 (2017), p. 055008. DOI: 10.1088/1361-6420/aa5e12.

[9] Vincent Duval and Gabriel Peyré. "Exact Support Recovery for Sparse Spikes Deconvolution". In: *Foundations of Computational Mathematics* 15.5 (2015), pp. 1315–1355. DOI: 10.1007/s10208-014-9228-6.

[10] Nicolas Keriven et al. "Sketching for large-scale learning of mixture models". In: *Information and Inference* 7.3 (2018), pp. 447–508. DOI: 10.1093/imaiai/iax015.

[11] Walter Rudin. *Real and Complex Analysis*. 3rd. McGraw-Hill, 1987.

[12] Walter Rudin. *Functional Analysis*. 2nd ed. McGraw-Hill, 1991.

[13] Tuomas Hytönen et al. *Analysis in Banach Spaces. Volume I: Martingales and Littlewood–Paley Theory*. Vol. 63. Ergebnisse der Mathematik und ihrer Grenzgebiete. 3. Folge. Springer, 2016.

[14] Heinz H. Bauschke and Patrick L. Combettes. *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. 2nd ed. Springer, 2017.

[15] Ivar Ekeland and Roger Temam. *Convex Analysis and Variational Problems*. Vol. 1. Studies in Mathematics and Its Applications. Reprinted by SIAM, 1999. North-Holland, 1976.

[16] Clarice Poon, Nicolas Keriven, and Gabriel Peyré. “The Geometry of Off-the-Grid Compressed Sensing”. In: *Foundations of Computational Mathematics* 23.1 (2023). arXiv:1802.08464, pp. 241–327. DOI: 10.1007/s10208-021-09545-5.

[17] Yohann De Castro, Rémi Gribonval, and Nicolas Jouvin. “Effective regions and kernels in continuous sparse regularisation, with application to sketched mixtures”. In: *Information and Inference: A Journal of the IMA* (2026). Forthcoming (August 2026); arXiv:2507.08444. Pinpoints follow the journal version.

[18] Emmanuel J. Candès and Carlos Fernandez-Granda. “Towards a Mathematical Theory of Super-Resolution”. In: *Communications on Pure and Applied Mathematics* 67.6 (2014), pp. 906–956. DOI: 10.1002/cpa.21455.

[19] Iosif Pinelis. “Optimum bounds for the distributions of martingales in Banach spaces”. In: *The Annals of Probability* 22.4 (1994), pp. 1679–1706. DOI: 10.1214/aop/1176988477.

[20] Mathieu Blondel, André F. T. Martins, and Vlad Niculae. “Learning with Fenchel-Young losses”. In: *Journal of Machine Learning Research* 21.35 (2020), pp. 1–69.

[21] Pierre-Cyril Aubin-Frankowski and Yohann De Castro. “Fenchel–Young Duality Gaps: Certified Early Stopping for Regularized Inverse Problems”. Preprint. 2026.

[22] Pierre-Cyril Aubin-Frankowski and Yohann De Castro. “Fenchel–Young Duality Gaps: Localization and Exact Support Recovery for the Generalized Beurling–Lasso”. Preprint. 2026.

[23] Quentin Denoyelle et al. “The Sliding Frank–Wolfe Algorithm and Its Application to Super-resolution Microscopy”. In: *Inverse Problems* 36.1 (2020), p. 014001. DOI: 10.1088/1361-6420/ab2a29.

[24] Lénaïc Chizat. “Sparse Optimization on Measures with Over-parameterized Gradient Descent”. In: *Mathematical Programming* 194.1–2 (2022), pp. 487–532. DOI: 10.1007/s10107-021-01636-z.

[25] Nicolás García Trillos and Dejan Slepčev. “On the Rate of Convergence of Empirical Measures in ∞-Transportation Distance”. In: *Canadian Journal of Mathematics* 67.6 (2015), pp. 1358–1383. DOI: 10.4153/CJM-2014-044-6.

[26] Yohann De Castro, Sébastien Gadat, and Clément Marteau. “Fast Spawn&Prune (FS&P): Global convergence of stochastic conic particle gradient descent via birth/death process”. Preprint, arXiv:2605.19784. 2026.