IOP Publishing Inverse Problems

Inverse Problems **36** (2020) 014001 (42pp) https://doi.org/10.1088/1361-6420/ab2a29

The sliding Frank–Wolfe algorithm and its application to super-resolution microscopy

Quentin Denoyelle[1], Vincent Duval[2], Gabriel Peyré[3] and Emmanuel Soubies[4]

[1] CEREMADE, Univ. Paris-Dauphine, Paris, France
[2] INRIA Paris, MOKAPLAN, Paris, France
[3] CNRS & ENS, Paris, France
[4] Biomedical Imaging Group, EPFL

E-mail: denoyelle@ceremade.dauphine.fr, vincent.duval@inria.fr, gabriel.peyre@ens.fr and emmanuel.soubies@epfl.ch

Received 13 November 2018, revised 29 April 2019
Accepted for publication 17 June 2019
Published 3 December 2019

Abstract

This paper showcases the theoretical and numerical performance of the Sliding Frank–Wolfe, which is a novel optimization algorithm to solve the BLASSO sparse spikes super-resolution problem. The BLASSO is a continuous (i.e. off-the-grid or grid-less) counterpart to the well-known ℓ^1 sparse regularisation method (also known as LASSO or basis pursuit). Our algorithm is a variation on the classical Frank–Wolfe (also known as conditional gradient) which follows a recent trend of interleaving convex optimization updates (corresponding to adding new spikes) with non-convex optimization steps (corresponding to moving the spikes). Our main theoretical result is that this algorithm terminates in a finite number of steps under a mild non-degeneracy hypothesis. We then target applications of this method to several instances of single molecule fluorescence imaging modalities, among which certain approaches rely heavily on the inversion of a Laplace transform. Our second theoretical contribution is the proof of the exact support recovery property of the BLASSO to invert the 1D Laplace transform in the case of positive spikes. On the numerical side, we conclude this paper with an extensive study of the practical performance of the Sliding Frank–Wolfe on different instantiations of single molecule fluorescence imaging, including convolutive and non-convolutive (Laplace-like) operators. This shows the versatility and superiority of this method with respect to alternative sparse recovery technics.

Keywords: super-resolution, convex optimization, Frank–Wolfe, microscopy, sparsity, BLASSO, Laplace transform

(Some figures may appear in colour only in the online journal)

1. Introduction

1.1. Super-resolution using the BLASSO

Super-resolution consists of retrieving the fine scale details of a possibly noisy signal from coarse scale information. The importance of recovering the high frequencies of a signal comes from the fact that there is often a physical blur in the acquisition process, such as diffraction in optical systems, wave reflection in seismic imaging or spikes recording from neuronal activity.

In resolution theory [27], the two-point resolution criterion defines the ability of a system to resolve two points of equal intensities. It is defined as a distance, namely the Rayleigh criterion, which only depends on the system. In the case of the ideal low-pass filter (i.e. convolution with the Dirichlet kernel) with cutoff frequency f_c, the Rayleigh criterion is $1/f_c$. Then, super-resolution in signal processing consists in developing techniques which enable to retrieve information below the Rayleigh criterion.

Let us introduce in a more formal way the problem which will be the framework of this article. Let X be a connected subset of $\mathbb{R}^d$ with non-empty interior or the d-dimensional torus $\mathbb{T}^d$ ($d \in \mathbb{N}^*$) and $\mathcal{M}(X)$ the Banach space of bounded Radon measures on X. The latter can be seen as the topological dual of $\mathscr{C}_0(X, \mathbb{R})$, the space of continuous functions on X that vanish at infinity. We consider a given integral operator $\Phi : \mathcal{M}(X) \to \mathcal{H}$, where $\mathcal{H}$ is a separable Hilbert space, whose kernel φ is supposed to be a smooth function (see definition 4 for the technical assumptions made on φ), i.e.

$$\forall m \in \mathcal{M}(X), \quad \Phi m \overset{\text{def.}}{=} \int_X \varphi(x) \mathrm{d}m(x). \tag{1}$$

The operator Φ models the acquisition process. It includes translation-invariant operators such as convolutions (i.e. $\varphi(x) = \tilde{\varphi}(\cdot - x)$) as well as non-translation invariant operators such as the Laplace transform ($X = \mathbb{R}_+^*$ and $\varphi(x) = (t \mapsto \mathrm{e}^{-tx}) \in L^2(\mathbb{R}_+)$) considered in the present paper.

The sparse spikes super-resolution problem aims at recovering an approximation of an unknown input discrete measure $m_{a_0,x_0} \overset{\text{def.}}{=} \sum_{i=1}^N a_{0,i}\delta_{x_{0,i}}$ from noisy measurements $y \overset{\text{def.}}{=} y_0 + w$ where $y_0 \overset{\text{def.}}{=} \Phi m_{a_0,x_0}$ and $w \in \mathcal{H}$ models the acquisition noise. Here $a_{0,i} \in \mathbb{R}$ are the amplitudes of the Dirac masses at positions $x_{0,i} \in X$. This is an ill-posed inverse problem and the BLASSO is a way to solve it in a stable way by introducing a sparsity-enforcing convex regularization.

1.1.1. From the LASSO to the BLASSO. The common practice in sparse spike recovery relies on ℓ^1 regularization which is known as LASSO in statistic [83] or basis pursuit in the signal processing community [17]. Given a grid of possible positions, the reconstruction problem is addressed as the minimization of a quadratic error subject to an ℓ^1 penalization. The ℓ^1 prior provides solutions with few nonzero coefficients and can be computed efficiently with convex optimization methods. Moreover, recovery guarantees have been proved under certain assumptions [29].

Following recent works (see for instance [6, 12, 15, 22, 33]), we consider instead sparse spike estimation methods which operate over a continuous domain, i.e. without resorting to some sort of discretization on a grid. The inverse problem is solved over the space of Radon measures which is a non-reflexive Banach space. This continuous 'grid-free' setting makes the mathematical analysis easier and allows us to make precise statement about the location of the recovered spikes.

The technique that we are considering in this paper consists in solving a convex optimization problem that uses the total variation norm, which is the counterpart of the ℓ^1-norm for measures. It favors the emergence of spikes in the solution and is defined by

$$\forall m \in \mathcal{M}(X), \quad |m|(X) \overset{\text{def.}}{=} \sup_{\psi \in \mathscr{C}_0} \left\{ \int_X \psi \mathrm{d}m ; \|\psi\|_{\infty,X} \leqslant 1 \right\}. \tag{2}$$

In particular, for $m_{a_0,x_0} \overset{\text{def.}}{=} \sum_{i=1}^N a_{0,i}\delta_{x_{0,i}}$,

$$|m_{a_0,x_0}|(X) = \|a_0\|_1,$$

which shows in a way that the total variation norm generalizes the ℓ^1-norm to the continuous setting of measures (i.e. no discretization grid is required).

When no noise is contaminating the data, one considers the classical basis pursuit, defined originally in [17] in a finite dimensional setting, written here over the space of Radon measures

$$\min_{m \in \mathcal{M}(X)} |m|(X) \quad \text{s.t.} \Phi m = y_0 \tag{$\mathcal{P}_0(y_0)$}.$$

This problem is studied in [15], in the case where Φ is an ideal low-pass filter on the torus $X = \mathbb{T}$.

When the signal is noisy, i.e. when one observes $y = y_0 + w$, with $w \in \mathcal{H}$, we may rather consider the problem

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

Here $\lambda > 0$ is a parameter that should be adapted to the noise level $\|w\|_{\mathcal{H}}$. This problem is coined as BLASSO [22].

1.1.2. BLASSO performance analysis. In order to quantify the recovery performance of the methods $\mathcal{P}_0(y_0)$ and $\mathcal{P}_\lambda(y)$, the following two questions arise:

(i) Does the solutions of $\mathcal{P}_0(y_0)$ recover the input measure m_{a_0,x_0}?
(ii) How close is the solution of $\mathcal{P}_\lambda(y)$ to the solution of $\mathcal{P}_0(y_0)$?

When the amplitudes of the spikes are arbitrary complex numbers, the answers to the above questions require a large enough minimum separation distance $\Delta(m_{a_0,x_0})$ between the spikes, where

$$\Delta(m_{a_0,x_0}) \overset{\text{def.}}{=} \min_{i \neq j} d_X(x_{0,i}, x_{0,j}). \tag{3}$$

When $X = \mathbb{T}$, d_X is the geodesic distance on the circle

$$\forall x, y \in \mathbb{R}, \quad d_X(x + \mathbb{Z}, y + \mathbb{Z}) = \min_{k \in \mathbb{Z}} |x - y + k|. \tag{4}$$

In [15], the authors show that for the ideal low-pass filter, m_{a_0,x_0} is the unique solution of $\mathcal{P}_0(y_0)$ provided that $\Delta(m_{a_0,x_0}) \geqslant \frac{C}{f_c}$ where $C > 0$ is a universal constant and f_c the cutoff frequency of the ideal low-pass filter. In the same paper, it is shown that $C \leqslant 2$ when $a_0 \in \mathbb{C}^N$ and $C \leqslant 1.87$ when $a_0 \in \mathbb{R}^N$. In [41], the constant C is further refined to $C \leqslant 1.26$ when $a_0 \in \mathbb{R}^N$. Suboptimal lower bounds on C were given in [33, 81]. Moreover, it was recently shown in [20] that necessarily $C \geqslant 1$ in the sense that for all $\varepsilon > 0$, and for f_c large enough, there exist measures with $\Delta(m_{a_0,x_0}) \geqslant (1 - \varepsilon)/f_c$ which are not identifiable using $(\mathcal{P}_0(y_0))$.

The second question receives partial answers in [3, 12, 14, 40]. In [12], it is shown that if the solution of $\mathcal{P}_0(y_0)$ is unique then the measures recovered by $\mathcal{P}_\lambda(y)$ converge in the weak-$*$

sense to the solution of $\mathcal{P}_0(y_0)$ when $\lambda \to 0$ and $\|w\|_{\mathcal{H}} / \lambda \to 0$. In [14], the authors measure the reconstruction error using the L^2-norm of an ideal low-pass filtered version of the recovered measures. In [3, 40], error bounds are given on the locations of the recovered spikes with respect to those of the input measure m_{a_0,x_0}. However, those works provide little information about the geometrical structure of the measures recovered by $\mathcal{P}_\lambda(y)$. That point is addressed in [33] where the authors show that under the *non degenerate source condition*, there exists a unique solution to $\mathcal{P}_\lambda(y)$ with the exact same number of spikes as the original measure provided that λ and $\|w\|_{\mathcal{H}} / \lambda$ are small enough. Moreover in that regime, this solution converges to the original measure when the noise drops to zero.

BLASSO for positive spikes. For positive spikes (i.e. $a_{0,i} > 0$), the picture is radically different. Exact recovery of m_{a_0,x_0} without noise (i.e. $(w, \lambda) = (0, 0)$) holds whatever the distance between the spikes [22], but stability constants explode as $\Delta(m_{a_0,x_0}) \to 0$. However, the authors in [68] show that stable recovery is obtained if the signal-to-noise ratio grows faster than $O(1/\Delta^{2N})$. This closely matches the optimal lower bounds of $O(1/\Delta^{2N-1})$ obtained by combinatorial methods [25].

Finally, provided a certain nondegeneracy condition, it was recently shown in [28] that support recovery is guaranteed in the presence of noise if the signal-to-noise ratio grows faster than $O(1/\Delta^{2N-1})$.

1.2. Solving the BLASSO

As the BLASSO is an optimization problem over the infinite dimensional space of Radon measures $\mathcal{M}(X)$, its resolution is challenging. We review in this section the existing approaches to tackle this problem. They can be roughly divided into three main families although there exists a flurry of generalizations and extensions that must be considered separately.

1.2.1. Fixed spatial discretization. A common approach consists in constraining the measure to be supported on a grid. This leads to a finite dimensional convex optimization problem—known as LASSO [83] or basis pursuit [17]—for which there exist numerous solvers. These include the block-coordinate descent (BCD) algorithm [86, 87], the homotopy/LARS algorithm [36, 80], or proximal forward-backward splitting algorithms [18] such as the iterative soft thresholding (IST) [21]. Although simples to implement, the latters are in general slow to converge (the error in the objective function is typically of the order of $O(1/k)$, where k is the number of iterations) [21, 30, 42]. However, there exist accelerated versions such as FISTA [4], which benefit from a better non-asymptotic rate of convergence ($O(1/k^2)$). Finally, it is noteworthy that these proximal methods enjoy a linear asymptotic rate (see for instance [64]), but this regime takes time to reach.

The main limitation of these grid-based methods is that, in order to go below the Rayleigh limit and perform super-resolution, the grid must be thin enough. This leads to theoretical and practical issues. Indeed, refining the grid not only increases the computational cost of each iteration, but it also deteriorates the conditioning of the linear operator to invert. Hence, in practice, these methods provide solutions which are composed of small clusters of non-zero coefficents around each 'true' spike. A way to mitigate this issue is to perform a post processing by replacing each cluster of spikes by its center of mass, as proposed in [43, 82]. This drastically reduces the number of false positive spikes although it is hard to analyze theoretically and can be unstable. Instead, one can also consider methods based on safe rules [39] which perform a progressive pruning of the grid and keep only active sets of weights [67]. Finally, it

has been shown in [34, 35] that the solution of the LASSO, in a small noise regime and when the step size tends to zero, contains pairs of spikes around the true ones.

1.2.2. Fixed spectral discretization and semidefinite programming (SDP) formulation. In [15], the authors propose a reformulation of the basis Pursuit for measures into an equivalent finite dimensional SDP for which solvers exist. Similarly, one can get an SDP formulation of the BLASSO. However, these equivalences are only true in a 1D setting. In higher dimensions ($d \geqslant 2$), one needs to use the so-called Lasserre's hierarchy [61, 62]. This principle has been used for the super-resolution problem in [23].

The resolution of SDPs can be tackled through proximal splitting methods [84] as well as interior point methods [11]. However, the overall complexity of the latter is polynomial in $O(f_c^{2d})$, where d is the dimension of the domain X, which restricts its application to small dimensional problems. This limitation has led to recent developments [16] where the authors proposed a relaxed low rank SDP formulation of the BLASSO in order to use a Frank–Wolfe-type method (see below). The resulting method enjoys the better overall complexity of $O(f_c^d \log(f_c))$ per iteration.

Finally, note that these SDP-based approaches are restricted to certain type of forward operators (typically Fourier measurements). In contrast, grid-based proximal methods as well as Frank–Wolfe (directly on the BLASSO, see below) can be used for a larger class of operators Φ.

1.2.3. Optimization over the space of measures. In order to directly solve the BLASSO, one needs to design algorithms that do not use any Hilbertian structure and can instead deal with measures. The benefit is the fact that one can exploit advantageously the continuous setting of the problem (typically moving continuously spikes over the domain). In contrast to fixed spatial or spectral discretization methods, these algorithms proceed by iteratively adding new spikes, i.e. Dirac masses, to the recovered measure.

The Frank–Wolfe (FW) algorithm [44] (see section 4), also called the conditional gradient method (CGM) [63], solves optimization problems of the form $\min_{m \in C} f(m)$, where C is a weakly compact convex set of a topological vector space and f is a differentiable convex function (in the case of the BLASSO, m is a Radon measure). It proceeds by iteratively minimizing a linearized version of f. No Hilbertian structure is used, which makes it well suited to work on the space of Radon measures. It has been proven under a curvature condition on f (which holds on a Banach space for smooth functions having a Lipschitz gradient) that the rate of convergence of this algorithm in the objective function is $O(1/k)$ (see for instance [26]). However, it is possible to improve the convergence speed of FW by replacing the current iterate by any 'better' candidate $m \in C$ that further decreases the objective function f. This simple idea has led to several successful variations of the standard FW algorithm. For instance, the authors of [12] proposed a modified Frank–Wolfe algorithm for the BLASSO where the final step updates the amplitudes and positions of spikes by a gradient descent on a non-convex optimization problem. Moving the spikes positions takes advantage of the continuous framework of the problem (the domain X is not discretized) which is the main ingredient that leads to a typical N-step convergence observed empirically. Finally, this approach has later been used in [10] and provides state of the art results in many sparse inverse problems such as matrix completion or single molecule localization microscopy (SMLM) [75, 76].

Let us mention that, as observed in [37], the Frank–Wolfe algorithm for the resolution of the (constrained variant of the) BLASSO is equivalent to the exchange method (introduced by Remez in the 1930s, see [71]) applied to its dual problem. Since the main focus of the present

paper is on solving the primal problem ($\mathcal{P}_\lambda(y)$), we refer to this method as the Frank–Wolfe algorithm in the rest of the paper.

1.3. Other methods for super-resolution

The Prony method [24] and its successors such as MUSIC (MUltiple SIgnal classification) [77], ESPRIT (estimation of signal parameters by rotational invariance techniques) [59], or matrix pencil [51], are spectral methods which perform spikes localization from low frequency measurements. They do not need any discretization and enable to recover exactly the initial signal in the noiseless case as long as there are enough observations compared to the number of distinct frequencies [65]. Extensions to deal with noise have been developped in [13, 19] and stability is known under a minimum separation distance [65]. Greedy algorithms constitute another class of popular methods for sparse super-resolution. The matching pursuit (MP) [66] adds new spikes by finding the ones that best correlate with the residual. The orthogonal matching pursuit (OMP) [49, 79, 85] is similar to MP but imposes that the current estimate of the observations, i.e. $\Phi(\sum_{i=1}^{k} a_i \delta_{x_i})$, is always orthogonal to the residual. Hence, the amplitudes of the Dirac masses are updated by an orthogonal projection after every support update (i.e. addition of a new spike). It is noteworthy that there exist many generalizations/variants of OMP. For instance, the results of OMP can be improved with a backtracking step at each iteration, allowing to remove non reliable spikes from the support of the reconstructed measure [53].

These greedy pursuit algorithms can be applied without grid discretization [56] which enables the use of local optimizations over the spikes' positions [38]. Finally, let us mention the class of nonconvex optimization methods which include the well known iterative hard thresholding (IHT) [7, 8]

1.4. Contributions

Our first set of contributions, detailed in section 3, studies the BLASSO performance in the special case of several types of Laplace transforms. This theoretical study is motivated by the use of these Laplace transform for certain types of fluorescence microscopy imaging devices. Our main finding is that for positive spikes, these operators can be stably inverted without minimum separation distance. This study makes use of the theoretical tools developed in our previous work [28].

Our algorithmic contributions are detailed in section 4, where we introduce the Sliding Frank–Wolfe, which is an extension of the initial FW solver proposed in [12]. Proposition 5 shows that this algorithm, used to solve the BLASSO, enjoys the same convergence property as the classical Frank–Wolfe algorithm (weak-$*$ convergence with a rate in the objective function of $O(1/k)$). Our main theoretical contribution is theorem 3 which proves that our algorithm converges towards the unique solution of the BLASSO in a finite number of iterations.

Section 5 makes the connection between these two sets of contributions, by showcasing the SFW algorithm for 3D PALM/STORM super-resolution fluorescence microscopy. We study its performance for several imaging operators, among which some relies on the inversion of a Laplace transform along the depth axis.

The code to reproduce the numerical illustrations of this article can be found online at https://github.com/qdenoyelle.

1.5. Notations and definitions

This section gathers some useful notations and definitions.

1.5.1. Ground space and measures. We frame our theoretical and numerical analysis of the BLASSO on the space of Radon measure over a set X.

Definition 1 (Set X of positions of spikes). The set of positions of spikes, denoted X, is supposed to be a subset of $\mathbb{R}^d$ with non-empty interior $\overset{\circ}{X}$, or $\mathbb{T}^d$ with $d \in \mathbb{N}^*$.

Definition 1 covers the particular case of $X = \mathbb{R}^d$, $X = \mathbb{T}^d$ or any compact subset with non-empty interior of $\mathbb{R}^d$.

Definition 2 (Continuous functions on X). Let $(Y, \|\cdot\|_Y)$ be a normed space. We denote by $\mathscr{C}_c(X, Y)$ the space of Y-valued continuous functions with compact support, by $\mathscr{C}_0(X, Y)$ the set of continuous functions that vanish at infinity i.e.

$$\forall \varepsilon > 0, \exists K \subset X \text{ compact}, \quad \sup_{x \in X \setminus K} \|\varphi(x)\|_Y \leqslant \varepsilon,$$

and by $\mathscr{C}^k(X, Y)$ the set of k-times differentiable functions on X. Note that when X is compact, $\mathscr{C}_c(X, Y)$ and $\mathscr{C}_0(X, Y)$ are simply the set $\mathscr{C}(X, Y)$ of continuous functions on X.

Now we can define rigorously the space of real bounded Radon measures on X.

Definition 3 (Set $\mathcal{M}(X)$ of radon measures). We denote by $\mathcal{M}(X)$ the set of real bounded Radon measures on X which is the topological dual of $\mathscr{C}_0(X, \mathbb{R})$ endowed with $\|\cdot\|_{\infty,X}$ (the supremum norm for functions defined on X).

By the Riesz representation theorem, $\mathcal{M}(X)$ is also the set of regular real Borel measures with finite total mass on X. See [73] for more details on Radon measures.

1.5.2. Kernels. This paragraph details the assumptions that we use in the following on the kernel φ. We recall that the operator $\Phi : \mathcal{M}(X) \to \mathcal{H}$, which models the acquisition process of the source signal, has the form:

$$\forall m \in \mathcal{M}(X), \quad \Phi m \overset{\text{def.}}{=} \int_X \varphi(x) \mathrm{d}m(x). \tag{5}$$

The above quantity is well-defined (as a Bochner integral) as soon as φ is continuous and bounded. In order to apply some results of [28], we add the hypotheses that are summarized below.

Definition 4 (Admissible kernels φ). We denote by $\mathrm{KER}^{(k)}$, the set of admissible kernels of order k. A function $\varphi : X \to \mathcal{H}$ belongs to $\mathrm{KER}^{(k)}$ if:

- $\varphi \in \mathscr{C}^k(X, \mathcal{H})$,
- For all $p \in \mathcal{H}$, $x \in X \mapsto \langle \varphi(x), p \rangle_{\mathcal{H}}$ vanishes at infinity,
- for all $0 \leqslant i \leqslant k$, $\sup_{x \in X} \left\| D^i \varphi(x) \right\|_{\mathcal{H}} < +\infty$.

where $D^i\varphi$ is the ith differential of φ.

1.5.3. Operators. Given $x = (x_1, \ldots, x_N) \in \overset{\circ}{X}^N$, we denote by $\Phi_x : \mathbb{R}^N \to \mathcal{H}$ the linear operator such that:

$$\forall a \in \mathbb{R}^N, \quad \Phi_x(a) \overset{\text{def.}}{=} \sum_{i=1}^{N} a_i \varphi(x_i), \tag{6}$$

and by $\Gamma_x : (\mathbb{R}^N \times \underbrace{\mathbb{R}^N \times \cdots \times \mathbb{R}^N}_{d}) \to \mathcal{H}$ the linear operator defined by

$$\forall (a, b_1, \ldots, b_d) \in \mathbb{R}^N \times (\mathbb{R}^N)^d, \quad \Gamma_x \begin{pmatrix} a \\ b_1 \\ \vdots \\ b_d \end{pmatrix} \overset{\text{def.}}{=} \sum_{i=1}^{N} \left(a_i \varphi(x_i) + \sum_{j=1}^{d} b_{j,i} \partial_j \varphi(x_i) \right). \tag{7}$$

We may also write $\Gamma_x = \begin{pmatrix} \Phi_x & (\Phi_x)^{(1)} \end{pmatrix}$, where $(\Phi_x)^{(1)}$ (sometimes denoted by $\Phi_x{}'$) stacks all the first order derivatives of φ for the different positions x_i. Similarly we define $(\Phi_x)^{(k)}$ for $k \geqslant 1$ by stacking all the derivatives of order k. Finally, Γ_x^+ refers to the pseudo-inverse of Γ_x.

When $d = 1$, given $x_c \in \overset{\circ}{X}$, we denote by $\varphi_k \in \mathcal{H}$ the kth derivative of φ at x_c, i.e.

$$\varphi_k \overset{\text{def.}}{=} \varphi^{(k)}(x_c). \tag{8}$$

In particular, $\varphi_0 = \varphi(x_c)$. Given $k \in \mathbb{N}$, we then define

$$\Psi_k \overset{\text{def.}}{=} \begin{pmatrix} \varphi_0 & \varphi_1 & \ldots & \varphi_k \end{pmatrix}. \tag{9}$$

1.5.4. Injectivity assumption. In order to avoid degeneracy issues we sometimes assume the following injectivity assumption of the operator when restricted to discrete spikes.

Definition 5. Let $\varphi : X \to \mathcal{H}$. For all $k \in \mathbb{N}$, we say that the hypothesis $\mathcal{I}_k$ holds at $x_c \in \overset{\circ}{X}$ if and only if

$\varphi \in \mathrm{KER}^{(k)}$ and $(\varphi_0, \ldots, \varphi_k)$ are linearly independent in $\mathcal{H}$. $(\mathcal{I}_k)$

1.5.5. Norms. We use the ℓ^∞-norm, $|\cdot|_\infty$, for vectors of $\mathbb{R}^N$ or $\mathbb{R}^{2N}$, whereas the notation $\|\cdot\|$ refers to an operator norm (on matrices, or bounded linear operators). $\|\cdot\|_{\mathcal{H}}$ is the norm on $\mathcal{H}$ associated to the inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. $\|\cdot\|_{\infty,X}$ denotes the L^∞-norm for functions defined on X.

2. Reminders on the BLASSO

2.1. Recovery of the support in presence of noise.

Let $x_0 \in \overset{\circ}{X}{}^N$, $a_0 \in (\mathbb{R} \setminus \{0\})^N$ and $m_{a_0,x_0} = \sum_{i=1}^N a_{0,i} \delta_{x_{0,i}}$. The BLASSO is the variational problem

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

where $y \overset{\text{def.}}{=} \Phi m_{a_0,x_0} + w$ are the noisy observations of a measure composed of a sum of Dirac masses. The optimality of a measure m_λ for $\mathcal{P}_\lambda(y)$ is characterized by the fact that the function

$$\eta_\lambda \overset{\text{def.}}{=} \Phi^* p_\lambda \quad \text{where} \quad p_\lambda \overset{\text{def.}}{=} \frac{1}{\lambda}(y - \Phi m_\lambda) \tag{10}$$

satisfies $\|\eta_\lambda\|_{\infty,X} \leqslant 1$. The function η_λ is then called a dual certificate.

When one is interested in the recovery of the support, i.e. finding a solution $m_{a,x}$ of $\mathcal{P}_\lambda(y)$ composed of exactly the same number of Dirac masses as the initial measure m_{a_0,x_0}, in a small noise regime, an important object is the so-called vanishing derivatives precertificate introduced in [33].

Definition 6 (Vanishing derivatives precertificate [33]). If Γ_{x_0} has full column rank, there is a unique solution to the problem

$$\inf\left\{\|p\|_{\mathcal{H}}\,;\forall i=1,\dots,N,(\Phi^* p)(x_{0,i})=\operatorname{sign}(a_{0,i}),(\Phi^* p)'(x_{0,i})=0_{\mathbb{R}^d}\right\}.$$

Its solution p_V is given by

$$p_V = (\Gamma_{x_0}^+)^* \begin{pmatrix} \operatorname{sign}(a_0) \\ 0_{(\mathbb{R}^d)^N} \end{pmatrix}, \tag{11}$$

and we define the vanishing derivatives precertificate as $\eta_V \overset{\text{def.}}{=} \Phi^* p_V$ (Γ_{x_0} is defined in equation (7)).

One can show that if $\|\eta_V\|_{\infty,X} \leqslant 1$ then η_V is a so-called valid certificate, which ensures that m_{a_0,x_0} is a solution to the constrained problem (corresponding to setting $w=0$ and $\lambda \to 0$ in $\mathcal{P}_\lambda(y)$)

$$\min_{\Phi m = y_0} |m|(X) \quad \text{where} \quad y_0 \overset{\text{def.}}{=} \Phi m_{a_0,x_0} \tag{$\mathcal{P}_0(y_0)$}.$$

More importantly, if it satisfies a stronger nondegeneracy condition detailed in definition 7 below, then η_V also ensures the stable recovery of the support in a small noise regime when solving the BLASSO. This result proved in [33] is stated in theorem 1.

Definition 7 (Nondegeneracy of η_V, [33]). We say that η_V is *nondegenerate* if

$$\begin{cases} \forall x \in X \setminus \bigcup_{i=1}^N \{x_{0,i}\}, & |\eta_V(x)| & < 1, \\ \forall i \in \{1,\dots,N\}, & \det(D^2\eta_V(x_{0,i})) & \neq 0. \end{cases} \tag{12}$$

Theorem 1 (Exact support recovery [33]). *Assume that $\varphi \in \mathrm{KER}^{(2)}$, Γ_{x_0} has full column rank and η_V is nondegenerate. Then there exists $C>0$ such that if $(\lambda, w) \in \mathbb{R}_+^* \times \mathcal{H}$ satisfies:*

$$\max\left(\lambda, \|w\|_{\mathcal{H}}/\lambda\right) \leqslant C,$$

then there is a unique solution $m_{a,x}$ to $\mathcal{P}_\lambda(y)$ composed of N Dirac masses such that $(a,x) = g(\lambda, w)$ where g is $\mathscr{C}^1$. In particular, by taking the regularization parameter $\lambda = \|w\|_{\mathcal{H}}/C$ proportional to the noise level, one obtains

$$|(a,x)-(a_0,x_0)|_\infty = O(\|w\|_{\mathcal{H}}),$$

where $|\cdot|_\infty$ is the ℓ^∞-norm for vectors.

Figure 9 displays some example of η_V associated to several Φ operators for 3D super-resolution fluorescence microscopy. This shows that for these inverse problems, the BLASSO stably recovers the support of the input measure if the noise level is not too high.

2.2. The super-resolution problem

In this section, X is considered to be 1D and we now tackle the super-resolution problem in presence of noise using the BLASSO. In this setting, we assume that the Dirac masses of the initial measure have positive amplitudes and cluster at some point $x_c \in \mathring{X}$. We parametrize this cluster as

$$m_{a_0,tz_0} \overset{\text{def.}}{=} \sum_{i=1}^{N} a_{0,i}\delta_{x_c+tz_{0,i}} \quad \text{where} \quad a_{0,i} > 0,\ z_{0,i} \in \mathbb{R},$$

and where the parameter $t > 0$ controls the separation distance between the spikes of the input measure. This problem in a multidimensional setup has been studied in [69].

In [28], the authors proved that the recovery of the support in presence of noise in the limit $t \to 0$ is controlled by the $2N-1$ vanishing derivatives precertificate.

Proposition 1 ($2N-1$ vanishing derivatives precertificate [28]). *If $\mathcal{I}_{2N-1}$ holds at x_c (see Definiton 5), there is a unique solution to the problem*

$$\inf\left\{ \|p\|_{\mathcal{H}} ; (\Phi^* p)(x_c) = 1, (\Phi^* p)'(x_c) = 0, \dots, (\Phi^* p)^{(2N-1)}(x_c) = 0 \right\}.$$

We denote by p_W its solution, given by

$$p_W = (\Psi^+_{2N-1})^* \delta_{2N} \quad \text{where} \quad \delta_{2N} \overset{\text{def.}}{=} (1, 0, \dots, 0)^T \in \mathbb{R}^{2N}, \tag{13}$$

and we define the $2N-1$ vanishing derivatives precertificate as $\eta_W \overset{\text{def.}}{=} \Phi^* p_W$ (see equation (9) for the definition of Ψ_{2N-1}).

Figure 1 shows η_W in the case of a Gaussian convolution kernel.

Remark 1. From proposition 1, one can easily see that η_W can equivalently be written as

$$\forall x \in X, \quad \eta_W(x) = \sum_{k=0}^{2N-1} \alpha_k \partial_2^{(k)} c_\varphi(x, x_c), \tag{14}$$

where c_φ is the correlation kernel associated to the correlation operator $\Phi^*\Phi$, namely $c_\varphi(x,x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}}$, and the coefficients α_k are defined by the equations

$$\forall k \in \{0, \dots, 2N-1\}, \quad \eta_W^{(k)}(x_c) = \delta_0^k. \tag{15}$$

If η_W satisfies some nondegeneracy property (see definition 8) then one can prove that the recovery of the support in a small noise regime when $t \to 0$ is possible. Theorem 2 (see [28]) makes this statement precise by quantifying the scaling between the noise level and the separation t to ensure the recovery.

Definition 8. Assume that $\mathcal{I}_{2N-1}$ holds at x_c and $\varphi \in \mathrm{KER}^{(2N)}$. We say that η_W is $(2N-1)$-nondegenerate if $\eta_W^{(2N)}(x_c) \neq 0$ and for all $x \in X \setminus \{x_c\}, |\eta_W(x)| < 1$.

Theorem 2. *Suppose that $\varphi \in \mathrm{KER}^{(2N+1)}$ and that η_W is $(2N-1)$-nondegenerate. Then there exist positive constants t_0,C,M (depending only on φ, a_0 and z_0) such that for all $0 < t < t_0$, for all $(\lambda, w) \in \mathrm{B}\left(0, Ct^{2N-1}\right)$ with $\|w\|_{\mathcal{H}} / \lambda \leqslant C$,*

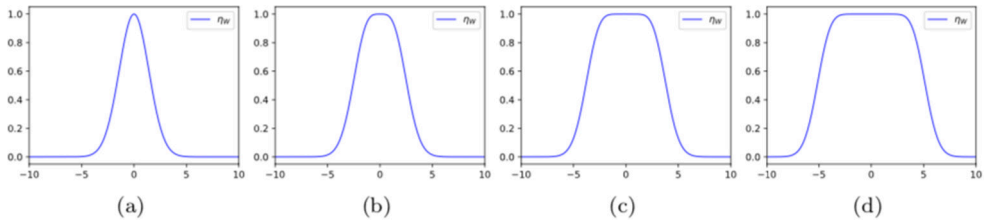

Figure 1. η_W for a Gaussian convolution ($x \in \mathbb{R}$, $\varphi(x) = \mathrm{e}^{-\frac{(\cdot - x)^2}{2\sigma^2}}$) for several numbers of spikes and $\sigma = 1$.

- *the BLASSO has a unique solution,*
- *that solution has exactly N spikes, and it is of the form* $m_{a,x_\mathrm{c}+tz}$, *with* $(a, z) = g(\lambda, w)$ *(where g is a* $\mathscr{C}^{2N}$ *function),*
- *the following inequality holds*

$$|(a, z) - (a_0, z_0)|_\infty \leqslant M \left(\frac{|\lambda|}{t^{2N-1}} + \frac{\|w\|_{\mathcal{H}}}{t^{2N-1}} \right).$$

In the next section, we prove that the main assumption of theorem 2 (the nondegeneracy of η_W) is satisfied for some operators Φ associated to Laplace measurements.

3. BLASSO for Laplace inversion

Most existing theoretical studies of super-resolution are focussed on translation-invariant operator Φ (convolution or Fourier measurements), see section 1.1. In contrast, this section presents new results for one of the most fundamental non-translation invariant operator: the Laplace transform (and variants).

The behavior of the Laplace transform is radically different from the one of the Fourier transform, and understanding the impact of the lack of translation invariance on super-resolution is relevant for many applications in imaging, including those considered in section 5. A first argument in favor of the BLASSO for the Laplace transform is the study provided in [32]. It essentially shows that the recovery of N positive spikes with stability of the support is possible using at least $2N$ measurements, regardless of the spacing of the spikes (and the spacings of the samples). The stability is asserted by showing that $\eta_{V,t}$ and η_W are nondegenerate, using abstract T-systems arguments.

Our strategy here is different, as we provide closed form expressions for η_W for these operators in order to show its nondegeneracy. The results presented here are thus complementary to those of [32], providing additional theoretical guarantees which backup our numerical observations. The main differences are

- we provide closed form expressions to η_W,
- some of the impulse responses we consider are L^2-normalized, a case which is not covered by the theory of [32],
- we cannot deal with arbitrary samplings μ, contrary to [32].

In this section, we suppose that N spikes are clustered at the position $x_\mathrm{c} \in \overset{\circ}{X}$ (which appears in the following results because of the non translation invariance of the kernel).

In the next section, we first detail the different continuous operators considered. Then, section 3.3 gives explicit formulas for η_W in two different setups and shows that η_W is $(2N-1)$-nondegenerate. Finally, section 3.4 provides some numerical material concerning η_W when the continuous kernels are approximated by a sampling.

3.1. Laplace operators

We suppose in this section that $X = [x_{\min}, x_{\max}] \subset \mathbb{R}_+^*$ is a compact interval, and that $\mathcal{H} = \mathrm{L}^2(\mathbb{R}_+, \mu)$ for some Radon measure μ on $\mathbb{R}_+$. A generic Laplace measurement kernel is defined as

$$\forall x \in X, \quad \varphi(x) \overset{\text{def.}}{=} \left(s \mapsto \xi(x)\mathrm{e}^{-sx}\right) \in \mathcal{H}. \tag{16}$$

This choice ensures that φ defines a valid operator Φ for all the the Laplace-like transform models presented below, provided $\mathrm{e}^{-x_{\min}s}\mathrm{d}\mu(s)$ has sufficiently many finite moments (in the following we require finite moments of order $4N-1$). The kernel is parametrized by a positive Radon measure μ on X (which models the sampling pattern) and a non-negative weighting function $\xi \in \mathscr{C}(X, \mathbb{R})$ (which takes care of the normalization of the measurement). The adjoint operator is thus defined as

$$(\Phi^* p)(x) = \xi(x) \int_{\mathbb{R}_+} \mathrm{e}^{-sx} p(s) \mathrm{d}\mu(s).$$

The choice of μ is let to the experimentalist and corresponds to the way samples are chosen. A discrete measure $\mu = \sum_{k=1}^K \mu_k \delta_{s_k}$ corresponds to using a finite set of samples values s_k. In this case, one can equivalently consider finite-dimensional observations $\mathcal{H} = \mathbb{R}^K$ and define $\varphi(x) \overset{\text{def.}}{=} (\xi(x)\mu_k \mathrm{e}^{-s_k x})_{k=1}^K \in \mathcal{H}$. A continuous measure $\mathrm{d}\mu(s) = h_\mu(s)\mathrm{d}s$ is a mathematical idealization, where a high value of $h_\mu(s)$ indicates that a high number of measurements have been taken for the index s (or equivalently that there is less noise for this measurement). On contrast, a value $\mu(s) = 0$ indicates that this measurement is not available.

In contrast, ξ can be freely chosen but strongly impacts the BLASSO problem by weighting the contribution of each position. The design of such a spatially-varying weighting is crucial (and non trivial) here because the operator Φ is not translation-invariant. The most frequent normalization for LASSO-type problems is

$$\xi(x)^2 = \frac{1}{\int_{\mathbb{R}_+} \mathrm{e}^{-2sx} \mathrm{d}\mu(s)}, \tag{17}$$

which guarantees that $\|\varphi(x)\|_{\mathcal{H}} = 1$ for all $x \in X$. See section 3.3.2 for more details for this normalization.

Note that both μ and ξ can be independently chosen, since they operate separately on the input and output variables x and s.

3.1.1. Correlation kernel. The properties of the BLASSO problem (and also the implementation of BLASSO solvers) only depend on the correlation operator $\Phi^*\Phi$ (rather than on the operator Φ itself). This operator reads $(\Phi^*\Phi m)(x) = \int_X c_\varphi(x, x')\mathrm{d}m(x')$ where c_φ is a symmetric positive kernel. For Laplace-type operators, it reads

$$\forall x, x' \in X, \quad c_\varphi(x, x') = \xi(x)\xi(x') \int_{\mathbb{R}_+} \mathrm{e}^{-(x+x')s} \mathrm{d}\mu(s).$$

The choice of normalization (17) ensures that $c_\varphi(x,x) = 1$.

We now detail in the following sections several particular cases covered by equation (16) and study the associated η_W.

3.2. Preliminaries results

This section gathers preliminary results useful for the computation of η_W.

One begins with two elementary lemmas. Their proofs are left to the reader. The first one is a simple consequence of the Faa di Bruno formula.

Lemma 1. *Let I, $I' \subset \mathbb{R}$ be open intervals, and $h : I' \to I$ be a smooth diffeomorphism. Let $x_c \in I$, $t_c := h^{-1}(x_c) \in I'$, and let $\eta : I \to \mathbb{R}$ be a smooth function. Then η satisfies*

$$\eta(x_c) = 1,\ \eta'(x_c) = 0, \dots, \eta^{(2N-1)}(x_c) = 0, \tag{18}$$

if and only if $\nu \overset{\text{def.}}{=} \eta \circ h$ satisfies

$$\nu(t_c) = 1,\ \nu'(t_c) = 0, \dots, \nu^{(2N-1)}(t_c) = 0. \tag{19}$$

Moreover, in that case, $\nu^{(2N)}(t_c) = \eta^{(2N)}(x_c)(h'(t_c))^{2N}$.

The next one follows from the general Leibniz rule.

Lemma 2. *Let I be an open interval, $t_c \in I$ and let $g : I \to \mathbb{R}$, $\eta : I \to \mathbb{R}$ be two smooth functions. If η satisfies:*

$$\eta(x_c) = 1,\ \eta'(x_c) = 0, \dots, \eta^{(2N-1)}(x_c) = 0, \tag{20}$$

then $P \overset{\text{def.}}{=} \eta \times g$ satisfies:

$$P(x_c) = g(x_c),\ P'(x_c) = g'(x_c), \dots,\ P^{(2N-1)}(x_c) = g^{(2N-1)}(x_c). \tag{21}$$

In particular, if $P \in \mathbb{R}_{2N-1}[T]$, then P is the Taylor expansion of g at x_c of order $2N-1$, and $\eta_W^{(2N)}(x_c) = -g^{(2N)}(x_c)/g(x_c)$ provided that $g(x_c) \neq 0$.

3.3. Explicit formulas for η_W in continuous settings

3.3.1. Classical Laplace operator. We suppose that $\mu = \mathcal{L}$, where $\mathcal{L}$ is the Lebesgue measure on $\mathbb{R}_+$, and $\xi = 1$. Then one has

$$c_\varphi(x,x') = \frac{1}{x+x'}. \tag{22}$$

The following proposition provides a formula for η_W in this unnormalized continuous setting and proves that it is nondegenerate.

Proposition 2. *η_W is $(2N-1)$-nondegenerate. More precisely, we have*

$$\forall x \in X, \quad \eta_W(x) = 1 - \left(\frac{x - x_c}{x + x_c}\right)^{2N}. \tag{23}$$

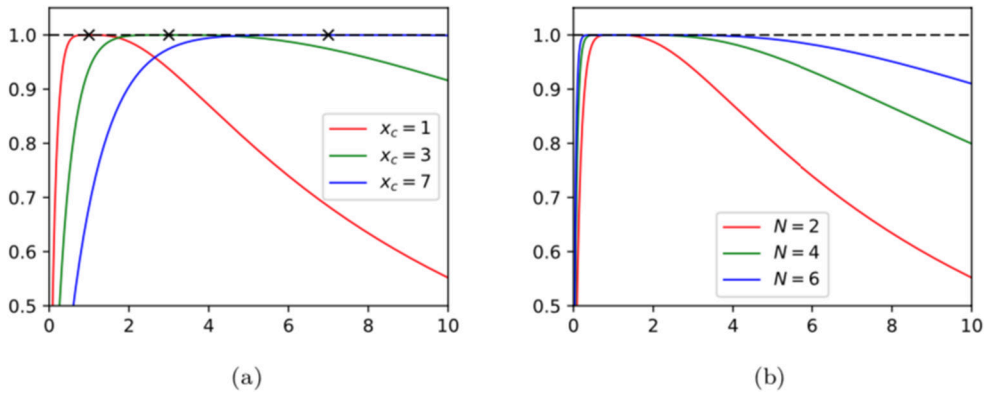

Figure 2. η_W for the unnormalized Laplace model for a varying x_c with fixed $N = 2$ and a fixed $x_c = 1$ with varying $N \in \{2, 4, 6\}$.

In figure 2, one sees that when the position x_c where the spikes cluster increases, the curvature of η_W at x_c decreases. This means that it is harder in this situation to perform the recovery. It reflects the exponential decay of the kernel φ.

Proof of proposition 2. From equations (14) and (22), one sees that η_W has the form

$$\eta_W(x) = \sum_{k=1}^{2N} \frac{\beta_k}{(x + x_c)^k}, \quad \text{where} \quad \beta_k \in \mathbb{R}.$$

We set $h : t \mapsto (1/t - x_c)$, $\nu \overset{\text{def.}}{=} \eta \circ h$ so that

$$\nu(t) = \sum_{k=1}^{2N} \beta_k t^k,$$

is a polynomial with degree at most $2N$ with $\nu(0) = 0$. By lemma 1, ν satisfies (19) at $t_c \overset{\text{def.}}{=} \frac{1}{2x_c}$. As a result, $\nu(t) = 1 + \beta_{2N}(t - t_c)^{2N}$. The constant β_{2N} is fixed by the condition $\nu(0) = 0$, so that $\nu(t) = 1 - \left(\frac{t - t_c}{t_c}\right)^{2N}$, and η_W is given by (23).

The $2N$ derivative is $\nu^{(2N)}(t_c) = -\frac{(2N)!}{(t_c)^{2N}}$, so that $\eta_W(x_c) = -\frac{(2N)!}{(2x_c)^{2N}} < 0$. □

3.3.2. L^2-normalized Laplace operator. We choose $\mu = \mathcal{L}$, where $\mathcal{L}$ is the Lebesgue measure on $\mathbb{R}_+$, and

$$\forall x \in X, \quad \xi(x) = \sqrt{\frac{1}{\int_{\mathbb{R}_+} \mathrm{e}^{-2sx} \mathrm{d}s}} = \sqrt{2x},$$

so that for all $x \in X$, $\varphi(x) : s \mapsto \sqrt{2x}\mathrm{e}^{-sx}$ and $\|\varphi(x)\|_{\mathcal{H}} = 1$. One gets

$$\forall x, x' \in X, \quad c_\varphi(x, x') \overset{\text{def.}}{=} \langle \varphi(x), \varphi(x') \rangle_{\mathcal{H}} = \frac{2\sqrt{xx'}}{x + x'}. \tag{24}$$

The following proposition provides a formula for η_W in this normalized setting and proves that it is nondegenerate.

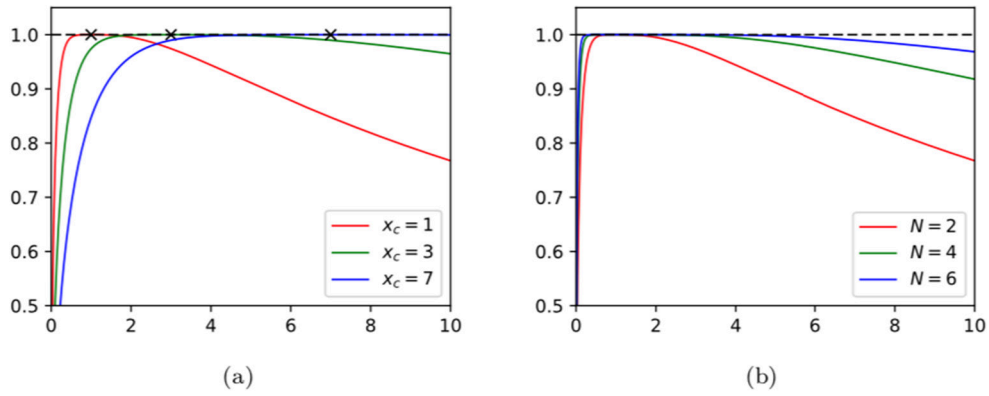

Figure 3. η_W for the normalized Laplace model for a varying x_c with fixed $N = 2$ and a fixed $x_c = 1$ with varying $N \in \{2, 4, 6\}$. (a) $N = 2$. (b) $x_c = 1$.

Proposition 3. *η_W is $(2N-1)$-nondegenerate. More precisely, we have the following formula:*

$$\forall x \in X, \quad \eta_W(x) = \frac{2\sqrt{x x_c}}{x + x_c} \sum_{k=0}^{N-1} \frac{(2k)!}{2^{2k}(k!)^2} \left(\frac{x - x_c}{x + x_c}\right)^{2k}. \tag{25}$$

In figure 3, one sees that when the position x_c where the spikes cluster increases then the curvature of η_W at x_c decreases. The interpretation is the same as in the previous paragraph.

Proof of proposition 3. From the general Leibniz rule, we have for all $n \in \{0, \dots, 2N-1\}$ and for all $x, x' \in X$:

$$\frac{\mathrm{d}^n}{\mathrm{d}x'^n}\left(c_\varphi(x, x')\right) = 2\sqrt{x} \sum_{k=0}^{n} \binom{n}{k} \frac{\mathrm{d}^{n-k}}{\mathrm{d}x'^{n-k}}\left(\sqrt{x'}\right) \frac{\mathrm{d}^k}{\mathrm{d}x'^k}\left(\frac{1}{x + x'}\right).$$

Evaluating this expression at $x' = x_c$, one gets that:

$$\partial_2^n c_\varphi(x, x_c) = \sqrt{x} \sum_{k=0}^{n} \frac{\alpha_k}{(x + x_c)^{k+1}},$$

for some coefficients $\alpha_k \in \mathbb{R}$. As a result, η_W is the unique function the form

$$\eta_W(x) = \sqrt{x} \sum_{k=0}^{2N-1} \frac{\beta_k}{(x + x_c)^{k+1}}$$

for some coefficients $\beta_k \in \mathbb{R}$, which satisfies (15). As before, we set $t = \frac{1}{x + x_c}$, that is $x = h(t) \overset{\text{def.}}{=} \frac{1}{t} - x_c$, and h is a diffeomorphism of $(0, 1/x_c)$ onto $(0, +\infty)$. Then:

$$\eta_W \circ h(t) = \sqrt{\frac{1}{t} - x_c}\, t P(t) = \sqrt{t - t^2 x_c}\, P(t),$$

where $P(T) = \sum_{k=0}^{2N-1} \beta_k T^k \in \mathbb{R}_{2N-1}[T]$.

By lemmas 1 and 2, P is the Taylor expansion of order $2N-1$ of $g : t \mapsto \frac{1}{\sqrt{t-t^2 x_c}}$ at $t_c = h^{-1}(x_c) = \frac{1}{2x_c}$. Setting $t = u + \frac{1}{2x_c}$, we note that:

$$\frac{1}{\sqrt{t-t^2x_c}} = \frac{2\sqrt{x_c}}{\sqrt{1-(2ux_c)^2}} \quad \text{and} \quad \frac{1}{\sqrt{1-z^2}} = \sum_{k=0}^{N-1} \frac{(2k)!}{2^{2k}(k!)^2} z^{2k} + o(z^{2N-1}).$$

One deduces that

$$\frac{1}{\sqrt{t-t^2x_c}} = 2\sqrt{x_c} \sum_{k=0}^{N-1} \frac{(2k)!}{2^{2k}(k!)^2} \left[2x_c(t-t_c)\right]^{2k} + o((t-t_c)^{2N-1}).$$

As a result, P is given by $P(t) = 2\sqrt{x_c}\sum_{k=0}^{N-1} \frac{(2k)!}{2^{2k}(k!)^2} \left[2x_c(t-t_c)\right]^{2k}$ and

$$\eta_W \circ h(t) = \sqrt{t-t^2x_c}P(t) \tag{26}$$

$$= 1 - \frac{\sum_{k=M}^{+\infty} \frac{(2k)!}{2^{2k}(k!)^2} \left[2x_c(t-t_c)\right]^{2k}}{\sum_{k=0}^{+\infty} \frac{(2k)!}{2^{2k}(k!)^2} \left[2x_c(t-t_c)\right]^{2k}}. \tag{27}$$

One sees that $|\eta_W \circ h(t)| < 1$ for all $t \in (0, \frac{1}{x_c}) \setminus \{\frac{1}{2x_c}\}$, and by lemma 2,

$$(\eta_W \circ h)^{(2N)}(t_c) = -g^{(2N)}(t_c)/g(t_c) = -\frac{((2N)!)^2}{(N!)^2} x_c^{2N} < 0 \tag{28}$$

so that $\eta_W \circ h$ (hence η_W) is $(2N-1)$-nondegenerate. One recovers η_W by composing with h^{-1}, noting that $2x_c(t-t_c) = \frac{x_c - x}{x + x_c}$. □

3.4. Sampled approximations

The previous two cases (normalized and unnormalized versions of the Laplace transform) correspond to mathematical idealizations. In practice, one needs to restrict the sampling patterns by limiting their ranges and considering discrete samples. The following two setups are involved in the application of section 5.

3.4.1. Discretized unnormalized Laplace. We assume that $\mu = \sum_{k=0}^{K-1} \delta_{s_k}$ and $\xi = 1$. Then $\varphi(x) = (\mathrm{e}^{-s_k x})_{k=0}^{K-1} \in \mathbb{R}^K$ and:

$$c_\varphi(x, x') = \sum_{k=0}^{K-1} \mathrm{e}^{-s_k(x+x')}.$$

3.4.2. Discretized L^2-normalized Laplace. We let $\mu = \sum_{k=0}^{K-1} \delta_{s_k}$ and $\xi(x) = \left(\sum_{k=0}^{K-1} \mathrm{e}^{-2s_k x}\right)^{-1/2}$. Then $\varphi(x) = \xi(x)(\mathrm{e}^{-s_k x})_{k=0}^{K-1} \in \mathbb{R}^K$, $\|\varphi(x)\|_{\mathcal{H}} = 1$ and:

$$c_\varphi(x, x') = \xi(x)\xi(x') \sum_{k=0}^{K-1} \mathrm{e}^{-s_k(x+x')}.$$

In contrast to the continuous setups of section 3.3, we do not have closed-form expressions for η_W. However, if a sequence of measures, e.g. $\mu_n = \sum_{k=0}^{K_n-1} \mu_{n,k}\delta_{s_{n,k}}$ converges in a suitable sense towards the Lebesgue measure $\mu = \mathcal{L}$, the following proposition shows that the corresponding η_W must be nondegenerate for n large enough. We consider both the unnormalized and L^2-normalized setups, corresponding respectively to

$$c_{\varphi_n}(x,x') = \int_{\mathbb{R}_+} \mathrm{e}^{-(x+x')s}\mathrm{d}\mu_n(s), \text{ and}$$
$$c_{\varphi_n}(x,x') = \xi_n(x)\xi_n(x')\int_{\mathbb{R}_+} \mathrm{e}^{-(x+x')s}\mathrm{d}\mu_n(s) \quad \text{where} \quad \xi_n(x) = \int_{\mathbb{R}_+} \mathrm{e}^{-2xs}\mathrm{d}\mu_n(s),$$

and similarly for c_φ and $\mu = \mathcal{L}$.

Proposition 4. *Let $(\mu_n)_{n\in\mathbb{N}}$ be a sequence of positive measures which converges towards the Lebesgue measure μ in the local weak-$*$ topology, i.e.*

$$\forall \psi \in \mathscr{C}_{\mathrm{c}}(\mathbb{R}_+), \quad \lim_{n\to+\infty} \int_{\mathbb{R}_+} \psi(s)\mathrm{d}\mu_n(s) = \int_{\mathbb{R}_+} \psi(s)\mathrm{d}s,$$

and such that

$$\sup_{n\in\mathbb{N}} \int_{\mathbb{R}_+} (1+s^{4N-1})\mathrm{e}^{-x_{\min}s}\mathrm{d}\mu_n(s) < +\infty. \tag{29}$$

Then, both in the unnormalized and the L^2-normalized case, for n large enough, the $2N-1$ vanishing derivatives precertificate $\eta_{W,n}$ is $(2N-1)$-nondegenerate.

Proof. Let us denote by $\Psi_{2N-1}^{[n]} = (\varphi_0, \dots, \varphi_{2N-1})$ (resp. Ψ_{2N-1}) the impulse response derivatives corresponding to μ_n (resp. $\mu = \mathcal{L}$), and by η_W the $2N-1$ vanishing derivatives precertificate for $\mu = \mathcal{L}$. First, in view of sections 3.3.1 and 3.3.2, we observe that the result follows immediately if we prove that

$$\lim_{n\to+\infty} \Psi_{2N-1}^{[n]*}\Psi_{2N-1}^{[n]} = \Psi_{2N-1}^{*}\Psi_{2N-1}, \tag{30}$$

(as it implies the linear independence of $(\varphi_0, \dots, \varphi_{2N-1})$ for n large enough), and that

$$\forall i \in \{0,1,\dots,2N\}, \quad \lim_{n\to+\infty} \left\| \eta_{W,n}^{(i)} - \eta_W^{(i)} \right\|_{\infty,X} = 0, \tag{31}$$

(as it implies $|\eta_{W,n}(x)| < 1$ for $x \neq x_{\mathrm{c}}$ and $\eta_{W,n}^{(2N)}(x_{\mathrm{c}}) < 0$ for n large enough).

We recall from (14) that $\eta_{W,n}$ is given by $\eta_{W,n}(x) = \sum_{i=0}^{2N-1} \alpha_i^{[n]}\partial_2^{(i)}c_{\varphi_n}(x,x_{\mathrm{c}})$ where $\alpha^{[n]} = (\Psi_{2N-1}^{[n]*}\Psi_{2N-1}^{[n]})^{-1}\delta_{2N}$ (provided the matrix is invertible), and the (i,j)-entry of $(\Psi_{2N-1}^{[n]*}\Psi_{2N-1}^{[n]})$ is $\partial_1^{(i)}\partial_2^{(j)}c_{\varphi_n}(x_{\mathrm{c}},x_{\mathrm{c}})$. As a consequence, both (30) and (31) are established if we can prove that

$$\lim_{n\to+\infty} \sup_{x,x'\in[x_{\min},x_{\max}]} \left| \partial_1^{(i)}\partial_2^{(j)}c_{\varphi_n}(x,x') - \partial_1^{(i)}\partial_2^{(j)}c_{\varphi}(x,x') \right| = 0, \tag{32}$$

for all $i \in \{0,\dots,2N\}$, $j \in \{0,\dots,2N-1\}$.

First, we prove (32) in the unnormalized case, i.e. $c_{\varphi_n}(x,x') = \int_{\mathbb{R}_+} \mathrm{e}^{-(x+x')s}\mathrm{d}\mu_n(s)$. The

dominated convergence theorem ensures that $\partial_1^{(i)}\partial_2^{(j)}c_{\varphi_n}(x,x') = \int_{\mathbb{R}_+} s^{i+j}\mathrm{e}^{-(x+x')s}\mathrm{d}\mu_n(s)$ (and similarly for c_φ and μ).

Let $(x,x') \in [x_{\min}, x_{\max}]^2$ and let $\psi \in \mathscr{C}_{\mathrm{c}}(\mathbb{R}_+)$ such that $\psi(s) = 1$ for $s \in [0,1]$, $\psi(s) = 0$ for $s \geqslant 2$, and $0 \leqslant \psi \leqslant 1$ on $\mathbb{R}_+$. We denote by C the supremum in (29).

Let $\varepsilon > 0$ and $A > 0$. Then,

$$\begin{aligned}
&\left|\int_{\mathbb{R}_+} s^{i+j}\mathrm{e}^{-(x+x')s}\mathrm{d}\mu_n(s) - \int_{\mathbb{R}_+} s^{i+j}\mathrm{e}^{-(x+x')s}\mathrm{d}s\right| \\
&\leqslant \underbrace{\left|\int_{\mathbb{R}_+} s^{i+j}\mathrm{e}^{-(x+x')s}\psi\left(\frac{s}{A}\right)\mathrm{d}\mu_n(s) - \int_{\mathbb{R}_+} s^{i+j}\mathrm{e}^{-(x+x')s}\psi\left(\frac{s}{A}\right)\mathrm{d}s\right|}_{\stackrel{\text{def.}}{=} a} \\
&+ \underbrace{\left|\int_{\mathbb{R}_+} s^{i+j}\mathrm{e}^{-(x+x')s}(1-\psi\left(\frac{s}{A}\right))\mathrm{d}\mu_n(s)\right|}_{=b} + \underbrace{\left|\int_{\mathbb{R}_+} s^{i+j}\mathrm{e}^{-(x+x')s}(1-\psi\left(\frac{s}{A}\right))\mathrm{d}s\right|}_{=c}.
\end{aligned}$$

We have

$$\begin{aligned}
c &\leqslant \int_A^{+\infty}(1+s^{4N-1})\mathrm{e}^{-2x_{\min}s}\mathrm{d}s, \\
\text{and}\quad b &\leqslant \mathrm{e}^{-x_{\min}A}\int_{\mathbb{R}_+}(1+s^{4N-1})\mathrm{e}^{-x_{\min}s}\mathrm{d}\mu_n(s) \leqslant \mathrm{e}^{-x_{\min}A}C.
\end{aligned}$$

We choose $A > 0$ sufficiently large so that $\int_A^{+\infty}(1+s^{4N-1})\mathrm{e}^{-2x_{\min}s}\mathrm{d}s \leqslant \varepsilon$ and $\mathrm{e}^{-x_{\min}A}C \leqslant \varepsilon$, hence $\max(b,c) \leqslant \varepsilon$.

Now, to prove that a is uniformly small for $(x,x') \in [x_{\min}, x_{\max}]^2$ as $n \to +\infty$, we apply lemma 3 to $((x,x'),s) \mapsto s^{i+j}\mathrm{e}^{-(x+x')s}\psi\left(\frac{s}{A}\right)$ defined on $[x_{\min}, x_{\max}]^2 \times [0, 2A]$. This yields the desired result.

The proof for the normalized case readily follows from the uniform convergence of the unnormalized case and the fact that the normalization factors $\xi_n(x) = \left(\int_{\mathbb{R}_+}\mathrm{e}^{-2sx}\mathrm{d}\mu_n(s)\right)^{-1/2} \leqslant \left(\int_{\mathbb{R}_+}\mathrm{e}^{-2sx_{\max}}\mathrm{d}\mu_n(s)\right)^{-1/2}$ are upper bounded by some positive constant independent of n. □

Lemma 3. *Let X and S be two compact metric spaces, and $\psi \in \mathscr{C}(X \times S)$. If $\{\mu_n\}_{n\in\mathbb{N}}$ and μ are Radon measures such that $\mu_n \xrightarrow{*} \mu$ in the weak-$*$ convergence of $\mathcal{M}(S)$, then*

$$\lim_{n\to+\infty}\int_S \psi(x,s)\mathrm{d}\mu_n(s) = \int_S \psi(x,s)\mathrm{d}\mu(s),$$

uniformly in $x \in X$.

Proof. We note that the mapping $(\eta,\nu) \mapsto \int_S \eta\mathrm{d}\nu$ is continuous on $\mathscr{C}(S) \times \mathcal{M}(S)$. Since $x \mapsto \psi(x,\cdot)$ is continuous from X to $\mathscr{C}(S)$, the mapping

$$F : (x,\nu) \longmapsto \int_S \psi(x,s)\mathrm{d}\nu(s) \tag{33}$$

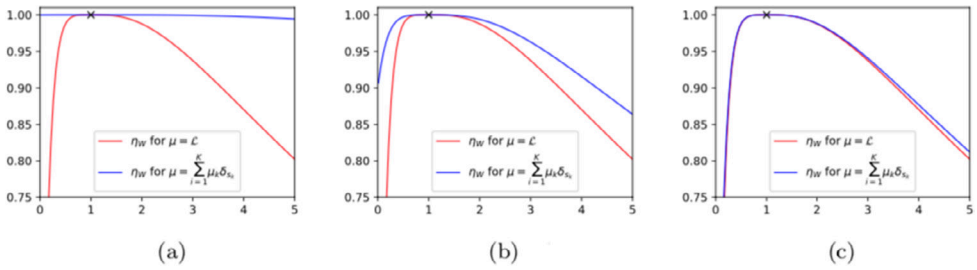

Figure 4. Approximation of η_W for the unnormalized continuous Laplace operator (see proposition 2) by the η_W obtained for discretized unnormalized Laplace operators. (a) $K = 10$. (b) $K = 120$. (c) $K = 800$.

is continuous on $X \times \mathcal{M}(S)$.

Now, since S is compact, $\mathcal{M}(S)$ is the dual of the Banach space $\mathscr{C}(S)$, and the Banach–Steinhaus theorem implies that there exists $R > 0$ such that $\sup_n |\mu_n|(S) \leqslant R$ (and $|\mu|(S) \leqslant R$).

The subspace $\mathcal{B}_R \overset{\text{def.}}{=} \{\nu \in \mathcal{M}(S); |\nu|(S) \leqslant R\}$ is metrizable for the weak-$*$ topology and compact. As a result, the mapping F is uniformly continuous on the compact $X \times \mathcal{B}_R$. In particular, as $\mu_n \to \mu$ in $\mathcal{B}_R$,

$$\sup_{x \in X} \left| \int_S \psi(x,s) \mathrm{d}\mu_n(s) - \int_S \psi(x,s) \mathrm{d}\mu(s) \right| \to 0. \qquad \square$$

Figure 4 illustrates this convergence between the precertificates in the unnormalized case.

4. The SFW algorithm

In this section, we present the Sliding Frank–Wolfe (see algorithm 2), a new version of the modified Frank–Wolfe algorithm introduced in [12]. Moreover, we prove in theorem 3 that it converges in a finite number of steps under mild assumptions. The code can be found in https://github.com/qdenoyelle.

We suppose in this section that $X \subset \mathbb{R}^d$ is compact, or $X = \mathbb{T}^d$ with $d \in \mathbb{N}^*$ and $\varphi \in \mathrm{KER}^{(2)}$ (see definition 4).

4.1. The algorithm

4.1.1. Frank–Wolfe algorithm. The Frank–Wolfe (FW) algorithm [44], also called the conditional gradient method (CGM) [63] solves the following optimization problem

$$\min_{m \in C} f(m), \tag{34}$$

where C is a weakly compact convex set of a Banach space, and f is a differentiable convex function. For instance, in the case of sparse recovery problems, m is a measure and C is a subset of $\mathcal{M}(X)$. A chief advantage of FW with respect to most first order optimization scheme (such as gradient descent or proximal splitting method) is that it does not rely on any underlying Hilbertian structure and only makes use of directional derivatives. It is thus particularly well adapted to optimize over the space of Radon measures. The algorithm is detailed in algorithm 1.

Algorithm 1. Frank–Wolfe algorithm.

1: **for** $k = 0, \dots, n$ **do**
2: Minimize: $s^{[k]} \ni \operatorname{argmin}_{s \in C} f(m^{[k]}) + \mathrm{d}f(m^{[k]})[s - m^{[k]}]$.
3: **if** $\mathrm{d}f(m^{[k]})[s^{[k]} - m^{[k]}] = 0$ **then**
4: $m^{[k]}$ solution of (34). Stop.
5: **else**
6: Step research: $\gamma^{[k]} \leftarrow \frac{2}{k+2}$ or $\gamma^{[k]} \ni \operatorname{argmin}_{\gamma \in [0,1]} f(m^{[k]} + \gamma(s^{[k]} - m^{[k]}))$.
7: Update: $m^{[k+1]} \leftarrow m^{[k]} + \gamma^{[k]}(s^{[k]} - m^{[k]})$.
8: **end if**
9: **end for**

Let us note that the FW algorithm is naturally endowed with a stopping criterion in step 3 (see for instance [26, chapter 3, section 1.2]) which is equivalent to the standard optimality condition for constrained convex problems

$$\forall s \in C, \quad \mathrm{d}f(m^{[k]})[s - m^{[k]}] \geqslant 0. \tag{35}$$

4.1.2. Frank–Wolfe for the BLASSO. The FW algorithm cannot be applied directly to the BLASSO because it is an optimization problem over $\mathcal{M}(X)$ which is not bounded and the objective function

$$\forall m \in \mathcal{M}(X), \quad T_\lambda(m) \overset{\text{def.}}{=} \frac{1}{2} \|\Phi m - y\|_{\mathcal{H}}^2 + \lambda |m|(X), \tag{36}$$

is not differentiable. Instead, we propose to consider an equivalent problem to the BLASSO, using an epigraphical lift (following an idea of [46]), which is presented in lemma 4.

Lemma 4. *The BLASSO*

$$\min_{m \in \mathcal{M}(X)} T_\lambda(m) \overset{\text{def.}}{=} \frac{1}{2} \|\Phi m - y\|_{\mathcal{H}}^2 + \lambda |m|(X). \qquad (\mathcal{P}_\lambda(y))$$

is equivalent to

$$\min_{(t,m) \in C} \tilde{T}_\lambda(m,t) \overset{\text{def.}}{=} \frac{1}{2} \|\Phi m - y\|_{\mathcal{H}}^2 + \lambda t, \qquad (\tilde{\mathcal{P}}_\lambda(y))$$

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

The equivalence stated in lemma 4 is to be understood in the following sense: m is a solution to $(\mathcal{P}_\lambda(y))$ if and only if (t,m) is a solution to $(\tilde{\mathcal{P}}_\lambda(y))$ for some $t \geqslant 0$. Moreover, in that case $t = |m|(X)$ and $\tilde{T}_\lambda(m,t) = T_\lambda(m)$. As a result, one can directly translate the FW algorithm (see algorithm 1) to $\mathcal{P}_\lambda(y)$.

Proof. Let m_* be a minimizer of T_λ on $\mathcal{M}(X)$, then we have

$$T_\lambda(m_*) \leqslant T_\lambda(0) = \lambda M. \tag{37}$$

Hence, one can restrict the BLASSO to the set of measures $m \in \mathcal{M}(X)$ such that $|m|(X) \leqslant M$ and $\tilde{\mathcal{P}}_\lambda(y)$ is obtained using an epigraphical representation. □

The next two remarks discuss the applicability of standard results on FW to the BLASSO.

Remark 2 (Well-posedness). The FW algorithm is well defined for $\tilde{\mathcal{P}}_\lambda(y)$. Indeed, $\tilde{T}_\lambda$ is a differentiable functional on the Banach space $\mathbb{R} \times \mathcal{M}(X)$, with differential

$$\mathrm{d}\tilde{T}_\lambda(t,m) : (t',m') \longmapsto \int_X \Phi^*(\Phi m - y)\mathrm{d}m' + \lambda t'. \tag{38}$$

Although C is not weakly compact (otherwise, by the Eberlein–Shmulyan theorem, $\mathcal{M}(X)$ would be reflexive), it is compact for the weak-$*$ topology: as $\mathrm{d}\tilde{T}_\lambda(t,m)$ is represented by $(\lambda, \Phi^*(\Phi m - y)) \in \mathbb{R} \times \mathscr{C}_0(X)$, it does reach its minimum on C.

Remark 3 (Rate of convergence). Let us note that $\mathrm{d}\tilde{T}_\lambda$ is Lipschitz continuous (because $\varphi \in \mathrm{KER}^{(2)}$), hence by classical results for the study of the convergence of the FW algorithm, one obtains the $O(1/k)$ rate of convergence in the objective function for any minimizing sequence for the BLASSO.

Lemma 5 ([26, theorem 3.1.7]). *Let $(t_k, m^{[k]})_{k\in\mathbb{N}}$ be a sequence generated by algorithm 1 applied to $\tilde{\mathcal{P}}_\lambda(y)$. Then, there exists $C_1 > 0$ such that for any m_* solution of $\mathcal{P}_\lambda(y)$ we have*

$$\forall k \in \mathbb{N}^*, \quad T_\lambda(m^{[k]}) - T_\lambda(m_*) \leqslant \frac{C_1}{k}. \tag{39}$$

Next, we discuss how the minimization step yields a greedy approach and a natural stopping criterion. The following two remarks also crucially relate the algorithm to the dual certificate of $(\mathcal{P}_\lambda(y))$.

Remark 4 (Greedy approach). Obviously, the FW algorithm is only interesting if, in step 2 of algorithm 1, one is able to minimize the linear form $s \mapsto \mathrm{d}\tilde{T}_\lambda(t^{[k]}, m^{[k]})[s]$ on C. That linear form reaches its minimum at least at one extreme point of C, i.e. $s = (0,0)$ or points of the form $s = (M, \pm M\delta_x)$ for $x \in X$. Finding a minimizer among those points amounts to finding a point x in

$$\mathrm{argmin}_{x\in X} \left(\pm \frac{1}{\lambda} \left(\Phi^*(y - \Phi m^{[k]}) \right)(x) + 1 \right) \lambda M,$$

$$\text{or equivalently in} \quad \mathrm{argmax}_{x\in X} \left(\left| \eta^{[k]}(x) \right| - 1 \right) \quad \text{where} \quad \eta^{[k]} \overset{\text{def.}}{=} \frac{1}{\lambda} \left(\Phi^*(y - \Phi m^{[k]}) \right)$$

(note the similarity of $\eta^{[k]}$ with the dual certificate defined in (10)).

As a consequence, at each step 7 of algorithm 1, a new spike is created at some point in $\mathrm{argmax}_X \left|\eta^{[k]}\right|$ (unless $s = (0,0)$ is optimal, which means that $\left\|\eta^{[k]}\right\|_{\infty,X} \leqslant 1$). This spike creation step is at the core of the algorithms in [12] and [10].

Remark 5 (Stopping criterion). It is interesting to relate the stopping criterion

$$(t^{[k]}, m^{[k]}) \in \mathrm{argmin}_{s\in C} \mathrm{d}\tilde{T}_\lambda(t^{[k]}, m^{[k]})[s],$$

with the dual certificates for $(\mathcal{P}_\lambda(y))$. As noted above (see equation (35)), the stopping cri-

terion is equivalent to $(t^{[k]}, m^{[k]})$ being a solution, hence $t^{[k]} = |m^{[k]}|(X)$. If $m^{[k]} \neq 0$, without loss of generality we write $m^{[k]} = \sum_{i=1}^{N^{[k]}} a_i^{[k]} \delta_{x_i^{[k]}}$ where the $x_i^{[k]}$'s are distinct, so that $t^{[k]} = |m^{[k]}|(X) = \sum_i \left|a_i^{[k]}\right|$. We also set $\varepsilon_i^{[k]} \overset{\text{def.}}{=} \operatorname{sign}(a_i^{[k]})$ and $L \overset{\text{def.}}{=} \mathrm{d}\tilde{T}_\lambda(t^{[k]}, m^{[k]})$.

Assume first that $|m^{[k]}|(X) < M$, so that the smallest face of C which contains $(t^{[k]}, m^{[k]})$ is

$$F \overset{\text{def.}}{=} \operatorname{conv}\left\{(0,0), (M, M\varepsilon_1^{[k]}\delta_{x_1^{[k]}}), \dots, (M, M\varepsilon_{N^{[k]}}^{[k]}\delta_{x_{N^{[k]}}^{[k]}})\right\}.$$

Since $\operatorname{argmin}_{s\in C} L$ is a face of C containing $(t^{[k]}, m^{[k]})$ (see [72, section 18]), it must contain F. Hence

$$L(0,0) = L(M, M\varepsilon_1^{[k]}\delta_{x_1^{[k]}}) = \cdots = L(M, M\varepsilon_{N^{[k]}}^{[k]}\delta_{x_{N^{[k]}}^{[k]}}) = \min_C L. \tag{40}$$

Now, if $|m^{[k]}|(X) = M$, it means that $\tilde{T}_\lambda(t^{[k]}, m^{[k]}) = \tilde{T}_\lambda(0,0)$, so that by convexity of $\tilde{T}_\lambda$ and optimality of $(t^{[k]}, m^{[k]})$ one has $L(t^{[k]}, m^{[k]}) = \mathrm{d}\tilde{T}_\lambda(t^{[k]}, m^{[k]})[t^{[k]}, m^{[k]}] = 0 = L(0,0)$. As the smallest face which contains $(t^{[k]}, m^{[k]})$ is

$$F' \overset{\text{def.}}{=} \operatorname{conv}\left\{(M, M\varepsilon_1^{[k]}\delta_{x_1^{[k]}}), \dots, (M, M\varepsilon_{N^{[k]}}^{[k]}\delta_{x_{N^{[k]}}^{[k]}})\right\},$$

we deduce as above that (40) holds.

In particular, $L(0,0) \leqslant \inf_{x\in X} L(M, \pm M\delta_x)$ yields

$$0 \leqslant \inf_{x\in X} \left(-\left|\eta^{[k]}(x)\right| + 1\right), \tag{41}$$

that is $\left\|\eta^{[k]}\right\|_{\infty,X} \leqslant 1$. Moreover $L(t^{[k]}, m^{[k]}) = \sum_{j=1}^{N^{[k]}} \left|a_j^{[k]}\right| L\left(M, M\varepsilon_j^{[k]}\delta_{x_j^{[k]}}\right) \leqslant \sum_{j=1}^{N} \left|a_j^{[k]}\right| L(M, \pm M\delta_{x_j^{[k]}})$, yields

$$-\sum_{j=1}^{N^{[k]}} a_j^{[k]}\eta^{[k]}(x_j^{[k]}) \leqslant -\sum_{j=1}^{N^{[k]}} \left|a_j^{[k]}\right| \left|\eta^{[k]}(x_j^{[k]})\right|,$$

from which we deduce $\eta^{[k]}(x_j^{[k]}) = \operatorname{sign}(a_j^{[k]})$.

As a result, when the FW algorithm stops (if it does), we observe that *the quantity $\eta^{[k]}$ it has constructed is the dual certificate* for $(\mathcal{P}_\lambda(y))$. If $m^{[k]} = 0$, the argument is similar (as (41) must hold).

4.1.3. The Sliding Frank–Wolfe *algorithm.* Applying directly algorithm 1 yields a sequence of measures $(m^{[k]})_{k\in\mathbb{N}}$ which weakly-$*$ converges towards some solution m_* in a greedy way. But the generated measures $m^{[k]}$ are not very sparse compared to m_*, each Dirac mass of m_* being approximated by a multitude of Dirac masses of $m^{[k]}$ with inexact positions. It is therefore suggested in [12], and strongly advocated in [10], to modify the Frank–Wolfe algorithm for the resolution of the BLASSO and to let the Dirac positions move.

One important feature of the FW algorithm (algorithm 1), as noted in [10, 57], is that in the update step 7, *the point $m^{[k+1]}$ may be replaced with any point $m \in C$ which has lower energy*, without breaking the convergence property and the convergence rate. The Frank–Wolfe algorithm with our modified update step is described in algorithm 2, we call it the

Sliding Frank–Wolfe (SFW) algorithm. The solvers used for the different steps of this algorithm are detailed in remark 9. Since the t variable is only auxiliary in $(\tilde{\mathcal{P}}_\lambda(y))$, we omit it and we formulate directly algorithm 2 in terms of m only.

Algorithm 2. Sliding Frank–Wolfe algorithm.

1: Initialize with $m^{[0]} = 0$ and $n = 0$.
2: **for** $k = 0, \dots, n$ **do**
3: $m^{[k]} = \sum_{i=1}^{N^{[k]}} a_i^{[k]} \delta_{x_i^{[k]}}$, $a_i^{[k]} \in \mathbb{R}$, $x_i^{[k]}$ pairwise distincts, find $x_*^{[k]} \in X$ s.t.:
$$x_*^{[k]} \in \arg\max_{x \in X} |\eta^{[k]}(x)| \quad \text{where} \quad \eta^{[k]} \stackrel{\text{def}}{=} \tfrac{1}{\lambda}\Phi^*(y - \Phi m^{[k]}),$$
4: **if** $|\eta^{[k]}(x_*^{[k]})| \leqslant 1$ **then**
5: $m^{[k]}$ is a solution of $\mathcal{P}_\lambda(y)$. Stop.
6: **else**
7: Obtain $m^{[k+1/2]} = \sum_{i=1}^{N^{[k]}} a_i^{[k+1/2]} \delta_{x_i^{[k]}} + a_{N^{[k]}+1}^{[k+1/2]} \delta_{x_*^{[k]}}$, s.t.:
$$a^{[k+1/2]} \in \arg\min_{a \in \mathbb{R}^{N^{[k]}+1}} \tfrac{1}{2} \left\| \Phi_{x^{[k+1/2]}} a - y \right\|_{\mathcal{H}}^2 + \lambda \|a\|_1$$
$$\text{where} \quad x^{[k+1/2]} = (x_1^{[k]}, \dots, x_{N^{[k]}}^{[k]}, x_*^{[k]})$$
8: Find a critical point $m^{[k+1]} = \sum_{i=1}^{N^{[k]}+1} a_i^{[k+1]} \delta_{x_i^{[k+1]}}$ by minimizing locally
$$(a, x) \in \mathbb{R}^{N^{[k]}+1} \times X^{N^{[k]}+1} \mapsto \tfrac{1}{2} \left\| \Phi_x a - y \right\|_{\mathcal{H}}^2 + \lambda \|a\|_1,$$
using as initial point $(a^{[k+1/2]}, x^{[k+1/2]})$.
9: Eventually remove zero amplitudes Dirac masses from $m^{[k+1]}$.
10: **end if**
11: **end for**

As we detail below, the algorithm slightly (but crucially) differs from the one in [10]. The main ingredient is to replace the final update with the local minimization of a non-convex function updating both the positions and the amplitudes of the spikes (whereas [10] update successively the amplitudes and the positions). It is crucial to note that it is only required in our algorithm that step 8 finds a critical point of the objective function $(a, x) \mapsto T_\lambda(m_{a,x})$.

Remark 6 (Links between FW applied to $\tilde{\mathcal{P}}_\lambda(y)$ and the SFW). Algorithm 2 is a valid variant of FW, as the update step decreases the energy more than the standard convex combination using $\gamma^{[k]}$. Indeed,

$$T_\lambda(m^{[k+1]}) \leqslant T_\lambda(m^{[k+1/2]}) \leqslant T_\lambda(m^{[k]} + \gamma^{[k]}(\operatorname{sign}(\eta^{[k]}(x_*^{[k]}))M\delta_{x_*^{[k]}} - m^{[k]})).$$

It is noteworthy that other forms were previously used in [10, 12], but, to our knowledge, the update procedure (steps 7 and 8) described in the present paper is new. As we show in theorem 3, optimizing over *both the amplitudes and the positions* is essential to prove the convergence of the algorithm in a finite number of iterations.

Remark 7 (Stopping criterion of the SFW). One may observe that the condition $\mathrm{d}f(m^{[k]})[s^{[k]} - m^{[k]}] = 0$ of algorithm 1 (or equivalently $m^{[k]} \in \operatorname{argmin}_{s \in C} \mathrm{d}f(m^{[k]})[s]$) has been replaced with $|\eta^{[k]}(x_*^{[k]})| \leqslant 1$. In fact the optimality conditions for the non-convex local descent (step 8) at iteration $k - 1$ imply

$$\forall i \in \{1, \dots, N^{[k]}\}, \quad \eta^{[k]}(x_i^{[k]}) = \operatorname{sign}(a_i^{[k]}),$$

whereas $|\eta^{[k]}(x_*^{[k]})| \leqslant 1$ implies $\left\|\eta^{[k]}\right\|_{\infty,X} \leqslant 1$, hence $\eta^{[k]}$ is a valid dual certificate.

With the words of remark 5, step 8 implies that $L\left(M, M\varepsilon_j^{[k]}\delta_{x_j^{[k]}}\right) = 0$ for $1 \leqslant j \leqslant N^{[k]}$, whereas the condition $|\eta^{[k]}(x_*^{[k]})| \leqslant 1$ means $0 = L(0,0) = \min_C L$. As m is a convex combination of those points, we deduce that $(|m^{[k]}|(X), m^{[k]}) \in \operatorname{argmin}_C L$, that is the optimality condition (35).

Remark 8 (Adaptation for the positive BLASSO). In many applications, one is often interested in recovering positive spikes (see for example in section 5). As a result, in these cases it is better to add a positivity constraint $m \geqslant 0$ to the BLASSO. This leads to several changes in algorithm 2

- the stopping condition $|\eta^{[k]}(x_*^{[k]})| \leqslant 1$ becomes $\eta^{[k]}(x_*^{[k]}) \leqslant 1$,
- the LASSO is solved on $\mathbb{R}_+^{N^{[k]}+1}$,
- the optimization problem of step 8 is solved on $\mathbb{R}_+^{N^{[k]}+1} \times X^{N^{[k]}+1}$.

Remark 9 (Implementation details). The SFW algorithm uses three different solvers for respectively steps 3, 7 and 8

- A Newton method, initialized by a grid search, is used to to find the maximum of $|\eta^{[k]}|$ over the compact domain X in step 3. The size of the grid depends on the operator Φ. For example, when Φ is the convolution by the Dirichlet kernel with cutoff frequency f_c, we choose a number of points proportional to f_c.
- The LASSO problem at step 7 is solved using the fast iterative shrinkage thresholding algorithm (FISTA) [4].
- To solve the non-convex optimization problem at step 8, we deploy a bounded BFGS. It allows to enforce the positions x_i to be in the compact domain X and to preserve the sign of the amplitudes a_i. These constraints ensure the differentiability of the objective function which is required by BFGS.

These are the choices that we made in our own implementation of the SFW algorithm. It is possible to use other solvers as long as they provide the same convergence guarantees as required in algorithm 2.

4.2. Study of the convergence of the SFW algorithm

We now study the convergence properties of the Sliding Frank–Wolfe algorithm presented last section (see algorithm 2). Our main result is theorem 3 where one shows that if $m_{a,x} = \sum_{i=1}^N a_i \delta_{x_i}$ is the unique solution of $\mathcal{P}_\lambda(y)$ and if $\eta_\lambda = \frac{1}{\lambda}\Phi^*(y - \Phi m_{a,x})$ is nondegenerate (see equation (42)), then algorithm 2 recovers $m_{a,x}$ in a finite number of iterations. But, first, one shows that our algorithm produces a sequence of measures $(m^{[k]})_{k\in\mathbb{N}}$ that converges

towards m_* (if $m_* \in \mathcal{M}(X)$ is the unique solution of the BLASSO) for the weak-$*$ topology on $\mathcal{M}(X)$.

Proposition 5. *Let $(m^{[k]})_{k\in\mathbb{N}}$ be the sequence obtained from the* Sliding Frank–Wolfe *algorithm. Then it has an accumulation point for the weak-$*$ topology on $\mathcal{M}(X)$, and that point is a solution to ($\mathcal{P}_\lambda(y)$).*

Proof. By remark 6, we know that $(m^{[k]})_{k\in\mathbb{N}}$ is a sequence obtained by applying algorithm 1 to $\tilde{\mathcal{P}}_\lambda(y)$ where the final update is steps 7 and 8 of the SFW. As a result, using lemma 5, one gets that for any m_* solution of $\mathcal{P}_\lambda(y)$,

$$\forall k \in \mathbb{N}, \quad T_\lambda(m^{[k]}) - T_\lambda(m_*) \leqslant \frac{C_1}{k}.$$

Hence $(m^{[k]})$ is a bounded minimizing sequence. One can extract from it a subsequence that converges towards some $m \in \mathcal{M}(X)$ (with $|m|(X) \leqslant M$) for the weak-$*$ topology. Since T_λ is convex and l.s.c., it is also weak-$*$ l.s.c. so that one obtains:

$$T_\lambda(m) = T_\lambda(m_*).$$

Hence m is a solution of $\mathcal{P}_\lambda(y)$. □

From this proposition, one easily deduces the following corollary.

Corollary 1. *If $m_* \in \mathcal{M}(X)$ is the unique solution of $\mathcal{P}_\lambda(y)$ then $(m^{[k]})_{k\in\mathbb{N}}$ weak-$*$ converges towards m_*.*

In fact, under mild assumptions, our algorithm even converges towards the solution of the BLASSO in a finite number of iterations, thanks to the displacement of the spikes over the continuous domain X. For the sake of clarity, we state and prove this theorem in the case of $d = 1$ but the changes for $d \in \mathbb{N}^*$ can be easily done.

Theorem 3. *Suppose that $\varphi \in \mathrm{KER}^{(2)}$, that $m_{a,x} = \sum_{i=1}^N a_i \delta_{x_i}$ is the unique solution of $\mathcal{P}_\lambda(y)$, and that $\eta_\lambda = \frac{1}{\lambda}\Phi^*(y - \Phi m_{a,x})$ is nondegenerate, i.e.*

$$\forall x \in X \setminus \bigcup_{i=1}^N \{x_i\}, \quad |\eta_\lambda(x)| < 1 \quad \text{and} \quad \forall i \in \{1,\dots,N\}, \quad \eta_\lambda''(x_i) \neq 0. \tag{42}$$

Then algorithm 2 recovers $m_{a,x}$ after a finite number of steps (i.e. there exists $k \in \mathbb{N}$ such that $m^{[k]} = m_{a,x}$).

Proof. Since $m_{a,x}$ is the unique solution of $\mathcal{P}_\lambda(y)$, one knows by corollary 1 that the sequence $(m^{[k]})_{k\in\mathbb{N}}$ produced by algorithm 2 converges for the weak-$*$ topology towards $m_{a,x}$.

As Φ is weak-$*$ to weak continuous and by defining $p^{[k]} \overset{\text{def.}}{=} \frac{1}{\lambda}(y - \Phi m^{[k]})$, one gets that $(p^{[k]})_{k\in\mathbb{N}}$ converges towards p_λ in the weak topology of $\mathcal{H}$ and that $\eta^{[k]} \overset{\text{def.}}{=} \Phi^* p^{[k]}$ converges pointwise towards η_λ. Then one can show that Φ^* is a compact operator. Indeed, for any bounded subset $A \in \mathcal{H}$, one can check easily that $\Phi^* A$ is equicontinuous and pointwise relatively compact so that by Ascoli theorem $\Phi^* A$ is relatively compact for the strong topology of $\mathscr{C}_0(X, \mathbb{R})$. As a result one can extract a subsequence of $(\eta^{[k]})_{k\in\mathbb{N}}$ that converges towards η_λ in uniform norm. η_λ is then the unique accumulation point in the uniform norm of the bounded

sequence $(\eta^{[k]})_{k\in\mathbb{N}}$ hence its convergence towards η_λ in uniform norm. One can repeat this argument for $(\eta^{[k]\prime})_{k\in\mathbb{N}}$ and $(\eta^{[k]\prime\prime})_{k\in\mathbb{N}}$ (since $\varphi \in \mathrm{KER}^{(2)}$), obtaining for all $j \in \{0,1,2\}$

$$(\eta^{[k]})^{(j)} \xrightarrow[k\to+\infty]{\|\cdot\|_{\infty,X}} \eta_\lambda^{(j)}. \tag{43}$$

Because η_λ is nondegenerate, there exists a small neighborhood around each x_i on which $\eta_\lambda'' \neq 0$. Hence, we deduce from equation (43) that there exist $\varepsilon > 0$ and $k_1 \in \mathbb{N}$ such that:

$$\forall k \geqslant k_1, \forall i \in \{1,\dots,N\}, \forall x \in]x_i-\varepsilon, x_i+\varepsilon[, \quad \eta^{[k]\prime\prime}(x) \neq 0.$$

We denote in the following

$$I_{x_i,\varepsilon} \overset{\text{def.}}{=}]x_i-\varepsilon, x_i+\varepsilon[, \quad \forall i \in \{1,\dots,N\}.$$

Since $m^{[k]}$ converges towards $m_{a,x}$ in the weak-$*$ topology and $|m_{a,x}|$ does not charge the boundary of $I_{x_i,\varepsilon}$, we have

$$\forall i \in \{1,\dots,N\}, \quad m^{[k]}(I_{x_i,\varepsilon}) \to m_{a,x}(I_{x_i,\varepsilon}) = a_i \neq 0,$$

so that there exists $k_2 \in \mathbb{N}$ such that for all $k \geqslant k_2$, $m^{[k]}$ has at least one spike in each $I_{x_i,\varepsilon}$. In particular $m^{[k]}$ has at least N spikes.

Again, from equation (43), since $(\eta^{[k]})_{k\in\mathbb{N}}$ converges uniformly towards η_λ, one deduces that there exists $k_3 \in \mathbb{N}$ such that for all $k \geqslant k_3$:

$$\mathrm{Sat}^{\pm}(\eta^{[k]}) \subset \left(\mathrm{Sat}^{\pm}(\eta_\lambda)\right) \oplus (]-\varepsilon,\varepsilon[\times\{0\}),$$

where the set of saturation points of a given $\eta \in \mathscr{C}_0(X,\mathbb{R})$ is defined as:

$$\mathrm{Sat}^{\pm}(\eta) \overset{\text{def.}}{=} \{(x,v) \in X \times \{-1,1\};\ \eta(x) = v\}.$$

Moreover,

$$\forall x \in X \setminus \bigcup_{i=1}^{N} I_{x_i,\varepsilon}, \quad |\eta^{[k]}(x)| < 1.$$

In particular for $k \geqslant k_3$, $m^{[k]}$ has no spikes in $X \setminus \bigcup_{i=1}^{N} I_{x_i,\varepsilon}$ because it would contradict the optimality conditions of step 8 of algorithm 2: for all $i \in \{1,\dots,N^{[k]}\}$, $\eta^{[k]}(x_i^{[k]}) = \mathrm{sign}(a_i^{[k]})$.

Suppose now that $k \geqslant \max(k_1,k_2,k_3)$. Then $m^{[k]}$ has at least one spike in each neighborhood of x_i and no spikes outside. Moreover $|\eta^{[k]}| < 1$ outside the neighborhoods and $\eta^{[k]\prime\prime} \neq 0$ inside. Let $i \in \{1,\dots,N\}$ and denote $x_j^{[k]} \in I_{x_i,\varepsilon}$ a position of a spike of $m^{[k]}$. From the optimality conditions of step 8, one has also that $\eta^{[k]\prime}(x_j^{[k]}) = 0$. This combined with $\eta^{[k]\prime\prime} \neq 0$ in $I_{x_i,\varepsilon}$ implies that $|\eta^{[k]}| < 1$, except at $x_j^{[k]}$. Hence, $m^{[k]}$ has exactly one spike in this neighborhood. As a consequence, we proved that $m^{[k]}$ has exactly N spikes (one inside each neighborhood) and:

$$\forall x \in X \setminus \bigcup_{i=1}^{N} \{x_i^{[k]}\}, \quad |\eta^{[k]}(x)| < 1.$$

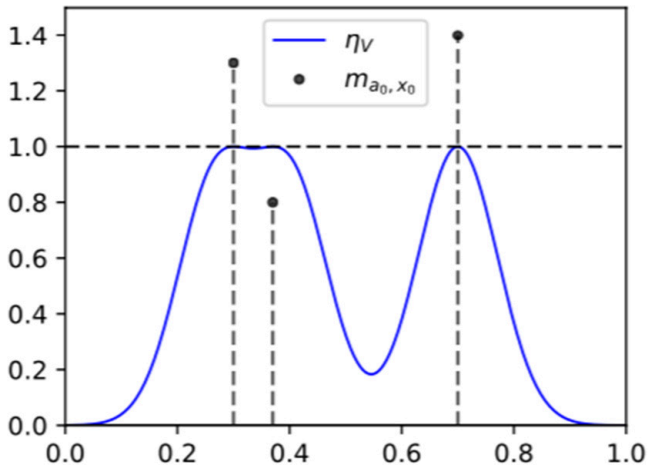

Figure 5. η_V for $m_{a_0,x_0} = 1.3\delta_{0.3} + 0.8\delta_{0.37} + 1.4\delta_{0.7}$.

Hence $m^{[k]}$, composed of N spikes, is a solution of $\mathcal{P}_\lambda(y)$. Since $m_{a,x}$ is supposed to be the unique solution of $\mathcal{P}_\lambda(y)$, one concludes that:

$$m^{[k]} = m_{a,x},$$

i.e. the algorithm recovers $m_{a,x}$ in a finite number of iterations. □

Note that one proved the convergence in a *finite* number of iterations but not exactly N iterations if $m_{a,x}$ is composed of N spikes. However in practice this is exactly what we observe.

4.3. Illustration of the N-steps convergence of the SFW

We now illustrate how the algorithm works and we show that it converges in exactly N iterations in practice (when the noise level and the regularization parameter are appropriate, i.e. $\max(\lambda, \|w\|_{\mathcal{H}}/\lambda)$ is low enough).

We consider $X = [0, 1]$ and a convolution operator with a sampled Gaussian kernel for Φ

$$\Phi : m \in \mathcal{M}(X) \mapsto \int_{[0,1]} \varphi \mathrm{d}m \in \mathbb{R}^K \quad \text{where} \quad \varphi(x) = \left(\frac{1}{\sqrt{2\pi\sigma^2}} \mathrm{e}^{-\frac{(\frac{i-1}{K-1} - x)^2}{2\sigma^2}} \right)_{1 \leqslant i \leqslant K}.$$

We set $\sigma = 0.05$ and $K = 100$. The initial measure used is $m_{a_0,x_0} = 1.3\delta_{0.3} + 0.8\delta_{0.37} + 1.4\delta_{0.7}$ and the noise is small ($w = 10^{-4} w_0$ where $w_0 = \mathrm{randn}(K)$).

Figure 5 shows η_V for this configuration. One can see that it is nondegenerate. Hence, in a small noise now regime, with the appropriate choice of λ, there is a unique measure solution of $\mathcal{P}_\lambda^+(y)$ which is composed of the same number of spikes as m_{a_0,x_0}. Moreover, by theorem 3, the SFW algorithm recovers it in a finite number of iterations.

The decrease of the objective function throughout the algorithm iterations (cumulative iterations of BFGS) is presented in figure 6. As indicated by the two vertical black lines, which show the intermediate iterations, the algorithm converges in exactly 3 iterations. One can observe an important decrease of the objective function each time a spike is added. Also, it is noteworthy that BFGS converges with very few iterations when $k = 0$ and $k = 1$ (first two spikes added) and that the main computational load for the non-convex step occurs for $k = 2$ (more iterations of BFGS).

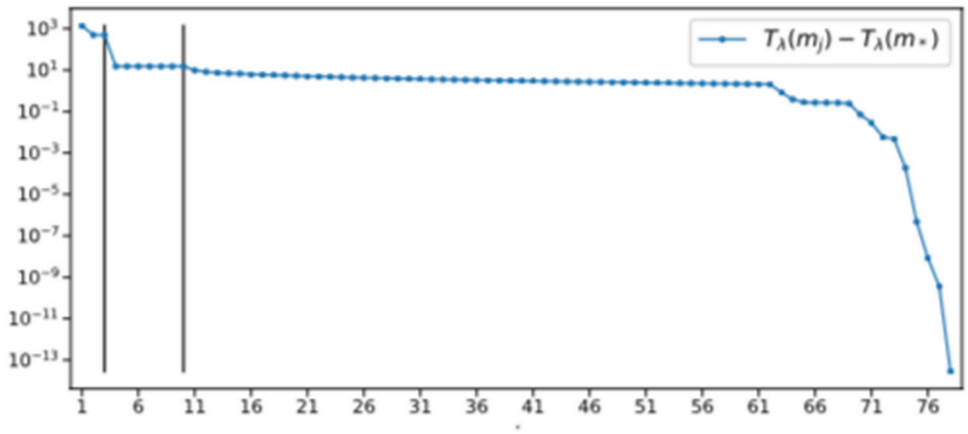

Figure 6. Values of the objective function throughout the SFW algorithm (cumulative iterations of the BFGS). The vertical black lines separate the main outer iterations of the algorithm.

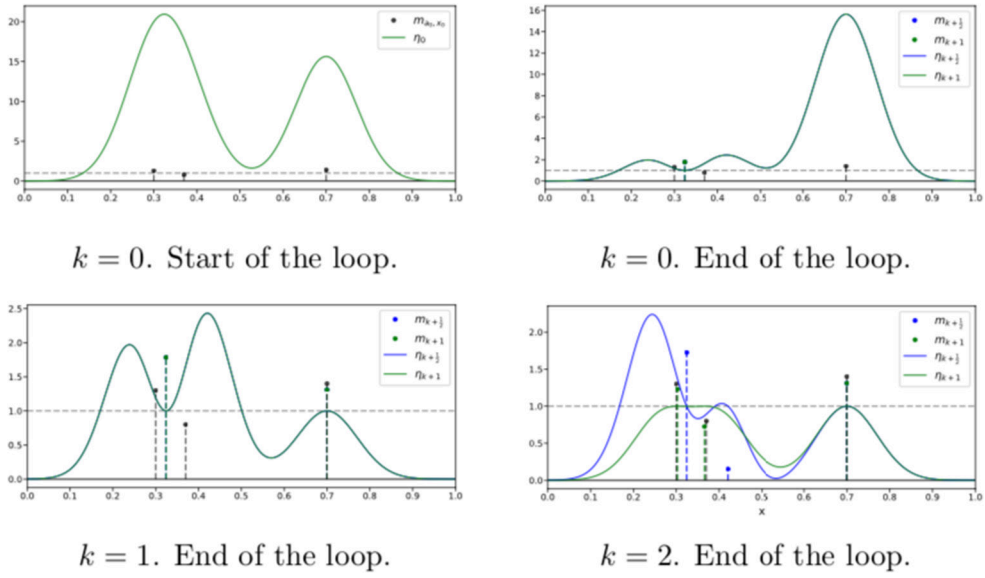

Figure 7. Main steps of the SFW algorithm.

Figure 7 shows $m^{[k]}$ and $\eta^{[k]}$ at different times of the algorithm. More precisely, for $k \in \{0, 1, 2\}$, we display the initial measure m_{a_0,x_0}, the recovered measure, and the associated η. Moreover, we present them after the LASSO step (i.e. $m^{[k+1/2]}$ and $\eta^{[k+1/2]}$) as well as after the BFGS step (i.e. $m^{[k+1]}$ and $\eta^{[k+1]}$) .

One remarks, as expected, that for all i, $\eta^{[k+1/2]}(x_i) = 1$, $\eta^{[k+1]}(x_i) = 1$ and $\eta^{[k+1]\prime}(x_i) = 0$. In the first two main iterations, the spikes are almost not moved by the BFGS. However, at the last iteration, the displacement of the positions and amplitudes of the spikes is crucial to obtain $\eta^{[k+1]} \in \partial|m^{[k+1]}|(X)$, and thus recover the solution of $\mathcal{P}_\lambda^+(y)$ in three steps.

5. Single molecule localization microscopy

The field of fluorescent microscopy has experienced an important revolution during the past two decades with the emergence of super-resolution techniques. These modalities, such as structured illumination microscopy (SIM) [45], stimulated emission depletion (STED) [47], or single molecule localization microscopy (SMLM)—which includes photoactivated localization microscopy (PALM) [5, 50] and stochastic optical reconstruction microscopy (STORM) [74]—bypass the diffraction limit so as to reach unprecedented nanoscale resolution. The main principle behind these methods relies on a combined use of optics and numerical processing, which is commonly called computational imaging. The resolution improvement is thus directly related to the performance of the reconstruction algorithms employed to process the acquired data.

SMLM techniques use photoactivables fluorescent probes to sequentially image a subset of activated molecules. Then, dedicated algorithms are deployed to precisely extract the position of these molecules. While the difficulty of the localization problem increases with the density of activated molecules per acquisitions, low density activations drastically reduce the temporal resolution of the system which makes the method limited for live imaging. Hence, current trends in SMLM concern the development of efficient algorithms dealing with high density data for which classical point-spread function (PSF) fitting or centroid localization methods [48] fail. In particular, off-the-grid sparse regularized methods have shown their efficiency for high density settings [10, 55]. For a complete review and comparisons of existing methods, we refer the reader to the two recent SMLM challenges [75, 76].

Initially introduced for two-dimensional imaging, SMLM has been extended to 3D thanks to PSF engineering. The principle relies on the design of PSFs which vary in the axial direction (i.e. z) in order to encode an information about the depth of molecules. Conventional PSF models include astigmatism [52] and double-helix [70]. An alternative to PSF engineering is to record simultaneously multiple focal planes, as in the biplane modality [58]. It is noteworthy that these two approaches can also be combined as in [54] where the authors use both an astigmatism PSF and multi-focal acquisitions.

In this section, we study the performance of the SFW algorithm on both astigmatism and double-helix modalities with various number of focal planes (typically from 1 to 4). We emphasize that conventional astigmatism and double-helix SMLM devices—in particular commercial ones—use a single focal plane. As opposed to single-focal acquisitions, multi-focal acquisitions require to mount and synchronize several cameras in parallel. To the best of our knowledge, such a setting has only been reported by Huang *et al* [54] for the astigmatism SMLM. Moreover, we propose to compare these two modalities to an alternative approach where depth information is extracted from multi-angle total internal reflection fluorescence (MA-TIRF) microscopy acquisitions. Such an approach has never been reported yet and we expect our numerical simulations to serve as a proof of concept for further developments. One of the main interest in combining SMLM with MA-TIRF is that classical PSFs, which are better localized laterally than astigmatism or double-helix, can be used. This would reduce the difficulty of lateral molecule localization for high density settings while recovering the depth through the MA-TIRF acquisitions.

5.1. Forward operators

In this section, we define the forward operator Φ for the three modalities considered in this paper. The first two correspond to conventional three-dimensional SMLM with astigmatism

or double-helix PSFs. The third one, on the contrary, uses a MA-TIRF excitation in order to get an information about the depth of molecules. The operator $\Phi : \mathcal{M}(X) \to \mathbb{R}^{N_1 N_2 K}$ maps the Radon measures $m \in \mathcal{M}(X)$ to the discrete noiseless measurements $\Phi m \in \mathbb{R}^{N_1 N_2 K}$,

$$\Phi m = \int_X \varphi(x) \mathrm{d}m(x). \tag{44}$$

It is fully characterized by the function $\varphi : X \to \mathbb{R}^{N_1 N_2 K}$. Hence, for each modality, we only have to define φ. In the following, $X \overset{\text{def.}}{=} [0, b_1] \times [0, b_2] \times [0, b_3]$ is a subset of $\mathbb{R}^3$, and we write $x = (x_1, x_2, x_3) \in X$. Then, we consider a camera containing $N_1 \times N_2$ pixels and we denote the center of the ith pixel by $(c_{i,1}, c_{i,2})$. Finally, we provide expressions of φ which enclose the integration over camera pixels

$$\Omega_i \overset{\text{def.}}{=} (c_{i,1}, c_{i,2}) + \left[-\frac{b_1}{2N_1}, \frac{b_1}{2N_1} \right] \times \left[-\frac{b_2}{2N_2}, \frac{b_2}{2N_2} \right] \subset \Omega \overset{\text{def.}}{=} [0, b_1] \times [0, b_2].$$

5.1.1. Astigmatism model. This modality provides depth information using an astigmatism deformation of the PSF with respect to the axial direction z. It is customary to model the latter with a Gaussian function whose variances σ_1 and σ_2 vary with z according to [55, 60]

$$\sigma_1(z) \overset{\text{def.}}{=} \sigma_0 \sqrt{1 + \left(\frac{\alpha z - \beta}{d} \right)^2} \quad \text{and} \quad \sigma_2(z) \overset{\text{def.}}{=} \sigma_1(-z). \tag{45}$$

The constants involved in (45) can be calibrated from real data [52, 60]. Then, integrating this Gaussian model over camera pixels, we have for all $i \in \{1, \dots, N_1 N_2\}$ and $k \in \{1, \dots, K\}$

$$[\varphi(x)]_{i,k} \overset{\text{def.}}{=} \frac{1}{2\pi \sigma_1(x_3 - z_k)\sigma_2(x_3 - z_k)} \int_{\Omega_i} \mathrm{e}^{-\left(\frac{(x_1 - s_1)^2}{2\sigma_1^2(x_3 - z_k)} + \frac{(x_2 - s_2)^2}{2\sigma_2^2(x_3 - z_k)} \right)} \mathrm{d}s_1 \mathrm{d}s_2,$$

where $(z_k)_{k=1}^K$ are the positions of the considered focal planes.

5.1.2. Double-helix model. Here, depth information is obtained by using a PSF formed out of two lobes which coil around each other along z to form a double-helix shape. In this paper, we model these lobes by two Gaussian functions with fixed variances $\sigma_1 = \sigma_2$, and with a center whose lateral position (r_1, r_2) (respectively, $(-r_1, -r_2)$) varies with z according to

$$r_1(z) \overset{\text{def.}}{=} \frac{\omega}{2} \cos(\theta(z)) \text{ and } r_2(z) \overset{\text{def.}}{=} -\frac{\omega}{2} \sin(\theta(z)) \text{ where } \theta(z) = \theta_{\text{speed}} z. \tag{46}$$

Parameters $\omega > 0$ and $\theta_{\text{speed}} > 0$ correspond to the distance between the two Gaussian and the rotation speed of the double-helix (rad/nm), respectively. Then, integrating this model over camera pixels, we have for all $i \in \{1, \dots, N_1 N_2\}$ and $k \in \{1, \dots, K\}$

$$[\varphi(x)]_{i,k} \overset{\text{def.}}{=} \frac{1}{2\pi\sigma_1\sigma_2} \sum_{u \in \{-1,1\}} \int_{\Omega_i} \mathrm{e}^{-\left(\frac{(x_1 + u r_1 (x_3 - z_k) - s_1)^2}{2\sigma_1^2} + \frac{(x_2 + u r_2 (x_3 - z_k) - s_2)^2}{2\sigma_2^2} \right)} \mathrm{d}s_1 \mathrm{d}s_2,$$

where $(z_k)_{k=1}^K$ are the positions of the considered focal planes.

5.1.3. MA-TIRF model. With this modality, each activated set of molecules is imaged using $K \in \mathbb{N}$ TIRF illuminations with incident angles $(\alpha_k)_{k=1}^K$. Let $n_i > 0$ and $n_t > 0$ be the refractive

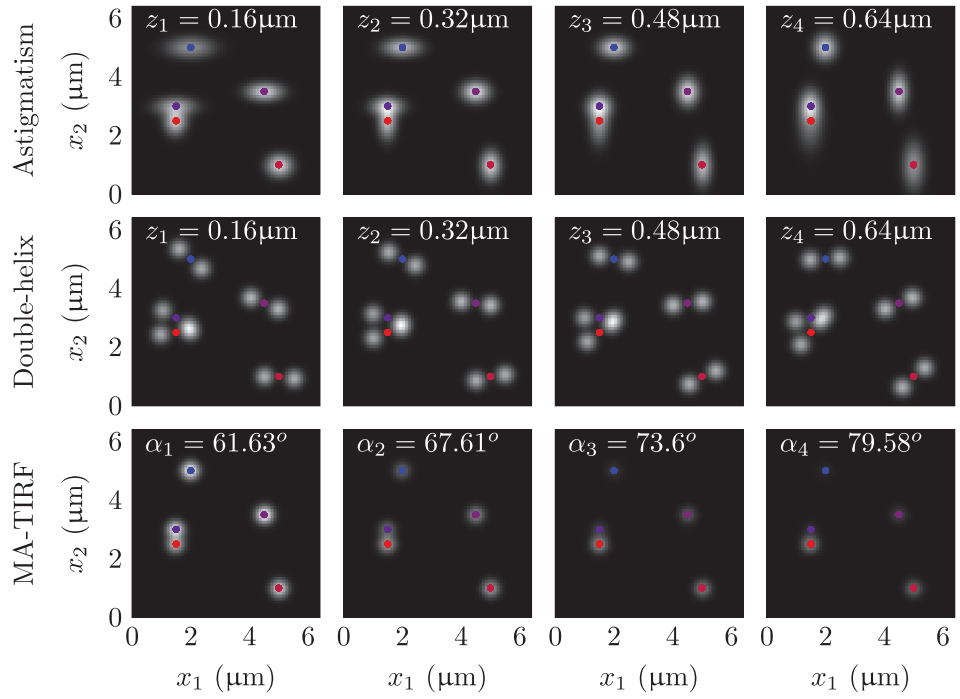

Figure 8. Noiseless acquisitions y_0 for the measure m_{a_0,x_0} given in (48) and $K = 4$. The parameters used for these simulations are given in table 1. The color of the molecules represent their depths: 0 (red)—0.8 μm (blue).

indices of the incident (i.e. glass coverslip) and the transmitted (i.e. sample) medium, respectively. A TIRF excitation is obtained when the incident angle α is greater than the critical angle $\alpha_c = \arcsin(n_t/n_i)$ for which we have total internal reflection of the light within the incident medium. This phenomenon produces an evanescent wave which decays in the transmitted medium as $\exp(-sx_3)$, where $s = (4\pi n_i)/\lambda_\ell \left(\sin^2(\alpha) - \sin^2(\alpha_c)\right)$ is the penetration depth and λ_ℓ is the wavelength of the incident laser beam [1, 2]. Because the decay of this evanescent excitation vary with the incident angle, the depth of biological structures can be recovered with a nanometric precision from multi-angle acquisitions [9, 31, 89]. Combining this principle with SMLM techniques lead to a forward model Φ defined, for all $i \in \{1, \dots, N_1 N_2\}$ and $k \in \{1, \dots, K\}$, by

$$[\varphi(x)]_{i,k} \overset{\text{def.}}{=} \frac{\xi(x_3)\mathrm{e}^{-s_k x_3}}{2\pi\sigma_1\sigma_2} \int_{\Omega_i} \mathrm{e}^{-\left(\frac{(x_1-s_1)^2}{2\sigma_1^2} + \frac{(x_2-s_2)^2}{2\sigma_2^2}\right)} \mathrm{d}s_1 \mathrm{d}s_2, \tag{47}$$

where $\xi(z) = \left(\sum_{k=1}^K \mathrm{e}^{-2s_k z}\right)^{-1/2}$. This model comes from the combination of a lateral convolution with the axial TIRF excitation. Here the PSF of the system is assumed to be a Gaussian with variances $\sigma_1 = \sigma_2$, and to be constant along x_3 (because only a thin layer of few hundred nanometers is excited by the evanescent wave). The values $(s_k)_{k=1}^K$ correspond to the penetration depths associated to the incident angles $(\alpha_k)_{k=1}^K$.

Remark 10. One particularity of the MA-TIRF modality is that the kernel φ in (47) is separable. This can be exploited numerically to reduce the overall algorithm complexity.

Table 1. Parameters used for data simulation.

	Parameter	Value	Description
All modalities	$b_1 = b_2$	6.4 μm	Region of interest
	b_3	0.8 μm	Maximal depth of molecules
	$N_1 = N_2$	64	Detector grid size
	NA	1.49	Objective numerical aperture
	n_i	1.515	Refractive index incident medium
	n_t	1.333	Refractive index transmitted medium
	λ_ℓ	0.66 μm	Excitation wavelength
	n_{photon}	1000	Photon budget
	σ	10^{-4}	Variance of Gaussian noise
Astigmatism	σ_0	$0.42\lambda_\ell/\text{NA}$	PSF variance at focus
	β	0.2 μm	Depth for which the variance is minimal
	d	$\lambda_\ell n_i/(2\text{NA}^2)$	Parameter related to the depth-of-field
	α	-0.79	Scaling constant
	$(z_k)_{k=1}^K$	$kb_3/(K+1)$	Focal planes
Double-helix	$\sigma_1 = \sigma_2$	$0.42\lambda_\ell/\text{NA}$	PSF variance
	ω	1 μm	Distance between the two PSF lobes
	θ_{speed}	0.3846π rad μm^{-1}	Rotation speed of the PSF
	$(z_k)_{k=1}^K$	$kb_3/(K+1)$	Focal planes
MA-TIRF	$\sigma_1 = \sigma_2$	$0.42\lambda_\ell/\text{NA}$	PSF variance
	$(\alpha_k)_{k=1}^K$	$\alpha_c + \frac{\alpha_{\max}-\alpha_c}{K-1}(k-1)$	Incident angles
	$\alpha_{\max}$	$\sin^{-1}(\text{NA}/n_i)$	Maximal incident angle

5.1.4. Illustrations and numerical computation of η_V. Examples of noiseless measurements $y_0 = \Phi m_{a_0,x_0}$ with

$$m_{a_0,x_0} = \delta_{(1.5,2.5,0.1)} + \delta_{(1.5,3,0.5)} + \delta_{(2,5,0.7)} + \delta_{(4.5,3.5,0.4)} + \delta_{(5,1,0.2)} \tag{48}$$

are presented in figure 8 for the three modalities. The parameters used for these simulations are provided in table 1. One can observe the effect of the three modalities on molecules at different depths. For the astigmatism modality, the orientation along which the PSF is defocuced indicates the position of the molecule with respect to the focal plane (above/below). Moreover, the larger is this defocucing, the deeper is the molecule. In the case of the double-helix modality, we can clearly see the rotation of the PSF with depth. Finally, for the MA-TIRF modality, we can observe that the recorded intensities for deep molecules decrease, with the incident angle, faster than the intensity for molecules which are close to the glass coverslip (i.e. $x_3 = 0$).

Although, for these three-dimensional models, an explicit expression of η_V seems challenging to come by, the latter can be computed numerically for specific points $x \in X$. A representation of η_V for the measure given in (48) at $x_3 = 0.1$ and $x_3 = 0.5$ is depicted in figure 9. For the three modalities, we have that $\eta_V(1.5, 2.5, 0.1) = \eta_V(1.5, 3, 0.5) = 1$ and otherwise η_V is smaller than 1. Hence, η_V seems nondegenerate and a measure composed of the same number of Dirac masses as m_{a_0,x_0} can be recovered by the SFW algorithm.

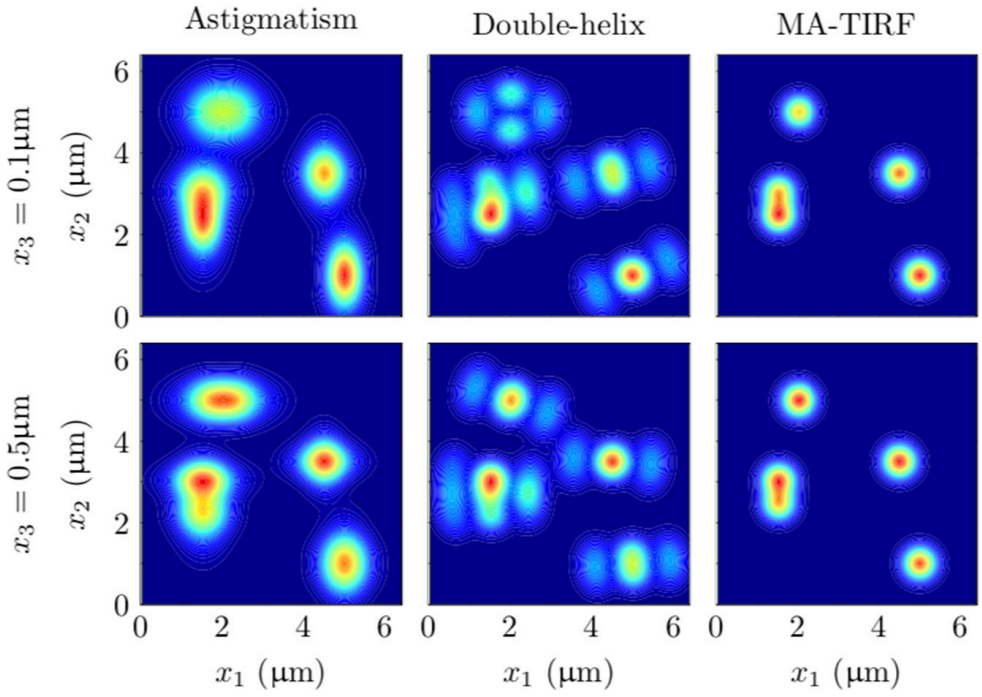

Figure 9. Numerical computation of η_V at $x_3 = 0.1$ (top) and $x_3 = 0.5$ (bottom) for the three models and the measure m_{a_0,x_0} given in (48). The colormap ranges from 0 (blue) to 1 (red).

5.2. Simulation setting

5.2.1. Imaged structure. Simulations were performed using the microtubules-like structure depicted in figure 10. It has been generated within the volume

$$X = [0, b_1] \times [0, b_2] \times [0, b_3] \subset \mathbb{R}^3 \quad \text{where} \quad b_1 = b_2 = 6.4\ \mu\text{m and } b_3 = 0.8\ \mu\text{m}. \tag{49}$$

The filaments were obtained by randomly sampling many points along four curves defined by polynomial equations. To ensure a uniform distribution of the points along the curves, we first parametrized each curve by a piecewise linear function (with very small steps). Then, in order to give a width to the filaments, each point $x \in X$ randomly chosen on one of the curves is replaced by a point randomly chosen in a ball centered at x with radius 10 nm. Thus, simulated filaments have a diameter of 20 nm.

5.2.2. Simulation of noiseless acquisitions. The $N_{\text{tot}} \in \mathbb{N}^*$ molecules of the simulated structure are divided into $n \in \mathbb{N}^*$ sparse set of $N \in \mathbb{N}^*$ molecules using a random permutation (i.e. $N_{\text{tot}} = n \times N$). This models the sequential stochastic activation of fluorophores used in SMLM. For each of the n subsets of molecules, we define a Radon measure composed of a sum of Dirac masses—located at the position of the molecules—with positive amplitudes

$$m_{a_0,x_0} = \sum_{i=1}^{N} a_{0,i}\delta_{x_{0,i}} \quad \text{where} \quad a_{0,i} > 0 \quad \text{and} \quad x_{0,i} \in X.$$

The amplitudes are randomly generated within $[1, 1.5]$. An example of a set of activated molecules is shown in figure 10 (black crosses). Now let $(N_1 \times N_2)$ be the size of the grid of pixels

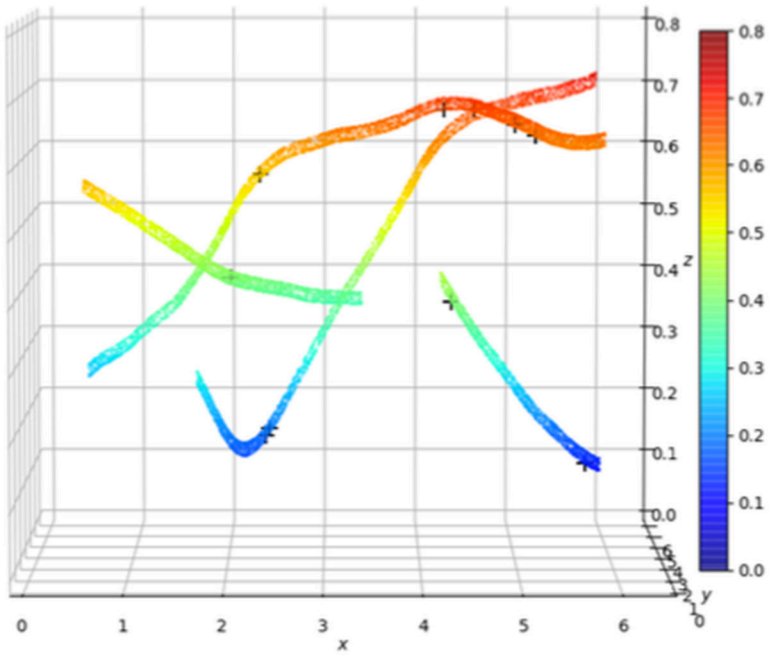

Figure 10. Microtubules structure used for the simulations. The diameter of the filaments is 20 nm. The color encodes the depth of molecules within the range 0–0.8 μm. Black crosses represent a subset of activated molecules (i.e. a measure m_{a_0,x_0}).

on the detector plane, and K be the number of focal planes (or the number of TIRF 'angles', see section 5.1) which are recorded. Then, the noiseless measurements y_0 for an activated measure m_{a_0,x_0} follow the model

$$y_0 = \Phi m_{a_0,x_0}, \tag{50}$$

where Φ is defined in (44).

Finally, it is noteworthy that in practice the number of activated molecules varies from one activation to another around an average value (which depends on the power of the excitation laser beam). However, fixing this number to N for each activated set of molecules allows us to better control the density of spikes in order to study the behaviour of the algorithm when the latter increases.

5.2.3. Noise model. There are two predominant sources of noise in microscopy data.

- The shot noise which is inherent to the quantum nature of light (random emissions of photons). It is well modeled by a Poisson distribution whose intensity is the number of photon collected at each pixel. Given the noiseless acquisition y_0, we normalize it such that

$$\max_{i\in\{1,\dots,N_1N_2\}} \left(\sum_{k=1}^{K} [y_0]_{i,k} \right) = n_{\text{photon}}, \tag{51}$$

  where $n_{\text{photon}} > 0$ denotes the maximal photon budget per pixel and controls the noise level. Then, each entry of y_0 is replaced by a realization of a Poisson distribution $\mathcal{P}$ with

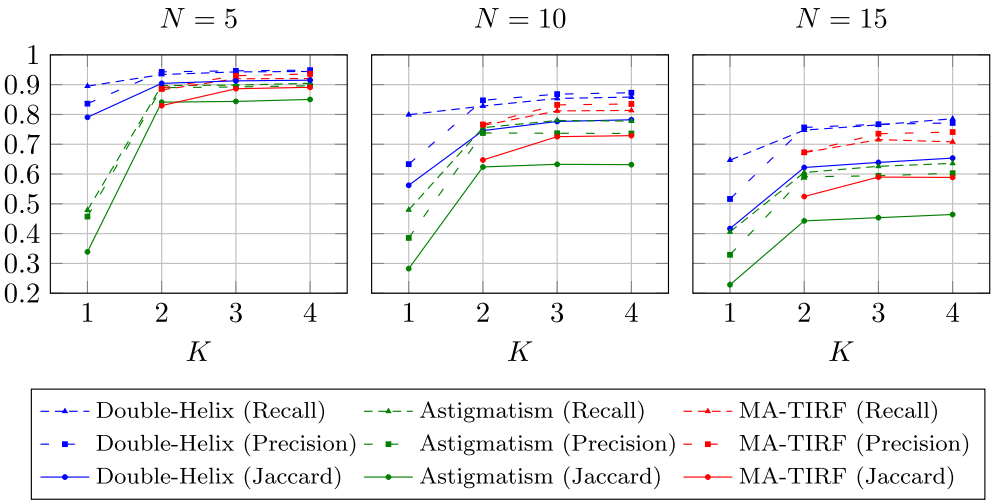

Figure 11. Evolution of Jaccard, Recall and Precision metrics with respect to K, for a radius of detection $r = 0.02$ (20 nm).

parameter $[y_0]_{i,k}$. It is noteworthy from (51) that the level of noise not only increases as n_{photon} decreases, but it also increases with K.

- The readout noise w_G of the camera. It is usually modeled by a Gaussian distribution with variance σ^2.

Finally the noisy data are given by

$$y = \mathcal{P}(y_0) + w_G. \tag{52}$$

5.3. Results

For each of the three modalities presented in section 5.1 (double-helix, astigmatism, MA-TIRF), acquisitions where simulated using the optical parameters gathered in table 1. These parameters have been tuned according to the experimental PSF used in the SMLM challenge [76]. Finally, we generated different experiments by varying the density of molecules $N \in \{5, 10, 15\}$ as well as the number of focal planes (or angles for the TIRF model) $K \in \{1, 2, 3, 4\}$.

5.3.1. Metrics for evaluation. In order to assess the quality of the reconstructed volumes, we consider standard metrics which reflect both the detection rate and the localization error [75, 76]. Given a recovered frame and a tolerance radius $r > 0$, we pair estimated molecules and ground truth (GT) molecules when the distance between them is lower than r. Paired estimated molecules are then referred as true positive (TP) while unpaired ones as false positive (FP). Finally, the unpaired GT molecules are identified as false negative (FN). These quantities being determined for each frame, we can compute the Jaccard index (Jac), the Recall (Rec) and the Precision (Pre) metrics,

$$\text{Jac} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP} + \#\text{FN}} \quad \text{Rec} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FN}} \quad \text{Pre} = \frac{\#\text{TP}}{\#\text{TP} + \#\text{FP}}. \tag{53}$$

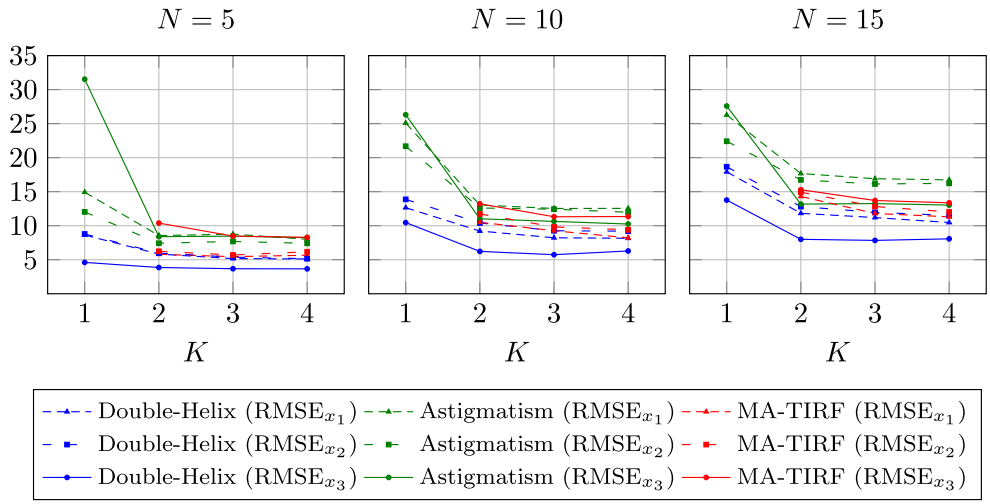

Figure 12. Evolution of the RMSE (nm) with respect to K, for a radius of detection $r = 0.1$ (100 nm).

The Jaccard index measures the overall performance of detection by giving a measure of similarity between the two sets of points. The Recall and Precision metrics can then be used to measure the ability of an algorithm to minimize FN and FP detection, respectively. Finally, the TP molecules are used to compute the root mean squared error (RMSE) along each dimension

$$\mathrm{RMSE}_{x_1} = \sqrt{\frac{1}{\#\mathrm{TP}} \sum_{i \in \mathrm{TP}} ([x_i]_1 - [x_{0,i}]_1)^2}, \tag{54}$$

and similarly for RMSE_{x_2} and RMSE_{x_3}. Note that, by construction, the RMSE is bounded by the radius r. Hence, in the following, we use different values for r depending on the metric of interest.

5.3.2. Choice of the regularization parameter λ. For each experiment (i.e. $N \in \{5, 10, 15\}$ and $K \in \{1, 2, 3, 4\}$), we choose the value of the regularization parameter λ which maximizes the Jaccard index for a radius of $r = 0.02$ (i.e. 20 nm). This training step was performed over a small subset of initial measures m_{a_0,x_0} (i.e. frames). Then the recovery was done on the complete dataset using the optimal λ found.

5.3.3. Discussion. The evolution of Jaccard, Recall, and Precision metrics with respect to K are depicted in figure 11. As expected, they all increase with K. However, although the improvement is significant from $K = 1$ to $K = 2$, higher values only provide marginal gains. This can be explained by the fact that the photon budget n_{photon} is distributed over the K acquisitions (see equation (51)). Hence, the additional axial information brought by increasing the number of acquisitions per activation should be balanced by the higher noise corrupting the data. Another observation from these plots concerns the degradation of the performance as the density (i.e. the number of molecules N) increases.

These results also bring useful information in order to improve existing systems. Let us recall that current commercial systems includes Astigmatism and double-helix modalities with one focal plane (i.e. $K = 1$). Hence, it can be inferred from our simulations that recording

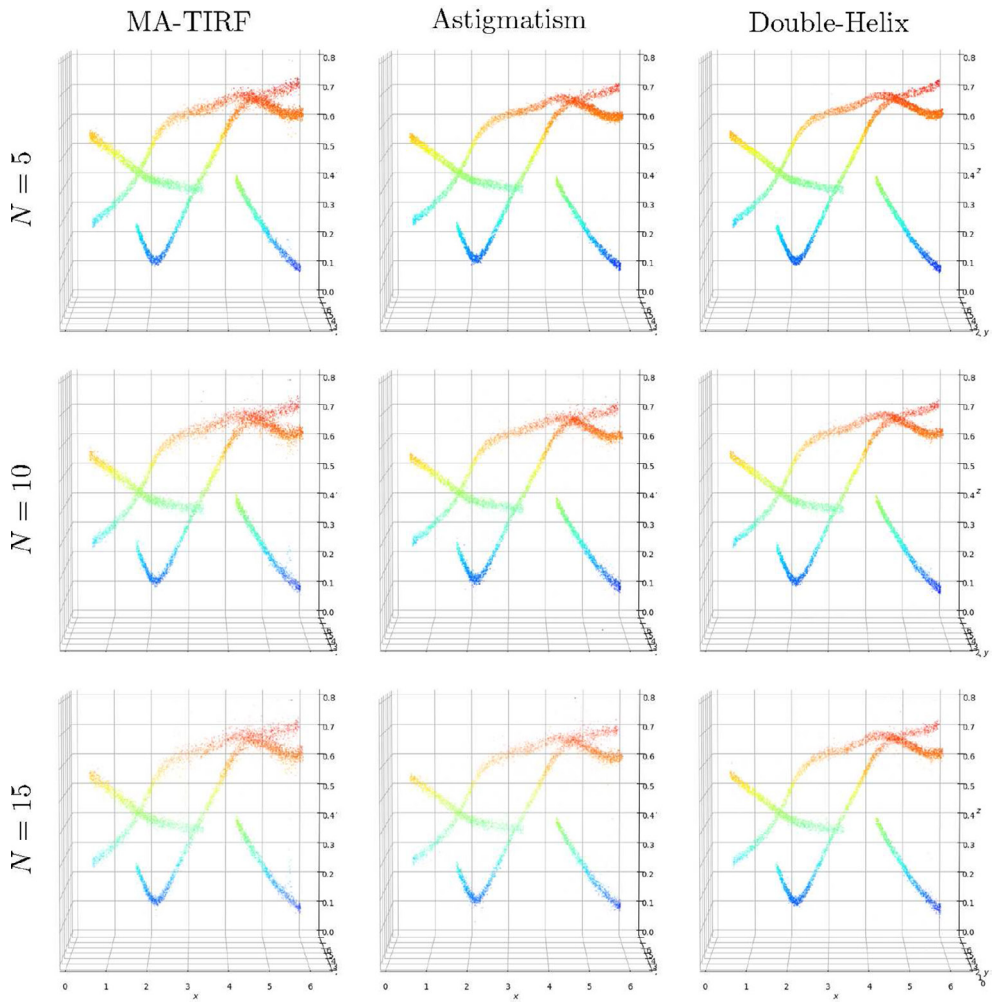

Figure 13. Recovered structures for $K = 4$.

an image at two focal planes for each activation of molecules would not only improve significantly the reconstruction quality but make the reconstructions more robust when the density of molecules increases. These observations corroborate the study in [54] where the authors use a multi-focus astigmatism system. However, to preserve a reasonable temporal resolution, multi-focal acquisitions require to synchronize several cameras [54] which can be expensive and lead to delicate calibration procedures (e.g. alignment and PSF aberrations for each camera). In that respect, the proposed combination of SMLM with MA-TIRF offers an interesting alternative to improve existing systems. First, it has the potential to provide reconstructions whose quality compares favorably with the double-helix model while improving over the Astigmatism modality. Second, it only requires the use of galvanometric mirrors to control the incident angle [9]. It is noteworthy that commercial SMLM systems generally use a single TIRF illumination to limit the illumination depth. Finally, as for the multi-focus strategy, MA-TIRF requires some calibrations (e.g. incident angles) for which there exist dedicated procedures [9, 78].

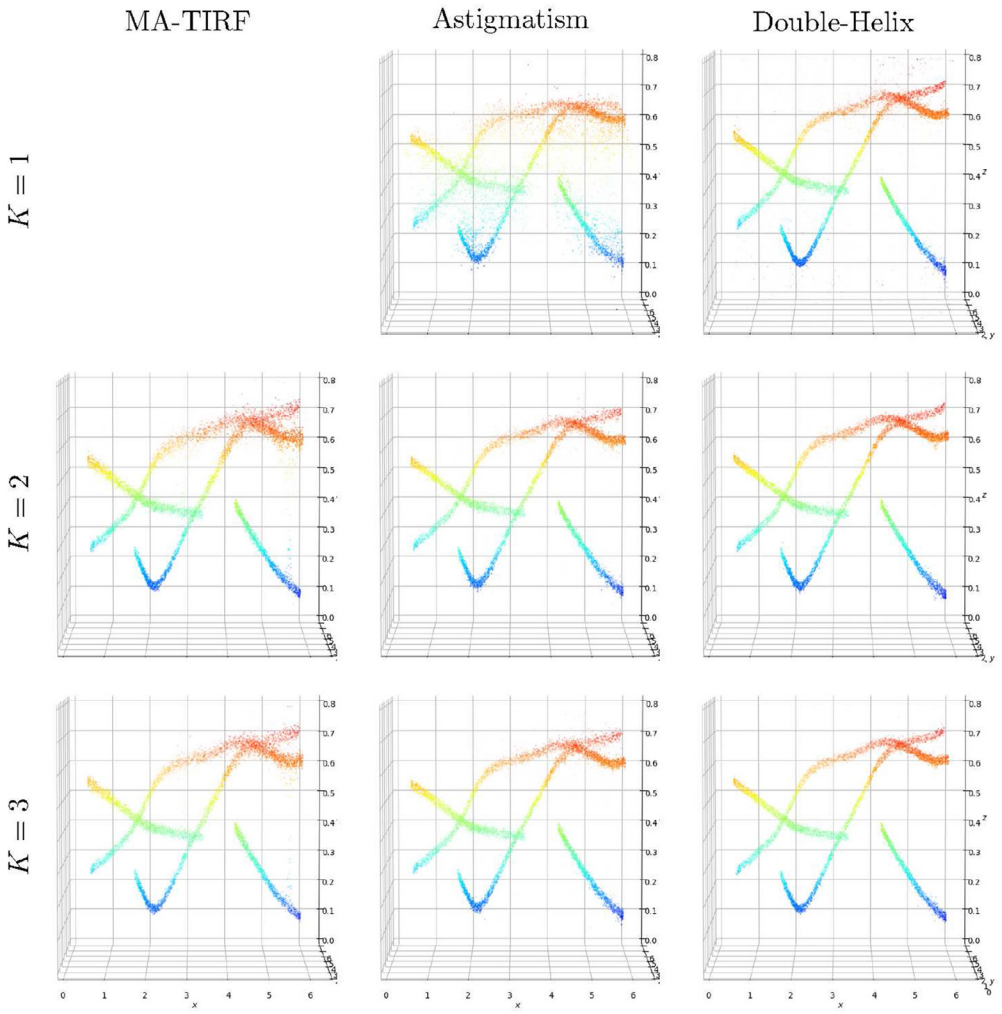

Figure 14. Recovered structures for $N = 10$.

Remark 11. Although the PSFs used for these simulations have been adjusted using experimental PSFs, they remain idealistic. This is particularly the case for the double-helix which in practice deviates from two Gaussian lobes that coil around each other along z [76]. In contrast, the Gaussian model yields a precise approximation of the MA-TIRF (i.e. widefield) PSF [88]. The main simplification for the latter lies in the fact that each molecule is activated only during one set of multi-angle acquisitions. This would not be the case with a real implementation of the system and the model should be improved by considering the temporal aspect of the acquisition. However, the present study constitutes a first proof-of-concept and future developments will consider a more sophisticated model.

The results in terms of RMSE presented in figure 12 lead to similar interpretations. First, the detection accuracy is increasing with K while decreasing with N. Second, we can observe that the differences between the double-helix and the MA-TIRF models mainly come from the

precision in x_3. Indeed, they both lead to the same lateral RMSE (around 5 nm when $N = 5$ and 12 nm at the highest density $N = 15$), but the double-helix enjoys a better axial RMSE. This reflects the challenging problem that constitues the inversion of the Laplace transform, which is related to the MA-TIRF model. Nevertheless, the SWF algorithm performs quite well at this task (see also figures 13 and 14). Another observation concerns the fact that the double-helix can reach a better axial than lateral RMSE. This fact, which was also observed in the recent SMLM challenge [76], can be explained by the large lateral support of the double-helix PSF as well as its good axial discrimination.

Finally, three-dimensional representations of the recovered structures are presented in figures 13 and 14 for a fixed $K = 4$ and $N = 10$, respectively. These figures complete and illustrate the observations made with the computed metrics.

6. Conclusion

This paper demonstrated from both theoretical and practical perspectives the Sliding Frank–Wolfe algorithm, in particular when facing a challenging non-translation invariant operator such as the Laplace kernels. Such operators lead to difficulties in estimating the spikes positions which is efficiently addressed by non-convex update step of the grid location. The BLASSO method, coupled with this Sliding Frank–Wolfe solver, is well adapted to these non-convolutive operators because it does not rely on spectral (Fourier) methods and can be analyzed theoretically through the prism of convex duality and vanishing certificates.

Acknowledgment

The authors would like to thank Laure Blanc-Féraud for initiating this collaboration and for stimulating discussions. The work of Gabriel Peyré has been supported by the European Research Council (ERC project NORIA). The work of Emmanuel Soubies has been supported by the European Research Council (ERC project GlobalBioIm).

ORCID iDs

Vincent Duval https://orcid.org/0000-0002-7709-256X
Gabriel Peyré https://orcid.org/0000-0002-4477-0387

References

[1] Axelrod D 1981 Cell-substrate contacts illuminated by total internal reflection fluorescence *J. Cell Biol.* **89** 141–5
[2] Axelrod D 2008 Total internal reflection fluorescence microscopy *Methods Cell Biol.* **89** 169–221
[3] Azaïs J M, de Castro Y and Gamboa F 2015 Spike detection from inaccurate samplings *Appl. Comput. Harmon. Anal.* **38** 177–95
[4] Beck A and Teboulle M 2009 A fast iterative shrinkage-thresholding algorithm for linear inverse problems *SIAM J. Imaging Sci.* **2** 183–202
[5] Betzig E, Patterson G H, Sougrat R, Lindwasser O W, Olenych S, Bonifacino J S, Davidson M W, Lippincott-Schwartz J and Hess H F 2006 Imaging intracellular fluorescent proteins at nanometer resolution *Science* **313** 1642–5
[6] Bhaskar B N, Tang G and Recht B 2013 Atomic norm denoising with applications to line spectral estimation *IEEE Trans. Signal Process.* **61** 5987–99

[7] Blumensath T and Davies M E 2008 Iterative thresholding for sparse approximations *J. Fourier Anal. Appl.* **14** 629–54
[8] Blumensath T and Davies M E 2009 Iterative hard thresholding for compressed sensing *Appl. Comput. Harmon. Anal.* **27** 265–74
[9] Boulanger J, Gueudry C, Münch D, Cinquin B, Paul-Gilloteaux P, Bardin S, Guérin C, Senger F, Blanchoin L and Salamero J 2014 Fast high-resolution 3d total internal reflection fluorescence microscopy by incidence angle scanning and azimuthal averaging *Proc. Natl Acad. Sci.* **111** 17164–9
[10] Boyd N, Schiebinger G and Recht B 2017 The alternating descent conditional gradient method for sparse inverse problems *SIAM J. Optim.* **27** 616–39
[11] Boyd S and Vandenberghe L 2004 *Convex Optimization* (Cambridge: Cambridge University Press)
[12] Bredies K and Pikkarainen H K 2013 Inverse problems in spaces of measures *ESAIM Control Optim. Calc. Var.* **19** 190–218
[13] Cadzow J A 1988 Signal enhancement-a composite property mapping algorithm *IEEE Trans. Acoust. Speech Signal Process.* **36** 49–62
[14] Candès E J and Fernandez-Granda C 2013 Super-resolution from noisy data *J. Fourier Anal. Appl.* **19** 1229–54
[15] Candès E J and Fernandez-Granda C 2014 Towards a mathematical theory of super-resolution *Commun. Pure Appl. Math.* **67** 906–56
[16] Catala P, Duval V and Peyré G 2017 A low-rank approach to off-the-grid sparse deconvolution (to appear in *SIAM J. Imaging Sci.*)
[17] Chen S S, Donoho D L and Saunders M A 1998 Atomic decomposition by basis pursuit *SIAM J. Sci. Comput.* **20** 33–61
[18] Combettes P L and Wajs V R 2005 Signal recovery by proximal forward-backward splitting *Multiscale Model. Simul.* **4** 1168–200
[19] Condat L and Hirabayashi A 2015 Cadzow denoising upgraded: a new projection method for the recovery of Dirac pulses from noisy linear measurements *Sampl. Theory Signal Image Process.* **14** 17–47
[20] Da Costa M F and Dai W 2018 A tight converse to the spectral resolution limit via convex programming *IEEE International Symposium on Information Theory (ISIT)* 901–5
[21] Daubechies I, Defrise M and De Mol C 2004 An iterative thresholding algorithm for linear inverse problems with a sparsity constraint *Commun. Pure Appl. Math.* **57** 1413–57
[22] de Castro Y and Gamboa F 2012 Exact reconstruction using Beurling minimal extrapolation *J. Math. Anal. Appl.* **395** 336–54
[23] De Castro Y, Gamboa F, Henrion D and Lasserre J B 2017 Exact solutions to super resolution on semi-algebraic domains in higher dimensions *IEEE Trans. Inform. Theory* **63** 621–30
[24] Hauer J F, Demeure C J and Scharf L L 1990 Initial results in Prony analysis of power system response signals *IEEE Trans. Power Syst.* **5** 80–9
[25] Demanet L and Nguyen N 2015 The recoverability limit for superresolution via sparsity (arXiv:1502.01385)
[26] Demyanov V F and Rubinov A M 1970 *Approximate Methods in Optimization Problems* vol 32 (Amsterdam: Elsevier)
[27] Den Dekker A and Van den Bos A 1997 Resolution: a survey *J. Opt. Soc. Am.* A **14** 547–57
[28] Denoyelle Q, Duval V and Peyré G 2017 Support recovery for sparse super-resolution of positive measures *J. Fourier Anal. Appl.* **23** 1153–94
[29] Donoho D L 1992 Super-resolution via sparsity constraints *SIAM J. Math. Anal.* **23** 1309–31
[30] Donoho D L and Johnstone I M 1995 Adapting to unknown smoothness via wavelet shrinkage *J. Am. Stat. Assoc.* **90** 1200–24
[31] Dos Santos M C, Déturche R, Vézy C and Jaffiol R 2016 Topography of cells revealed by variable-angle total internal reflection fluorescence microscopy *Biophys. J.* **111** 1316–27
[32] Duval V 2019 A characterization of the non-degenerate source condition in super-resolution *Inf. Inference: J. IMA* (https://doi.org/10.1093/imaiai/iaz002)
[33] Duval V and Peyré G 2015 Exact support recovery for sparse spikes deconvolution *Found. Comput. Math.* **15** 1315–55
[34] Duval V and Peyré G 2017 Sparse spikes super-resolution on thin grids I: the Lasso *Inverse Problems* **33** 055008
[35] Duval V and Peyré G 2017 Sparse spikes super-resolution on thin grids II: the continuous basis pursuit *Inverse Problems* **33** 095008

[36] Efron B, Hastie T, Johnstone I and Tibshirani R 2004 Least angle regression *Ann. Stat.* **32** 407–99 (with discussion, and a rejoinder by the authors)
[37] Eftekhari A and Thompson A 2018 A bridge between past and present: Exchange and conditional gradient methods are equivalent (arXiv:1804.10243)
[38] Eftekhari A and Wakin M B 2015 Greed is super: a fast algorithm for super-resolution (arXiv:1511.03385)
[39] El Ghaoui L, Viallon V and Rabbani T 2012 Safe feature elimination in sparse supervised learning *Pac. J. Optim.* **8** 667–98
[40] Fernandez-Granda C 2013 Support detection in super-resolution *Proc. of the 10th Int. Conf. on Sampling Theory and Applications* pp 145–8
[41] Fernandez-Granda C 2016 Super-resolution of point sources via convex programming *Inf. Inference* **5** 251–303
[42] Figueiredo M A T and Nowak R D 2003 An EM algorithm for wavelet-based image restoration *IEEE Trans. Image Process.* **12** 906–16
[43] Flinth A and Weiss P 2018 Exact solutions of infinite dimensional total-variation regularized problems *Inf. Inference: J. IMA* (https://doi.org/10.1093/imaiai/iay016)
[44] Frank M and Wolfe P 1956 An algorithm for quadratic programming *Nav. Res. Logist. Q.* **3** 95–110
[45] Gustafsson M G L 2000 Surpassing the lateral resolution limit by a factor of two using structured illumination microscopy *J. Microsc.* **198** 82–7
[46] Harchaoui Z, Juditsky A and Nemirovski A 2015 Conditional gradient algorithms for norm-regularized smooth convex optimization *Math. Program.* **152** 75–112
[47] Hell S W and Wichmann J 1994 Breaking the diffraction resolution limit by stimulated emission: stimulated-emission-depletion fluorescence microscopy *Opt. Lett.* **19** 780–2
[48] Henriques R, Lelek M, Fornasiero E F, Valtorta F, Zimmer C and Mhlanga M M 2010 Quickpalm: 3d real-time photoactivation nanoscopy image processing in imagej *Nat. Methods* **7** 339
[49] Herzet C, Drémeau A and Soussen C 2016 Relaxed recovery conditions for OMP/OLS by exploiting both coherence and decay *IEEE Trans. Inf. Theory* **62** 459–70
[50] Hess S, Girirajan T P K and Mason M 2007 Ultra-high resolution imaging by fluorescence photoactivation localization microscopy *Biophys. J.* **91** 4258–72
[51] Hua Y and Sarkar T K 1990 Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise *IEEE Trans. Acoust. Speech Signal Process.* **38** 814–24
[52] Huang B, Wang W, Bates M and Zhuang X 2008 Three-dimensional super-resolution imaging by stochastic optical reconstruction microscopy *Science* **319** 810–3
[53] Huang H and Makur A 2011 Backtracking-based matching pursuit method for sparse signal reconstruction *IEEE Signal Process. Lett.* **18** 391–4
[54] Huang J, Sun M, Gumpper K, Chi Y and Ma J 2015 3d multifocus astigmatism and compressed sensing (3d macs) based superresolution reconstruction *Biomed. Opt. Express* **6** 902–17
[55] Huang J, Sun M, Ma J and Chi Y 2017 Super-resolution image reconstruction for high-density three-dimensional single-molecule microscopy *IEEE Trans. Comput. Imaging* **3** 763–73
[56] Jacques L and De Vleeschouwer C 2008 A geometrical study of matching pursuit parametrization *IEEE Trans. Signal Process.* **56** 2835–48
[57] Jaggi M 2013 Revisiting frank-wolfe: projection-free sparse convex optimization *ICML (1)* pp 427–35
[58] Juette M, Gould T J, Lessard M, Mlodzianoski M, Nagpure B S, Thomas Bennett B, Hess S and Bewersdorf J 2008 Three-dimensional sub-100 nm resolution fluorescence microscopy of thick samples *Nat. Methods* **5** 527–9
[59] Kailath T 1990 ESPRIT-estimation of signal parameters via rotational invariance techniques *Opt. Eng.* **29** 296
[60] Kirshner H, Vonesch C and Unser M 2013 Can localization microscopy benefit from approximation theory? *IEEE 10th Int. Symp. on Biomedical Imaging* (IEEE) pp 588–91
[61] Lasserre J B 2000/01 Global optimization with polynomials and the problem of moments *SIAM J. Optim.* **11** 796–817
[62] Lasserre J B 2010 *Moments, Positive Polynomials and their Applications* (*Imperial College Press Optimization Series* vol 1) (London: Imperial College Press) p xxii
[63] Levitin E S and Polyak B T 1966 Constrained minimization methods *Zh. Vychisl. Mat. Mat. Fiz.* **6** 787–823
[64] Liang J, Fadili J and Peyré G 2017 Activity identification and local linear convergence of forward-backward-type methods *SIAM J. Optim.* **27** 408–37

[65] Liao W and Fannjiang A 2016 MUSIC for single-snapshot spectral estimation: stability and super-resolution *Appl. Comput. Harmon. Anal.* **40** 33–67
[66] Mallat S and Zhang Z 1994 Matching pursuit with time-frequency dictionaries *IEEE Trans. Signal Process.* **41** 3397–415
[67] Massias M, Gramfort A and Salmon J 2017 From safe screening rules to working sets for faster lasso-type solvers (arXiv:1703.07285)
[68] Morgenshtern V I and Candès E J 2016 Super-resolution of positive sources: the discrete setup *SIAM J. Imaging Sci.* **9** 412–44
[69] Poon C and Peyré G 2019 Multidimensional sparse super-resolution *SIAM J. Math. Anal.* **51** 1–44
[70] Rama Prasanna Pavani S, Thompson M A, Biteen J S, Lord S, Liu N, Twieg R, Piestun R and Moerner W 2009 Three-dimensional, single-molecule fluorescence imaging beyond the diffraction limit by using a double-helix point spread function *Proc. Natl Acad. Sci. USA* **106** 2995–9
[71] Reemtsen R and Rückmann J J 1998 *Semi-Infinite Programming* vol 25 (New York: Springer)
[72] Rockafellar R T 2015 *Convex Analysis* (Princeton, NJ: Princeton University Press)
[73] Rudin W 1987 *Real and Complex Analysis* 3rd edn (New York: McGraw-Hill)
[74] Rust M J, Bates M and Zhuang X 2006 Sub-diffraction-limit imaging by stochastic optical reconstruction microscopy (storm) *Nat. Methods* **3** 793–6
[75] Sage D, Kirshner H, Pengo T, Stuurman N, Min J, Manley S and Unser M 2015 Quantitative evaluation of software packages for single-molecule localization microscopy *Nat. Methods* **12** 06
[76] Sage D 2019 Super-resolution fight club: assessment of 2D and 3D single-molecule localization microscopy software *Nat. Methods* **16** 387
[77] Schmidt R 1986 Multiple emitter location and signal parameter estimation *IEEE Trans. Antennas Propag.* **34** 276–80
[78] Soubies E, Schaub S, Radwanska A, Van Obberghen-Schilling E, Blanc-Féraud L and Aubert G 2016 A framework for multi-angle tirf microscope calibration *IEEE 13th Int. Symp. on Biomedical Imaging* (IEEE) pp 668–71
[79] Soussen C, Gribonval R, Idier J and Herzet C 2013 Joint k-step analysis of orthogonal matching pursuit and orthogonal least squares *IEEE Trans. Inform. Theory* **59** 3158–74
[80] Soussen C, Idier J, Duan J and Brie D 2015 Homotopy based algorithms for ℓ_0-regularized least-squares *IEEE Trans. Signal Process.* **63** 3301–16
[81] Tang G 2015 Resolution limits for atomic decompositions via markov-bernstein type inequalities *Int. Conf. on Sampling Theory and Applications (SampTA) (*IEEE) pp 548–52
[82] Tang G, Bhaskar B N and Recht B 2013 Sparse recovery over continuous dictionaries-just discretize *Asilomar Conf. on Signals, Systems and Computers* pp 1043–7
[83] Tibshirani R 1996 Regression shrinkage and selection via the lasso *J. R. Stat. Soc.* B **58** 267–88
[84] Toh K C and Yun S 2010 An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems *Pac. J. Optim.* **6** 615–40
[85] Tropp J A and Gilbert A C 2007 Signal recovery from random measurements via orthogonal matching pursuit *IEEE Trans. Inf. Theory* **53** 4655–66
[86] Tseng P 2001 Convergence of a block coordinate descent method for nondifferentiable minimization *J. Optim. Theory Appl.* **109** 475–94
[87] Wu T T and Lange K 2008 Coordinate descent algorithms for lasso penalized regression *Ann. Appl. Stat.* **2** 224–44
[88] Zhang B, Zerubia J and Olivo-Marin J C 2007 Gaussian approximations of fluorescence microscope point-spread function models *Appl. Opt.* **46** 1819–29
[89] Zheng C, Zhao G, Liu W, Chen Y, Zhang Z, Jin L, Xu Y, Kuang C and Liu X 2018 Three-dimensional super-resolved live cell imaging through polarized multi-angle tirf *Opt. Lett.* **43** 1423–6