On the linear convergence rates of exchange and continuous methods for total variation minimization

Axel Flinth, Frédéric de Gournay, Pierre Weiss

Noname manuscript No.
(will be inserted by the editor)

On the linear convergence rates of exchange and continuous methods for total variation minimization

Axel Flinth · Frédéric de Gournay · Pierre Weiss

the date of receipt and acceptance should be inserted later

Abstract We analyze an exchange algorithm for the numerical solution total-variation regularized inverse problems over the space $\mathcal{M}(\Omega)$ of Radon measures on a subset Ω of $\mathbb{R}^d$. Our main result states that under some regularity conditions, the method eventually converges linearly. Additionally, we prove that continuously optimizing the amplitudes of positions of the target measure will succeed at a linear rate with a good initialization. Finally, we propose to combine the two approaches into an alternating method and discuss the comparative advantages of this approach.

Keywords: Total variation minimization, inverse problems, superresolution, semi-infinite programming.
MSC Classification: 49M25, 49M29, 90C34, 65K05.

Acknowledgement

The authors acknowledge support from ANR JCJC OMS.

1 Introduction

1.1 The problem

The main objective of this paper is to develop and analyze iterative algorithms to solve the following infinite dimensional problem:

$$\inf_{\mu \in \mathcal{M}(\Omega)} J(\mu) \stackrel{\text{def.}}{=} \|\mu\|_{\mathcal{M}} + f(A\mu), \tag{$\mathcal{P}(\Omega)$}$$

where Ω is a bounded open domain of $\mathbb{R}^d$, $\mathcal{M}(\Omega)$ is the set of Radon measures on Ω, $\|\mu\|_{\mathcal{M}}$ is the total variation (or mass) of the measure μ, $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ is a convex lower semi-continuous function with non-empty domain and $A : \mathcal{M}(\Omega) \to \mathbb{R}^m$ is a linear measurement operator.

An important property of Problem ($\mathcal{P}(\Omega)$) is that at least one of its solutions $\mu^\star$ has a support restricted to s distinct points with $s \leq m$ (see e.g. [33,14,3]), i.e. is of the form

$$\mu^\star = \sum_{i=1}^{s} \alpha_i^\star \delta_{\xi_i}, \tag{1}$$

Axel Flinth
University of Gothenburg and Chalmers University of Technology E-mail: flinth (at) chalmers.se

Frédéric de Gournay
IMT, Université de Toulouse, CNRS

Pierre Weiss
IMT, Université de Toulouse, CNRS

with $\xi_i \in \Omega$ and $\alpha_i^\star \in \mathbb{R}$. This property motivates us to study a class of *exchange* algorithms. They were introduced as early as 1934 [26] and then extended in various manners [25]. They consist in discretizing the domain Ω coarsely and then refining it adaptively based on the analysis of so-called dual certificates. If the refinement process takes place around the locations (ξ_i) only, these methods considerably reduce the computational burden compared to a finely discretized mesh.

Our main results consist in a set of convergence rates for this algorithm that depend on the regularity of f and on the non-degeneracy of a dual certificate at the solution. We also show the linear convergence rate for first order algorithms that continuously vary the coefficients α_i and x_i of a discrete measure. Finally, we show that algorithms alternating between an exchange step and a continuous method share the best of both worlds: the global convergence guarantees of exchange algorithms together with the efficiency of first order methods. This yields a fast adaptive method with strong convergence guarantees for total variation minimization and related problems.

1.2 Applications

Our initial motivation to study the problem $(\mathcal{P}(\Omega))$ stems from signal processing applications. We recover an infinite dimensional version of the *basis pursuit* problem [6] by setting

$$f(x) = \iota_{\{y\}}(x) = \begin{cases} 0 & \text{if } x = y \\ +\infty & \text{otherwise.} \end{cases}$$

Similarly, the choice $f(x) = \frac{\tau}{2}\|x - y\|_2^2$, leads to an extension of the LASSO [29] called Beurling LASSO [8]. Both problems proved to be extremely useful in engineering applications. They got a significant attention recently thanks to theoretical progresses in the field of super-resolution [8,28,5,12]. Our results are particularly strong for the quadratic fidelity term.

Another less popular application in approximation theory [14], which was revived recently [31], is "generalized" total variation minimization. Given a surjective Fredholm operator $L : B(\Omega) \to \mathcal{M}(\Omega)$, where $B(\Omega)$ is a suitably defined Banach space, we consider the following problem

$$\inf_{u \in B(\Omega)} \|Lu\|_{\mathcal{M}} + f(Au). \tag{2}$$

The solutions of this problem can be proved to be (generalized) splines with free knots [31]. Following [15] and letting L^+ denote a pseudo-inverse of L, solving this problem can be rephrased as

$$\inf_{\mu \in \mathcal{M}(\Omega), u_K \in \ker(L)} \|\mu\|_{\mathcal{M}} + f(A(L^+\mu + u_K)), \tag{3}$$

which is a variant of $(\mathcal{P}(\Omega))$ that can also be solved with the proposed algorithms.

1.3 Numerical approaches in signal processing

The progresses on super-resolution [8,28,5,12] motivated researchers from this field to develop numerical algorithms for the resolution of Problem $(\mathcal{P}(\Omega))$. By far the most widespread approach is to use a fine uniform discretization and solve a finite dimensional problem. The complexity of this approach is however too large if one wishes high precision solutions. This approach was analyzed from a theoretical point of view in [27,12] for instance. The first papers investigating the use of $(\mathcal{P}(\Omega))$ for super-resolution purposes advocated the use of semi-definite relaxations [28,5], which are limited to specific measurement functions and domains, such as trigonometric polynomials on the 1D torus $\mathbb{T}$. The limitations were significantly reduced in [9], where the authors suggested the use of Lasserre hierarchies. These methods are however currently unable to deal with large scale problems. Another approach suggested in [4], consists in adding one point to a discretization set iteratively, where a so-called dual certificate is maximal. The weights of a measure supported on the set of added points are then updated using an ad-hoc rule. The authors refer to this algorithm as a mix between a Frank-Wolfe (or conditional gradient) algorithm and a LASSO type method. More recently, [30] began investigating the use of methods that continuously vary the positions (x_i) and amplitudes (α_i) of discrete measures parameterized as $\mu = \sum_{i=1}^s \alpha_i \delta_{x_i}$. The authors gave sufficient conditions for a simple gradient descent on the product-space (α, x) to

converge. In [2] and [10], this method was used alternatively with a Frank-Wolfe algorithm, the idea being to first add Dirac masses roughly at the right locations and then to optimize their locations and position continuously, leading to promising numerical results. Surprisingly enough, it seems that the connection with the mature field of semi-infinite programming has been ignored (or not explicitly stated) in all the mentioned references.

1.4 Some numerical approaches in semi-infinite programming

A semi-infinite program [25,17] is traditionally defined as a problem of the form

$$\min_{\substack{q\in Q\\ c(x,q)\leq 0, x\in\Omega}} u(q) \tag{SIP[Ω]}$$

where Q and Ω are subsets of $\mathbb{R}^m$ and $\mathbb{R}^n$ respectively, $u : Q \to \mathbb{R}$ and $c : \Omega \times Q \to \mathbb{R}$ are functions. The term semi-infinite stems from the fact that the variable q is finite-dimensional, but it is subject to infinitely many constraints $c(x, q) \leq 0$ for $x \in \Omega$. In order to see the connection between the semi-infinite program (SIP[Ω]) and our problem ($\mathcal{P}(\Omega)$), we can formulate its *dual*, which reads as

$$\sup_{q\in\mathbb{R}^m, \|A^*q\|_\infty\leq 1} -f^*(q). \tag{$\mathcal{D}(\Omega)$}$$

This dual will play a critical role in all the paper and it is easy to relate it to a SIP by setting $Q = \mathbb{R}^m$, $u = f^*$ and $c(x, q) = |(A^*q)(x)| - 1$.

Many numerical methods have been and are still being developed for semi-infinite programs and we refer the interested reader to the excellent chapter 7 of the survey book [25] for more insight. We sketch below two classes of methods that are of interest for our concerns.

1.4.1 Exchange algorithms

A canonical way of discretizing a semi-infinite program is to simply control finitely many of the constraints, say $c(x, q) \leq 0$ for $x \in \Omega_0 \subseteq \Omega$, where Ω_0 is finite. The discretized problem SIP[Ω_0] can then be solved by standard proximal methods or interior point methods. In order to obtain convergence towards an exact solution of the problem, it is possible to choose a sequence (Ω_k) of nested sets such that $\bigcup_k \Omega_k$ is dense in Ω. Solving the problems SIP[Ω_k] for large k however leads to a high numerical complexity due to the high number of discretization points. The idea of exchange algorithms is to iteratively update the discretization sets Ω_k in a more clever manner than simply making them denser. A generic description is given by Algorithm 1.

Algorithm 1 A Generic Exchange Algorithm

1: **Input:** Objective function u, Constraint function c, Constraint sets Ω and Q, Initial discretization set Ω_0.
2: **while** Not converged **do**
3: Set $q_k \in \underset{\substack{q\in Q\\ c(x,q)\leq 0, x\in\Omega_k}}{\operatorname{argmin}} u(q)$
4:
5: Set $\Omega_{k+1} = \text{Update\_Rule}(\Omega_k, q_k, k)$.
6: **end while**
7: **Output:** The last iterate q_∞.

In this paper, we consider Update_Rules of the form

$$\Omega_{k+1} \subset \Omega_k \cup \{x_k^1, \dots, x_k^{p_k}\},$$

where the points x_k^i are *local maximizers* of $c(\cdot, q_k)$. At each iteration, the set of discretization points can therefore be updated by adding and dropping a few prescribed points, explaining the name 'exchange'. The simplest rule consists of adding the single most violating point, i.e.

$$\Omega_{k+1} = \Omega_k \cup \operatorname*{argmax}_{x\in\Omega} c(x, q_k). \tag{4}$$

It seems to be the first exchange algorithm and it first appeared under a less general form as the Remez algorithm in the 30's [26]. It also shares similarities with the Frank-Wolfe (a.k.a. conditional gradient) method [16], which iteratively adds a point at a location where the constraint is most violated. It however differs in the way the solution q_k is updated. The connection was discussed recently in [13] for problems where the total variation term is used as a constraint. The use of the Frank-Wolfe algorithm for penalized total variation problems was also clarified recently in [10] using an epigraphical lift.

The update rule (4) is sufficient to guarantee convergence in the generic case and to ensure a decay of the cost function in $O\left(\frac{1}{k}\right)$, see [20]. Although 'exchange' suggests that points are both added and subtracted, methods for which $\Omega_k \subseteq \Omega_{k+1}$ are also coined exchange algorithms. The use of such rules often leads to easier convergence analyses, since we get monotonicity of the objective values $u(q_k)$ for free [17]. Other examples [18] include only adding points if they exceed a certain margin, i.e. $c(x,y) \geq \epsilon_k$, or all local maxima of $c(q_k, \cdot)$. In the case of convex functions f, algorithms that both add and remove points can be derived and analyzed with the use of cutting plane methods. All these instances have their pros and cons and perform differently on different types of problems. Since a semi-infinite program basically allows to minimize *arbitrary* continuous and finite dimensional problems, a theoretical comparison should depend on additional properties of the problem.

1.4.2 Continuous methods

Every iteration of an exchange algorithm can be costly: it requires solving a convex program with a number of constraints that increases if no discretization point is dropped. In addition, the problems tend to get more and more degenerate as the discretization points cluster, leading to numerical inaccuracies. In practice it is therefore tempting to use the following two-step strategy: i) find an approximate solution $\mu_k = \sum_{i=1}^{p_k} \alpha_k^i \delta_{x_k^i}$ of the primal problem $(\mathcal{P}(\Omega))$ using k iterations of an exchange algorithm and ii) continuously move the positions $X = (x_i)$ and amplitudes $\alpha = (\alpha_i)$ starting from (α_k, X_k) to minimize $(\mathcal{P}(\Omega))$ using a nonlinear programming approach such as a gradient descent, a conjugate gradient algorithm or a Newton approach.

This procedure supposes that the output μ_k of the exchange algorithm has the right number $p_k = s$ of Dirac masses, that their amplitudes satisfy $\operatorname{sign}(\alpha_i) = \operatorname{sign}(\alpha_i^\star)$ and that μ_k lies in the basin of attraction of the optimization algorithm around the global minimum $\mu^\star$. To the best of our knowledge, knowing a priori when those conditions are met is still an open problem and deciding when to switch from an exchange algorithm to a continuous method therefore relies on heuristics such as detecting when the number of masses p_k stagnates for a few iterations. The cost of continuous methods is however much smaller than that of exchange algorithms since they amount to work over a small number $s(d+1)$ of variables. In addition, the instabilities mentioned earlier are significantly reduced for these methods. This observation was already made in [2,10] and proved in [30] for specific problems.

1.5 Contribution

Many recent results in the field of super-resolution provide sufficient conditions for a *non degenerate source condition* to hold [28,11,5,23]. The non degeneracy means that the solution $q^\star$ of $(\mathcal{D}(\Omega))$ is unique and that the *dual certificate* $|A^* q^\star|$ reaches 1 at exactly s points, where it is strictly concave. The main purpose of this paper is to study the implications of this non degeneracy for the convergence of a class of exchange algorithms and for continuous methods based on gradient descents. Our main results are as follows:

1. We show an eventual linear convergence rate of a class of exchange algorithms for convex functions f with Lipschitz continuous gradient. More precisely, we prove that after a finite number of iterations N the algorithm outputs vectors q_k such that the set

$$X_k \overset{\text{def.}}{=} \{x \in \Omega \,|\, x \text{ local maximizer of } |A^* q_k|\,,\ |A^* q_k|(x) \geq 1\} \tag{5}$$

   contains exactly s-points $(x_k^1, \ldots, x_k^s)$.
   Letting $\widehat{\mu}_k = \sum_{i=1}^s \alpha_i^k \delta_{x_i^k}$ denote the solution of the finite dimensional problem $\inf_{\mu \in \mathcal{M}(X_k)} \|\mu\|_{\mathcal{M}} + f(A\mu)$, we also show the linear convergence rate of the cost function $J(\widehat{\mu}_k)$ to $J(\mu^\star)$ and of the support in the following sense: after a number N of initial iterations, it will take no more that $k_\tau = C \log(\tau^{-1})$ iterations to ensure that the Hausdorff distance between the sets $X_{k_\tau + N}$ and ξ is smaller than τ. A similar statement holds for the coefficient vectors α^k.

Of importance, let us mention that similar results were derived under slightly different conditions by Pieper and Walter in [22]. The two works were carried out independently at the same time.
2. We also show that a well-initialized gradient descent algorithm on the pair (α, x) converges linearly to the true solution $\mu^\star$ and explicit the width of the basin of attraction.
3. We then show how the proposed guarantees may explain the success of methods alternating between exchange methods and continuous methods at each step, in a spirit similar to the sliding Frank-Wolfe algorithm [10].
4. We finally illustrate the above results on total variation based problems in 1D and 2D.

2 Preliminaries

2.1 Notation

In all the paper, Ω designs an open *bounded* domain of $\mathbb{R}^d$. The boundedness assumptions plays an important role to control the number of elements in the discretization procedures. A *grid* Ω_k is a finite set of points in Ω. Its cardinality is denoted by $|\Omega_k|$. The distance from a set Ω_2 to a set Ω_1 is defined by

$$\operatorname{dist}(\Omega_1|\Omega_2) = \sup_{x_2\in\Omega_2} \inf_{x_1\in\Omega_1} \|x_1 - x_2\|_2. \tag{6}$$

Note that this definition of distance is not symmetric: in general $\operatorname{dist}(\Omega_1|\Omega_2) \neq \operatorname{dist}(\Omega_2|\Omega_1)$.

We let $\mathcal{C}_0(\Omega)$ denote the set of continuous functions on Ω vanishing on the boundary. The set of Radon measures $\mathcal{M}(\Omega)$ can be identified as the dual of $\mathcal{C}_0(\Omega)$, i.e. the set of continuous linear forms on $\mathcal{C}_0(\Omega)$. For any sub-domain $\Omega_k \subset \Omega$, we let $\mathcal{M}(\Omega_k)$ denote the set of Radon measures supported on Ω_k. For $p \in [1, +\infty]$, the L^p-norm of a function $u \in \mathcal{C}_0(\Omega)$ is denoted by $\|u\|_p$. The total variation of a measure $\mu \in \mathcal{M}(\Omega)$ is denoted $\|\mu\|_{\mathcal{M}}$. It can be defined through duality as

$$\|\mu\|_{\mathcal{M}} = \sup_{\substack{u\in\mathcal{C}_0(\Omega)\\ \|u\|_\infty\leq 1}} \mu(u). \tag{7}$$

The ℓ^p-norm of a vector $x \in \mathbb{R}^m$ is also denoted $\|x\|_p$. The Frobenius norm of a matrix M is denoted by $\|M\|_F$.

Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ denote a convex lower semi-continuous function with non-empty domain $\operatorname{dom}(f) = \{x \in \mathbb{R}^m, f(x) < +\infty\}$. Its subdifferential is denoted ∂f. Its Fenchel transform f^* is defined by

$$f^*(y) = \sup_{x\in\mathbb{R}^m} \langle x, y\rangle - f(x).$$

If f is differentiable, we let $f' \in \mathbb{R}^m$ denote its gradient and if it is twice differentiable, we let $f'' \in \mathbb{R}^{m\times m}$ denote its Hessian matrix. We let $\|f'\|_\infty = \sup_{x\in\Omega} \|f'(x)\|_2$ and $\|f''\|_\infty = \sup_{x\in\Omega} \|f''(x)\|$, where $\|f''(x)\|$ is the largest singular value of $f''(x)$. A convex function f is said to be l-strongly convex if

$$f(x_2) \geq f(x_1) + \langle \eta, x_2 - x_1\rangle + \frac{l}{2}\|x_2 - x_1\|_2^2 \tag{8}$$

for all $(x_1, x_2) \in \mathbb{R}^m \times \mathbb{R}^m$ and all $\eta \in \partial f(x_1)$. A differentiable function f is said to have an L-Lipschitz gradient if it satisfies $\|f'(x_1) - f'(x_2)\|_2 \leq L\|x_1 - x_2\|_2$. This implies that

$$f(x_2) \leq f(x_1) + \langle f'(x_1), x_2 - x_1\rangle + \frac{L}{2}\|x_2 - x_1\|_2^2 \text{ for all } (x_1, x_2) \in \mathbb{R}^m \times \mathbb{R}^m. \tag{9}$$

We recall the following equivalence [19]:

Proposition 2.1 *Let $f : \mathbb{R}^m \to \mathbb{R} \cup \{+\infty\}$ denote a convex and closed function with non empty domain. Then the following two statements are equivalent:*

- *f has an L-Lipschitz gradient.*
- *f^* is $\frac{1}{L}$-strongly convex.*

The linear measurement operators A considered in this paper can be viewed as a collection of m continuous functions $(a_i)_{1\leq i\leq m}$. For $x \in \Omega$, the notation $A(x)$ corresponds to the vector $[a_1(x), \ldots, a_m(x)] \in \mathbb{R}^m$.

2.2 Existence results and duality

In order to obtain existence and duality results, we will now make further assumptions.

Assumption 1 *$f : \mathbb{R}^m \to \mathbb{R} \cup \{\infty\}$ is convex and lower bounded. In addition, we assume that either* $\mathrm{dom}(f) = \mathbb{R}^m$ *or that f is polyhedral (that is, its epigraph is a finite intersection of closed halfspaces).*

Assumption 2 *The operator A is weak-$*$-continuous. Equivalently, the* measurement functionals a_i^* *defined by $\langle a_i^*, \mu \rangle = (A(\mu))_i$ are given by*

$$\langle a_i^*, \mu \rangle = \int_\Omega a_i d\mu,$$

for functions $a_i \in \mathcal{C}_0(\Omega)$. In addition, we assume that A is surjective on $\mathbb{R}^m$.

The following results relate the primal and the dual.

Proposition 2.2 (Existence and strong duality) *Under Assumptions 1 and 2, the following statements are true:*

- *The primal problem ($\mathcal{P}(\Omega)$) and its dual ($\mathcal{D}(\Omega)$) both admit a solution.*
- *The following strong duality result holds*

$$\min_{\mu \in \mathcal{M}(\Omega)} \|\mu\|_{\mathcal{M}(\Omega)} + f(A\mu) = \max_{q \in \mathbb{R}^m, \|A^* q\|_\infty \leq 1} -f^*(q). \tag{10}$$

- *Let $(\mu^\star, q^\star)$ denote a primal-dual pair. They are related as follows*

$$A^* q^\star \in \partial_{\|\cdot\|_{\mathcal{M}}}(\mu^\star) \text{ and } -q^\star \in \partial f(A\mu^\star). \tag{11}$$

Proof The stated assumptions ensure the existence of a feasible measure μ. In addition, the primal function is coercive since f is bounded below. Since $\mathcal{M}(\Omega)$ can be viewed as the dual of the Banach space $\mathcal{C}_0(\Omega)$, we further have that bounded sets in $\mathcal{M}(\Omega)$ are compact in the weak-$*$-topology (this is the the Banach-Alaoglu theorem). Using these three facts, a standard argument now allows one to deduce the existence of a primal solution. The existence of a dual solution stems from the compactness of the set $\{q \in \mathbb{R}^m, \|A^* q\|_\infty \leq 1\}$ (which itself follows from the surjectivity of A) and the continuity of f^* on its domain. The strong duality result follows from [1, Thm 4.2]. The primal-dual relationship directly derives from the first order optimality conditions.

The left inclusion in equation (11) plays an important role, which is well detailed in [12]. It implies that the support of $\mu^\star$ satisfies: $\mathrm{supp}(\mu^\star) \subseteq \{x \in \Omega, |A^* q^\star(x)| = 1\}$.

3 An Exchange Algorithm and its convergence

3.1 The algorithm

We assume that an initial grid $\Omega_0 \subseteq \Omega$ is given (e.g. a coarse Euclidean grid). Given a discretization Ω_k, we can define a discretized primal problem ($\mathcal{P}(\Omega_k)$)

$$\inf_{\mu \in \mathcal{M}(\Omega_k)} \|\mu\|_{\mathcal{M}} + f(A\mu), \tag{$\mathcal{P}(\Omega_k)$}$$

and its associated dual ($\mathcal{D}(\Omega_k)$)

$$\sup_{q \in \mathbb{R}^m, |A^* q(x)| \leq 1, \ \forall x \in \Omega_k} -f^*(q). \tag{$\mathcal{D}(\Omega_k)$}$$

In this paper, we will investigate the exchange rule below:

$$\Omega_{k+1} = \Omega_k \cup X_k \text{ where } X_k \text{ is defined in (5).} \tag{12}$$

The implementation of this rule requires finding X_k, the set of all the local maximizers of $|A^* q_k|$ exceeding 1.

3.2 A generic convergence result

The exchange algorithm above converges under quite weak assumptions. For instance, it is enough to assume that the function f is differentiable.

Assumption 3 *The data fitting function $f : \mathbb{R}^m \to \mathbb{R}$ is differentiable with L-Lipschitz continuous gradient.*

Alternatively, we may assume that the initial set Ω_0 is fine enough, which in particular implies that $|\Omega_0| \geq m$.

Assumption 4 *The initial set Ω_0 is such that A restricted to Ω_0 is surjective.*

We may now present and prove our first result.

Theorem 3.1 (Generic convergence) *Under assumptions 1, 2 and 3 or 4, a subsequence of (μ_k, q_k) will converge in the weak-$*$-topology towards a solution pair $(\mu^\star, q^\star)$ of $(\mathcal{P}(\Omega))$ and $(\mathcal{D}(\Omega))$, as well as in objective function value. If the solution of $(\mathcal{P}(\Omega))$ and/or $(\mathcal{D}(\Omega))$ is unique, the entire sequence will converge.*

Proof First remark that the sequence $(\|\mu_k\|_{\mathcal{M}} + f(A\mu_k))_{k\in\mathbb{N}}$ is non-increasing since the spaces $\mathcal{M}(\Omega_k)$ are nested. Due to the boundedness below of f, the same must be true for $(\|\mu_k\|_{\mathcal{M}})$. Hence there exists a subsequence (μ_k), which we do not relabel, that weak-$*$ converges towards a measure μ_∞.

Now, we will prove that the sequence of dual variables $(q_k)_{k\in\mathbb{N}}$ is bounded. If Assumption 3 is satisfied, then f^* is strongly convex and since 0 is a feasible point, we must have $q_k \in \{q \in \mathbb{R}^m, f^*(q) \leq f^*(0)\}$, which is bounded. Alternatively, if Assumption 4 is satisfied, notice that $1 \geq \|A_k^* q_k\|_\infty \geq \|A_0^* q_k\|_\infty$. Since A_0 is surjective, the previous inequality implies that $(\|q_k\|_2)_{k\in\mathbb{N}}$ is bounded. Hence, in both cases, the sequence $(q_k)_{k\in\mathbb{N}}$ converges up to a subsequence to a point q_∞.

The key is now to prove that $\|A^* q_\infty\|_\infty \leq 1$. To this end, let us first argue that the family $(A^* q_k)_{k\in\mathbb{N}}$ is equicontiuous. For this, let $\epsilon > 0$ be arbitrary. Since the functions $a_i \in \mathcal{C}_0(\Omega)$ all are uniformly continuous, there exists a $\delta > 0$ with the property

$$\|x - y\|_2 < \delta \;\Rightarrow\; |a_i(x) - a_i(y)| < \frac{\epsilon}{\sup_k \|q_k\|_1} \text{ for all } i.$$

Consequently,

$$\begin{aligned}\|x - y\|_2 < \delta \;\Rightarrow\; |(A^* q_k)(x) - (A^* q_k)(y)| &= \left|\sum_{i=1}^m (a_i(x) - a_i(y)) q_k(i)\right| \leq \sum_{i=1}^m |a_i(x) - a_i(y)|\,|q_k(i)| \\ &< \frac{\epsilon}{\sup_k \|q_k\|_1} \sum_{i=1}^m |q_k(i)| \leq \epsilon. \end{aligned} \tag{13}$$

Due to the convergence of $(q_k)_{k\in\mathbb{N}}$, the sequence $(A^* q_k)_{k\in\mathbb{N}}$ is converging strongly to $A^* q_\infty$. We will now prove that $\|A^* q_\infty\|_\infty \leq 1$. If for some k, $\|A^* q_k\|_\infty \leq 1$, we will have $A^* q_\ell = A^* q_k$ for all $\ell \geq k$, and in particular $q_\infty = q_k$ and thus $\|A^* q_\infty\| \leq 1$. Hence, we may assume that $\|A^* q_k\|_\infty > 1$ for each k, i.e. that we add at least one point to Ω_k in each iteration.

Now, towards a contradiction, assume that $\|A^* q_\infty\|_\infty = 1 + 2\epsilon$ for an $\epsilon > 0$. Set δ as in (13). For each $k \in \mathbb{N}$, let $x_k^\star$ be the element in $\operatorname{argmax}_x |(A^* q_k)(x)|$ which has the largest distance to Ω_k. Due to $a_\ell \in \mathcal{C}_0(\Omega)$ for each k, there needs to exist a compact subset $C \subseteq \Omega$ such that $(x_k^\star)_k \subseteq C$. Indeed, there exists for each $\ell = 1, \ldots, m$ a C_ℓ such that $|a_\ell(x)| \leq (\sup_k \|q_k\|_1)^{-1}$ for all $x \notin C_\ell$. Now, if $x \notin C \overset{\text{def.}}{=} \bigcup_{\ell=1}^m C_\ell$, we get

$$|A^* q_k(x)| = \left|\sum_{i=1}^m a_i(x) q_k(i)\right| \leq \sum_{i=1}^m |a_i(x)|\,|q_k(i)| < \frac{1}{\sup_k \|q_k\|_1} \sum_{i=1}^m |q_k(i)| \leq 1$$

for every k. Since $|A^* q_k(x_k^\star)| > 1$, we conclude $(x_k^\star)_k \subseteq C$. Consequently, a subsequence (which we do not rename) of $(x_k^\star)$ must converge. Thus, for some k_0 and every $k > k_0$, we have $\|x_k^\star - x_{k_0}^\star\|_2 < \delta$. We then have

$$\|A^* q_k\|_\infty = |(A^* q_k)(x_k^\star)| < \left|(A^* q_k)(x_{k_0}^\star)\right| + \epsilon \leq 1 + \epsilon.$$

In the last estimate, we used the constraint of $(\mathcal{D}(\Omega_k))$ and the fact that $x^\star_{k_0} \in \Omega_k$. Since the last inequality holds for every $k \geq k_0$, we obtain

$$\|A^* q_\infty\|_\infty = \lim_{k\to\infty} \|A^* q_k\|_\infty \leq 1 + \epsilon,$$

where we used the fact that $(A^* q_k)_k$ converges strongly towards $A^* q_\infty$. This is a contradiction, and hence, we do have $\|A^* q_\infty\|_\infty \leq 1$.

Overall, we proved that the primal-dual pair (μ_∞, q_∞) is feasible. It remains to prove that it is actually a solution. To do this, let us first remark that $\|\mu_\infty\|_\mathcal{M} + f(A\mu_\infty) \geq -f^*(q_\infty)$ by weak duality. To prove the second inequality, first notice that the weak-$*$-continuity of A implies that $A\mu_k \to A\mu_\infty$. Assumption 1 furthermore implies that f is lower semi-continuous. As a supremum of linear functions, so is f^*. Since also $q_k \to q_\infty$, we conclude

$$f^*(q_\infty) + f(A\mu_\infty) \leq \liminf_{k\to\infty} f^*(q_k) + f(A\mu_k).$$

Assumptions 1, 2 together with Proposition 2.2 imply exact duality of the discretized problems. This means $f^*(q_k) + f(A\mu_k) = -\|\mu_k\|_\mathcal{M}$. Since the norm is weak-$*$-l.s.c. , we thus obtain

$$\liminf_{k\to\infty} f^*(q_k) + f(A\mu_k) = \liminf_{k\to\infty} -\|\mu_k\|_\mathcal{M} \leq -\liminf_{k\to\infty} \|\mu_k\|_\mathcal{M} \leq -\|\mu_\infty\|_\mathcal{M}.$$

Reshuffling these inequalities yields $\|\mu_\infty\|_\mathcal{M} + f(A\mu_\infty) \leq -f^*(q_\infty)$, i.e., the reverse inequality. Thus, μ_∞ and q_∞ fulfill the duality conditions, and are solutions. The final claim follows from a standard subsequence argument.

Remark 1 *Let us mention that the convergence result in Theorem 3.1 and its proof, is not new, see e.g. [24]. The proof technique can be applied to prove similar statements for other refinement rules. For instance, the result still holds if we add the single most violating point:*

$$\Omega_{k+1} \supseteq \Omega_k \cup \{x_k\} \text{ with } x_k \in \operatorname*{argmax}_{x\in\Omega} |A^* q_k|. \tag{14}$$

The result that we have just shown is very generally applicable. It however does not give us any knowledge of the convergence rate. The next section will be devoted to proving a linear convergence rate in a significant special case.

3.3 Non degenerate source condition

The idea behind adding points to the grid adaptively is to avoid a uniform refinement, which results in computationally expensive problems $(\mathcal{D}(\Omega_k))$. However, there is a priori no reason for the exchange rule not to refine in a uniform manner. In this section, we prove that additional assumptions improve the situation. First, we will from now on work under Assumption (3). It implies that the dual solutions q_k are unique for every k, since Proposition (2.1) ensures the strong convexity of the Fenchel conjugate f^*. We furthermore assume that the functions a_j are smooth.

Assumption 5 (Assumption on the measurement functionals) *The measurement functions a_j all belong to $\mathcal{C}^2_0(\Omega) \overset{\text{def.}}{=} \mathcal{C}_0(\Omega) \cap \mathcal{C}^2(\Omega)$ and their first and second order derivatives are uniformly bounded on Ω. We hence may define*

$$\kappa \overset{\text{def.}}{=} \sup_{\|q\|_2\leq 1} \|A^* q\|_\infty = \sup_{x\in\Omega} \|A(x)\|_2, \quad \kappa_\nabla \overset{\text{def.}}{=} \sup_{\|q\|_2\leq 1} \|(A^* q)'\|_\infty, \quad \kappa_{\text{hess}} \overset{\text{def.}}{=} \sup_{\|q\|_2\leq 1} \|(A^* q)''\|_\infty.$$

We also assume the following regularity condition on the solution $q^\star$ of $(\mathcal{D}(\Omega))$, and its corresponding primal solution $\mu^\star$.

Assumption 6 (Assumption on the primal-dual pair) *We assume that $(\mathcal{P}(\Omega))$ admits a unique s-sparse solution $\mu^\star$ supported on $\xi = (\xi_i)_{i=1}^s \in \Omega^s$:*

$$\mu^\star = \sum_{i=1}^s \alpha_i^\star \delta_{\xi_i}. \tag{15}$$

Let $q^\star$ denote the associated dual pair. We assume that the only points x for which $|A^*q^\star(x)| = 1$ are the points in ξ, and that the second derivative of $|A^*q^\star|$ is negative definite in each point ξ_i. It follows that there exists $\tau_0 > 0$ and $\gamma > 0$ such that

$$|A^*q^\star|''(x) \preccurlyeq -\gamma \operatorname{Id} \text{ and } |A^*q^\star|(x) \geq \frac{\gamma\tau_0^2}{2} \text{ for } x \text{ with } \operatorname{dist}(\xi|x) \leq \tau_0. \tag{16}$$

$$|(A^*q^\star)(x)| \leq 1 - \frac{\gamma\tau_0^2}{2} \text{ for } x \text{ with } \operatorname{dist}(\xi|x) \geq \tau_0. \tag{17}$$

We note that if Equations (16) and (17) are valid for some (γ, τ_0), they are also valid for any $(\tilde{\gamma}, \tilde{\tau}_0)$ with $\tilde{\gamma} \leq \gamma$ and $\tilde{\tau}_0 \leq \tau_0$.

Assumption (6) may look very strong and hard to verify in advance. Recent advances in signal processing actually show that it is verified under clear geometrical conditions. First, there will always exists at most m-sparse solutions to problem $(\mathcal{P}(\Omega))$, [33,14,3]. Therefore, the main difficulty comes from the uniqueness of the primal solution and from the two regularity conditions (16) and (17). These assumptions are called *non-degenerate source condition* of the *dual certificate* $A^*q^\star$ [12]. Many results in this direction have been shown for $f = \xi_{\{b\}}$ or $f(\cdot) = \frac{L}{2}\|\cdot - b\|_2^2$, where $b = A\mu_0$ with μ_0 a finitely supported measure. The papers [5,28,11] deal with different Fourier-type operators, whereas [23] provides an analysis for arbitrary integral operators sampled at random.

3.4 Auxiliary results

In this and the following sections, we always work under Assumptions 1, 2, 3 without further notice. We derive several lemmata that are direct consequences of the above assumptions. The first two rely strongly on the Lipschitz regularity of the gradient of f.

Lemma 3.2 (Boundedness of the dual variables) *Let $\bar{q} = \operatorname{argmin}_{q\in\mathbb{R}^m} f^*(q)$ denote the prox-center of f^*. For all $k \in \mathbb{N}$, we have*

$$\|q_k\|_2 \leq \sqrt{2L(f^*(0) - f^*(\bar{q}))} + \|\bar{q}\|_2 \overset{\text{def.}}{=} R. \tag{18}$$

Proof (Proof of Lemma 3.2) For all $k \in \mathbb{N}$, we have $0 \in \{q \in \mathbb{R}^m, \|A_k^*q\|_\infty \leq 1\}$, hence $f^*(q_k) \leq f^*(0)$. By strong convexity of f^* and optimality of $\bar{q}$ and q_k, we get:

$$f^*(0) \geq f^*(q_k) \geq f^*(\bar{q}) + \frac{1}{2L}\|q_k - \bar{q}\|_2^2. \tag{19}$$

Therefore $\|q_k - \bar{q}\|_2 \leq \sqrt{2L(f^*(0) - f^*(\bar{q}))}$ and the conclusion follows from a triangle inequality.

Proposition 3.3 *Let $q^\star$ be the solution of $(\mathcal{D}(\Omega))$. Let*

$$\rho \overset{\text{def.}}{=} \sqrt{\sup_{w\in\partial f^*(q^\star)} -L\langle w, q^\star\rangle}.$$

Then for any q, we have

$$f^*(q^\star) - f^*(q) + \frac{1}{2L}\|q - q^\star\|_2^2 \leq \rho^2 L^{-1}(\sup_{x\in\xi} |A^*q|(x) - 1).$$

Proof Let $M = \{q \in \mathbb{R}^m, f^*(q) \leq f^*(q^\star)\}$ denote the sub-level set of f^* and $D = \left\{q \in \mathbb{R}^n \mid \sup_{x\in\xi} |A^*q|(x) \leq 1\right\}$. We first claim that M and D only have the point $q^\star$ in common. Indeed $\mu^\star$ solves the problem $\mathcal{P}(\xi)$ and by strong duality of the problem restricted to $\mathcal{M}(\xi)$, $q^\star$ solves $\mathcal{D}(\xi)$. By strong convexity of f, $q^\star$ is the unique solution $\mathcal{D}(\xi)$, this exactly means $M \cap D = \{q^\star\}$.

The fact that $M \cap D = \{q^\star\}$ implies that there exists a separating hyperplane there. Since the hyperplane must be tangent to M, it can be written as $\{q \mid \langle w, q\rangle = \langle w, q^\star\rangle\}$ for a $w \in \partial f^*(q^\star)$, with $D \subset \{q \mid \langle w, q\rangle \geq \langle w, q^\star\rangle\}$. Consequently, letting $\epsilon = \sup_{x\in\xi} |A^*q(x)| - 1$, we have

$$(1+\epsilon)D \subset \{q \mid \langle w, q\rangle \geq (1+\epsilon)\langle w, q^\star\rangle\} = \{q \mid \langle w, q - q^\star\rangle \geq \epsilon\langle w, q^\star\rangle\}.$$

Now, the strong convexity of f^* implies for every $q \in (1+\epsilon)D \cap M$,

$$f^*(q) \geq f^*(q^\star) + \langle w, q - q^\star \rangle + \frac{1}{2L}\|q - q^\star\|_2^2 \geq f^*(q^\star) + \epsilon \langle w, q^\star \rangle + \frac{1}{2L}\|q - q^\star\|_2^2.$$

Rearranging this, we obtain

$$-\epsilon \langle w, q^\star \rangle \geq f^*(q^\star) - f^*(q) + \frac{1}{2L}\|q - q^\star\|_2^2.$$

which is the claim.

Before moving on, let us record the following proposition:

Proposition 3.4 *We have*

$$\|A(x) - A(y)\|_2 \leq \kappa_\nabla \|x - y\|_2 \quad \text{and} \quad \|A'(x) - A'(y)\|_F \leq \kappa_{\text{hess}} \|x - y\|_2. \tag{20}$$

Proof The proof of the first inequality of (20) is a standard Taylor expansion :

$$\begin{aligned}\|A(x) - A(y)\|_2 &= \sup_{\substack{q \in \mathbb{R}^m \\ \|q\|_2 = 1}} \langle q, A(x) - A(y) \rangle = \sup_{\substack{q \in \mathbb{R}^m \\ \|q\|_2 = 1}} |(A^*q)(x) - (A^*q)(y)| \\ &\leq \sup_{\substack{q \in \mathbb{R}^m \\ \|q\|_2 = 1}} \sup_{z \in [x,y]} \langle (A^*q)'(z), x - y \rangle \leq \sup_{\substack{q \in \mathbb{R}^m \\ \|q\|_2 = 1}} \|(A^*q)'\|_\infty \|x - y\|_2 \leq \kappa_\nabla \|x - y\|_2.\end{aligned}$$

The proof of the second part of (20) follows the same lines as the first part and is left to the reader.

The next two lemmata aim at transferring bounds from the geometric distances of the sets X_k, Ω_k and ξ to bounds on $|A^*q_k(\xi)|$. Using Proposition 3.3, we may then transfer these bounds to bounds on the errors of the dual solutions and the dual (or primal) objective values.

Lemma 3.5 *The following inequalities hold*

$$\|A^*q_k\|_\infty \leq 1 + \frac{R\kappa_{\text{hess}}}{2} \operatorname{dist}(\Omega_k | X_k)^2, \tag{21}$$

$$\begin{aligned} f^*(q^\star) - f^*(q_k) &\leq \frac{R\kappa_{\text{hess}}\rho^2}{2L} \operatorname{dist}(\Omega_k | X_k)^2, \\ \|q_k - q^\star\|_2 &\leq \operatorname{dist}(\Omega_k | X_k) \sqrt{R\kappa_{\text{hess}}}\rho. \end{aligned}$$

Proof (Proof of Lemma 3.5) To show (21), first notice that

$$\|A^*q_k\|_\infty \leq 1 + \|(A^*q_k)''\|_\infty \frac{\operatorname{dist}(\Omega_k | X_k)^2}{2}. \tag{22}$$

Indeed, by definition, the global maximum z of $|A^*q_k|$ lies in X_k and satisfies $(A^*q_k)'(z) = 0$. Furthermore, by construction, all points x in Ω_k satisfy $|A^*q_k(x)| \leq 1$. Using a Taylor expansion, we get for all $x \in \Omega$

$$|A^*q_k(x) - A^*q_k(z)| \leq \|(A^*q_k)''\|_\infty \frac{\|x - z\|_2^2}{2}.$$

Taking x as the point in Ω_k minimizing the distance to z leads to (22). In addition, we have $\|(A^*q_k)''\|_\infty \leq R\kappa_{\text{hess}}$ by Lemma 3.2, so that $\|A^*q_k\|_\infty \leq 1 + \epsilon$ with $\epsilon = R\kappa_{\text{hess}} \frac{\operatorname{dist}(\Omega_k | X_k)^2}{2}$.

Now, letting $C = \{q \,|\, \|A^*q\|_\infty \leq 1\}$, we have just proven that $q_k \in (1+\epsilon)C$. Furthermore, due to the optimality of q_k for the discretized problem and to the fact that $q^\star$ is feasible for that problem, we will have $f^*(q_k) \leq f^*(q^\star)$, i.e., q_k is included in the $f^*(q^\star)$-sub-level set of f^*: $M = \{q \in \mathbb{R}^m | f^*(q) \leq f^*(q^\star)\}$. An application of Proposition 3.3 now yields the result.

Lemma 3.6 *Suppose that* $\operatorname{dist}(X_k|\xi) \leq \delta$ *and* $\operatorname{dist}(\Omega_k|\xi) \leq \delta$. *Then*

$$f^*(q^\star) - f^*(q_k) \leq \frac{2R\kappa_{\text{hess}}\rho^2}{L} \cdot \delta \operatorname{dist}(\Omega_k|\xi)$$
$$\|q_k - q^\star\|_2 \leq \rho\sqrt{2R\kappa_{\text{hess}}}\sqrt{\delta \cdot \operatorname{dist}(\Omega_k|\xi)}.$$

Proof Let y_k^i (resp. x_k^i) be the point closest to ξ_i in Ω_k (resp. X_k). By assumption, we have $\|x_k^i - y_k^i\|_2 \leq 2\delta$. For all i, we have

$$|A^*q_k(\xi_i)| \leq |A^*q_k(y_k^i)| + \sup_{z\in[y_k^i,\xi_i]} \|(A^*q_k)'(z)\|_2\|\xi_i - y_k^i\|_2 \leq 1 + \sup_{z\in[y_k^i,\xi_i]} \|(A^*q_k)'(z)\|_2\|\xi_i - y_k^i\|_2. \tag{23}$$

Then, for all $z \in [y_k^i, \xi_i]$, using the fact that $(A^*q_k)'(x_k^i) = 0$, we get

$$\|(A^*q_k)'(z)\|_2 \leq R\kappa_{hess}\|z - x_k^i\|_2 \leq 2\delta R\kappa_{\text{hess}}.$$

Hence, we have $|A^*q_k(\xi_i)| \leq 1 + 2\delta R\kappa_{\text{hess}}\|\xi_i - y_k^i\|_2 \leq 1 + 2\delta R\kappa_{\text{hess}} \operatorname{dist}(\Omega_k|\xi)$. To conclude, we use Proposition 3.3 again.

The last assertion takes full advantage of Assumption 6 and the fact that the function $|A^*q^\star|$ is uniformly concave around its maximizers. It allows to transfer bounds from $\|q_k - q^\star\|_2$ to bounds on the distance from X_k to ξ.

Proposition 3.7 *Define* $c_q = \gamma \min\left(\frac{\tau_0^2}{2\kappa}, \frac{\tau_0}{\kappa_\nabla}, \frac{1}{\kappa_{\text{hess}}}\right)$ *and assume that* $\|q_k - q^\star\|_2 < c_q$, *then*

$$\operatorname{dist}(\xi|X_k) \leq \frac{\kappa_\nabla}{\gamma}\|q_k - q^\star\|_2.$$

Moreover, for each i, *if* B_i *is the ball or radius* τ_0 *around* ξ_i, *then* X_k *contains at most one point in* B_i *and* A^*q_k *has the same sign as* $A^*q^\star(\xi_i)$ *in* B_i.

Proof Define $\tau = \frac{\kappa_\nabla}{\gamma}\|q_k - q^\star\|$ and note that $\tau < \tau_0$. By Proposition 3.4, we have for each $x \in \Omega$

$$|(A^*q_k)(x) - (A^*q^\star)(x)| \leq \|A^*(q_k - q^\star)\|_\infty \leq \kappa\|q_k - q^\star\|_2 < \frac{\gamma\tau_0^2}{2}$$
$$\|(A^*q_k)'(x) - (A^*q^\star)'(x)\|_2 \leq \|(A^*(q_k - q^\star))'\|_\infty \leq \kappa_\nabla\|q_k - q^\star\|_2 = \gamma\tau$$
$$\|(A^*q_k)''(x) - (A^*q^\star)''(x)\|_2 \leq \|(A^*(q_k - q^\star))''\|_\infty \leq \kappa_{\text{hess}}\|q_k - q^\star\|_2 < \gamma.$$

The above inequalities together with Assumption 6 imply the following for all $1 \leq i \leq s$:

(i) For x with $\|x - \xi_i\|_2 \leq \tau_0$, we have $\operatorname{sign}(A^*q_k)(x) = \operatorname{sign}(A^*q^\star)(x) = \operatorname{sign}(A^*q^\star)(\xi_i)$.
(ii) For x with $\|x - \xi_i\|_2 \leq \tau_0$, we have $(|A^*q_k|)''(x) \prec (|A^*q^\star|)''(x) + \gamma\,\mathrm{id} \prec 0$.
(iii) For x with $\|x - \xi_i\|_2 \geq \tau_0$, we have $|(A^*q_k)(x)| < |(A^*q^\star)(x)| + \frac{\gamma\tau_0^2}{2} \leq 1 - \frac{\gamma\tau_0^2}{2} + \frac{\gamma\tau_0^2}{2} = 1$.
(iv) For x with $\tau < \|x - \xi_i\|_2 \leq \tau_0$, we have $\|(A^*q_k)'(x)\|_2 \geq \|(A^*q^\star)'(x)\|_2 - \gamma\tau > 0$.

The estimate $\|(A^*q^\star)'(x)\|_2 > \gamma\tau$ deserves a slightly more detailed justification than the others. Define $w = x - \xi_i$ and $g(\theta) = \langle(A^*q)'(\xi_i + \theta w), w\rangle$ for $\theta \in (0, 1)$. We may apply the mean value theorem to conclude that

$$g(1) - g(0) = g'(\hat\theta) = \left\langle (A^*q)''(\xi_i + \hat\theta w)w, w\right\rangle$$

for some $\hat\theta \in (0, 1)$. Since $g(0) = \langle(A^*q^\star)'(\xi_i), w\rangle = \langle 0, w\rangle = 0$, and $\left\langle (A^*q^\star)''(\xi_i + \hat\theta w)w, w\right\rangle \leq -\gamma\|w\|_2^2$, due to $(|A^*q^\star|)'' \preccurlyeq -\gamma\,\mathrm{id}$ in $\{x \in \Omega, \|x - \xi_i\|_2 \leq \tau_0\}$, we obtain

$$\|(A^*q^\star)'(x)\|_2 \geq \frac{1}{\|w\|_2}|\langle(A^*q^\star)'(x), w\rangle| = \frac{|g(1)|}{\|w\|_2} \geq \frac{\gamma\|w\|_2^2}{\|w\|_2} > \gamma\tau,$$

since $\|w\|_2 = \|x - \xi_i\|_2 > \tau$ by assumption. The last estimate was the claim (iv).

This implies a number of things. First, any local maximum of $|A^*q_k|$ with $|A^*q_k| \geq 1$ must lie within a distance of τ from the set ξ (since for all other points, we have $|A^*q_k| < 1$ – via (iii) – or $(Aq_k)' \neq 0$ – via (iv)). Since $|A^*q_k|$ is locally concave on the τ_0-neighborhoods of the ξ_i – this follows from (ii) – at most one local extremum furthermore exists in each such neighborhood. This is the claim.

3.5 Fixed grids estimates

In this section, we consider a fixed grid Ω_0 and ask what we need to assume about it in order to guarantee that the set of local maxima of $|A^* q_0(x)|$ is close to true support ξ. We express our result in terms of a geometrical property that we can control, the *width* of the grid $\mathrm{dist}(\Omega_0|\Omega)$.

Theorem 3.8 *Assume that* $\mathrm{dist}(\Omega_0|\Omega) \leq \frac{c_q}{\rho\sqrt{\kappa_{\mathrm{hess}}}}$, *then*

$$\begin{aligned}
\mathrm{dist}(\xi|X_0) &\leq \frac{\kappa_\nabla \sqrt{R\kappa_{\mathrm{hess}}}\rho}{2\gamma} \mathrm{dist}(\Omega_0|\Omega) \\
\|q_0 - q^\star\|_2 &\leq \rho\sqrt{R\kappa_{\mathrm{hess}}}\mathrm{dist}(\Omega_0|\Omega) \\
\inf(\mathcal{P}(\Omega_0)) &\leq \inf\left(\mathcal{P}(\Omega)\right) + \frac{R\kappa_{hess}\rho^2}{2L}\mathrm{dist}(\Omega_0|\Omega)^2
\end{aligned}$$

Proof It is trivial that $\mathrm{dist}(\Omega_0|X_0) \leq \mathrm{dist}(\Omega_0|\Omega)$. Applying Lemma 3.5, we immediately obtain the bound on $\|q_0 - q^\star\|_2$. By the same lemma,

$$\begin{aligned}
\inf(\mathcal{P}(\Omega_0)) = \sup(\mathcal{D}(\Omega_0)) = -f^*(q_0) &\leq -f^*(q^\star) + \frac{R\kappa_{hess}\rho^2}{2L}\mathrm{dist}(\Omega_0|X_0)^2 \\
&= \sup\left(\mathcal{D}(\Omega)\right) + \frac{R\kappa_{hess}\rho^2}{2L}\mathrm{dist}(\Omega_0|\Omega)^2 = \inf\left(\mathcal{P}(\Omega)\right) + \frac{R\kappa_{hess}\rho^2}{2L}\mathrm{dist}(\Omega_0|\Omega)^2.
\end{aligned}$$

In order to obtain the first bound, remark that $\|q_0 - q^\star\|_2 \leq c_q$ and use Proposition 3.7.

Remark 2 *Note that Theorem 3.8 allows to control* $\mathrm{dist}(\xi|X_0)$ *but not* $\mathrm{dist}(X_0|\xi)$. *Indeed each* $x \in X_0$ *is guaranteed to be close to a* ξ_i, *but not every* ξ_i *needs to have a point in* X_0 *closeby. Note however that the bounds on the optimal value indicates that in this case the missed* ξ_i *is not crucial to produce a good candidate for solving the primal problem. We will provide more insight on this, in the case of* f *being strongly convex, in Section 4.*

3.6 Eventual linear convergence rate

In this section, we provide an asymptotic convergence rate for the iterative algorithm. As a follow-up to Remark 2, the proof of convergence relies on the fact that the distances will eventually $\mathrm{dist}(X_k|\xi)$ and $\mathrm{dist}(\xi|X_k)$ become equal. To prove that this is the case is exactly the purpose of the next proposition.

Proposition 3.9 *Let* $B_i = \{x \in \Omega, \|x - \xi_i\|_2 < \tau_0\}$. *There exists a finite number of iterations* N, *such that for all* $k \geq N$, X_k *has exactly* s *points, one in each* B_i. *It follows that* $dist(X_k|\xi) = dist(\xi|X_k)$. *Moreover if* S_k *is the set of active points of* $\mathcal{D}(\Omega_k)$, *that is*

$$S_k = \{z \in \Omega_k \text{ s.t. } |A^* q_k(z)| = 1\},$$

then $S_k \subset \cup_i B_i$ *and for each* i, $B_i \cap S_k \neq \emptyset$.

Proof We first prove that B_i contains a point in S_k. To this end, define the set of measures $\mathcal{M}_- = \{\mu \in \mathcal{M}(\Omega), \exists i \in \{1, \dots, s\}, \mathrm{supp}(\mu) \cap B_i = \emptyset\}$ and

$$J_+ = \min_{\mu \in \mathcal{M}_-} \|\mu\|_{\mathcal{M}} + f(A\mu).$$

By assumption (6), $J_+ > J^\star$. Since $(J(\mu_k))_{k\in\mathbb{N}}$ converges to $J(\mu^\star)$, there exists $k_2 \in \mathbb{N}$ such that $\forall k \geq k_2$, $J(\mu_k) < J_+$. Hence μ_k must for each $1 \leq i \leq s$ have points $z_k^i \in \Omega_k$ such that μ_k has non-zero mass at z_k^i. Consequently, $|A^* q_k(z_k^i)| = 1$, hence, each B_i contains *at least* one point in Ω_k such that $|A^* q_k(z_k^i)| = 1$.

Notice that q_k converges to $q^\star$ by Theorem 3.1. Hence there a finite number of iterations k_1 such that $\|q_k - q^\star\| < c_q$ for all $k \geq k_1$. By item (iii) of the proof of Proposition 3.7, $|A^* q_k| < 1$ outside $\cup_i B_i$, and by item (ii), $|A^* q_k|$ is strictly concave in each B_i. Hence each B_i contains exactly one maximizer of $|A^* q_k|$ exceeding one.

We now move on to analyzing our exchange approach. Before formulating the main result, let us introduce a term: *δ-regimes.*

Definition 1 *We say that the algorithm enters a* δ-regime *at iteration* k_δ *if for all* $k \geq k_\delta$*, we have* $\operatorname{dist}(\xi|X_k) \leq \delta$*. In particular it means that only points with a distance at most* δ *from* ξ *are added to the grid.*

Lemma 3.10 *Let* $\bar{\tau}_0 = \frac{\kappa_\nabla}{\gamma} c_q$ *and* $A = 2^{d+1} d^{d/2} \left(\frac{\rho\sqrt{R\kappa_{hess}}\kappa_\nabla}{\gamma} \right)^{3d}$*. Let* N *be as in Proposition 3.9.*

1. *For any* τ*, the algorithm enters a* τ-regime *after a finite number of iterations.*
2. *Assume that* N *iterations have passed and that the algorithm is in a* τ*-regime with* $\tau \leq \bar{\tau}_0$*. Then for every* $\alpha \in (0,1)$ *it takes no more than* $\left\lceil \frac{A}{\alpha^{2d}} \right\rceil + 1$ *iterations to enter an* $\alpha\tau$*-regime.*

Proof Note that for any $\delta \leq \bar{\tau}_0$, if there exists $p \in \mathbb{N}$ such that

$$\|q_k - q^\star\|_2 \leq \frac{\gamma}{\kappa_\nabla}\delta \quad \text{for all } k \geq p, \tag{24}$$

we will enter an δ-regime after iteration p by applying Proposition 3.7.

To prove (1), note that we without loss of generality can assume that $\tau \leq \bar{\tau}_0$ (since entering a τ-regime means in particular entering a τ'-regime for any $\tau' \geq \tau$.) Then , since $\|q_k - q^\star\|_2$ tends to zero as k goes to infinity, (24) with $\delta = \tau$ is true after a finite number of iterations.

To prove (2), we proceed as follows : Proposition 3.9 ensures that in each iteration, exactly one point is added in each ball $\{x \in \Omega, \|x - \xi_i\|_2 \leq \tau\}$. Let k_0 be the actual iteration, a covering number argument [32] ensures, for any Δ that after $\delta_0 = \left\lceil 2d^{d/2} \left(\frac{\tau}{\Delta}\right)^d \right\rceil$ iterations, each point in X_k needs to lie at a distance at most Δ from Ω_k, i.e., $\operatorname{dist}(\Omega_k|X_k) \leq \Delta$.

Now, if we choose $\Delta = \left(\frac{\gamma}{\kappa_\nabla \rho\sqrt{R\kappa_{hess}}}\right)^3 \frac{\alpha^2\tau}{2}$, Lemma 3.5 together with Proposition 3.7 imply

$$\operatorname{dist}(\Omega_{k_0+\delta_0+1}|\xi) \leq \operatorname{dist}(X_{k_0+\delta_0}|\xi) \leq \frac{\kappa_\nabla}{\gamma}\rho\sqrt{R\kappa_{hess}}\operatorname{dist}(\Omega_{k_0+\delta_0}|X_{k_0+\delta_0}) \leq \left(\frac{\gamma\alpha}{\kappa_\nabla\rho}\right)^2 \frac{\tau}{2R\kappa_{hess}}$$

Since $\Omega_{k+1} \subset \Omega_k$ for all k, the distance $\operatorname{dist}(\Omega_k|\xi)$ is non-increasing. As a result $\operatorname{dist}(\Omega_k|\xi) \leq \left(\frac{\gamma\alpha}{\kappa_\nabla\rho}\right)^2 \frac{\tau}{2R\kappa_{hess}}$ for all $k \geq k_0 + \delta_0 + 1$. Since we are in τ-regime, we know that $\operatorname{dist}(X_k|\xi) \leq \tau$ and $\operatorname{dist}(\Omega_k|\xi) \leq \tau$. Hence we can apply Lemma 3.6 to obtain that

$$\|q_k - q^\star\|_2 \leq \sqrt{2R\kappa_{\text{hess}}\tau \cdot \operatorname{dist}(\Omega_k|\xi)}\rho \leq \frac{\gamma}{\kappa_\nabla}\alpha\tau.$$

Then inequality (24) is satisfied with $\delta = \alpha\tau$ and the algorithm enters a $\alpha\tau$-regime.

The main result will tell us how many iterations we need to enter a τ-regime.

Theorem 3.11 *Let* $\tau \leq \bar{\tau}_0 \overset{\text{def.}}{=} \frac{\kappa_\nabla}{R\gamma} c_q$ *and* k_0 *be the iteration on which the algorithm enters a* $\bar{\tau}_0$*-regime. Then* $k_0 < \infty$*, and the algorithm will enter a* τ*-regime after no more than* $k_0 + k_\tau$ *iterations, where*

$$k_\tau := \left\lceil e2^{d+1} d^{d/2} \left(\frac{\rho\sqrt{R\kappa_{\text{hess}}}\kappa_\nabla}{\gamma} \right)^{3d} + 1 \right\rceil \left\lceil 2d \log\left(\frac{\bar{\tau}_0}{\tau}\right) \right\rceil.$$

Additionally, we will have

$$\begin{aligned} \|q_k - q_*\|_2 &\leq \tau\sqrt{2R\kappa_{\text{hess}}}\rho \\ \inf\left(\mathcal{P}(\Omega_k)\right) &\leq \inf\left(\mathcal{P}(\Omega)\right) + \frac{2R\kappa_{\text{hess}}\rho^2}{L} \cdot \tau^2 \end{aligned} \tag{25}$$

for $k \geq k_0 + k_\tau + 1$*. In other words, the algorithm will eventually converge linearly.*

Proof The fact that $k_0 < \infty$ is the first assertion of Lemma 3.10. As for the other part, we argue as follows: Fix $\alpha \in (0,1)$. Since we have entered a $\bar{\tau}_0$-regime at iteration k_0, Lemma 3.10 implies that it will take no more than $\left\lceil \frac{A}{\alpha^{2d}} \right\rceil + 1$ additional iterations to enter a $\alpha\bar{\tau}_0$. Repeating this argument, we see that after no more than

$$n \cdot \left(\left\lceil \frac{A}{\alpha^{2d}} \right\rceil + 1 \right)$$

iterations, we will have entered a $\alpha^n \bar{\tau}_0$ regime. Choosing $\alpha = e^{-1/2d}$ and $n = \lceil 2d \log(\bar{\tau}_0/\tau) \rceil$, we obtain the first statement.

The second statement immediately follows from Lemma 3.6 (as in the proof of Theorem 3.8) and the fact that entering a τ-regime exactly amounts to that $\operatorname{dist}(X_k|\xi) \leq \tau$ for all future k, and therefore in particular $\operatorname{dist}(\Omega_{k+1}|\xi) \leq \tau$.

Remark 3 *Let us give some insights on Theorem 3.11.*

1. *Notice that the value k_τ depends exponentially on the ambient dimension d. This property cannot be improved with the current proof based on a covering number argument. We are unsure as if the exponential growth really is an artefact of the proof, or if it can be removed.*
2. *A popular variant of the algorithm consists in adding the single most violating maximizer, which can then be regarded as a variant of the conditional gradient descent. It is yet unclear whether the current proof can be adapted to this setting since our proof relies on systematically adding one point around every Dirac mass of the solution. We however believe that adding* all *the violating maximizers arguably makes more sense from a computational point of view. Indeed, all violating maximizers have to be explored to select the global maximizer. Hence some information is lost by adding only one point. For instance, in the context of super-resolution imaging, we will see that a variant of the proposed algorithm converges in a single iteration, while a similar variant of the conditional gradient would require s iterations.*
3. *An alternative proof covering the case of adding a single point and removing some was proposed in a work produced independently and roughly at the same time by Pieper and Walter [22]. In there, the authors consider a similar but more general framework allowing for vector valued total variation regularizers. Under an additional assumption of strong convexity of f, the authors also prove an eventual linear convergence rate. The proofs share a few similarities, but also some differences reflected by the additional assumption. In particular, the covering number argument does not appear. It is currently unclear to the authors which proof leads to the better rate.*

On a practical level, the algorithm contains two main difficulties: i) computing the dual solution q_k and ii) finding the local maximizers of $|A^* q_k|$. As for i), the Lipschitz continuity assumption on ∇f makes the dual problem strongly convex. This is a helpful feature that allows to certify the precision of iterative algorithms: we can generate points $\tilde{q}_k$ within a prescribed distance to the actual solution q_k. With some additional work, this could most probably lead to certified algorithms with an inexact resolution of the duals $\mathcal{D}(\Omega_k)$. Point ii) is arguably more problematic: unless the measurement functions a_i have a specific structure such as polynomials, certifying that the maximizers X_k are well evaluated is out of reach. Unfortunately, forgetting points in X_k can break the convergence to the actual solution. In practice, this evaluation proved to require some attention, but well designed particle flow algorithms initialized with a sufficiently large amount of particles seemed to solve any instance of the super-resolution experiments provided later.

The inequality (25) is an upper-bound on the cost function for the problem $(\mathcal{P}(\Omega_k))$. Unfortunately, the numerical resolution of this problem is hard since Ω_k contains clusters of points and in practice it is beneficial to solve the simpler discrete problem

$$\widehat{\mu}_k = \operatorname*{argmin}_{\mu \in \mathcal{M}(X_k)} \|\mu\|_{\mathcal{M}} + f(A\mu) \qquad (\mathcal{P}(X_k))$$

For this measure, we also obtain an a posteriori estimate of the convergence rate.

Proposition 3.12 *Define $\widehat{\mu}_k$ as the solution of $(\mathcal{P}(X_k))$, if $\operatorname{dist}(X_k|\xi) \leq \tau$, we have*

$$J(\widehat{\mu}_k) \leq J(\mu^\star) + \left(\|\alpha^\star\|_1 \frac{\kappa_{\mathrm{hess}} \|q^\star\|_2}{2} + \frac{L}{2} \|\alpha^\star\|_1^2 \kappa_\nabla^2 \right) \tau^2. \qquad (26)$$

Proof For any i, denote x_k^i a point in X_k closest to ξ_i and define $\tilde{\mu}_k = \sum_{i=1}^s \alpha_i^\star \delta_{x_k^i}$. We have $J(\widehat{\mu}_k) \leq J(\tilde{\mu}_k)$ and $\|\tilde{\mu}_k\|_{\mathcal{M}} \leq \|\mu^\star\|_{\mathcal{M}}$. Furthermore, we have

$$f(A\tilde{\mu}_k) \leq f(A\mu^\star) + \langle \nabla f(A\mu^\star), A\tilde{\mu}_k - A\mu^\star \rangle + \frac{L}{2}\|A\tilde{\mu} - A\mu^\star\|_2^2.$$

The last term in the inequality is dealt with the following estimate:

$$\|A\tilde{\mu} - A\mu^\star\|_2 \leq \sum_{i=1}^s |\alpha_i^\star| \|A(x_k^i) - A(\xi_i)\|_2 \leq \sum_{i=1}^s |\alpha_i^\star| \kappa_\nabla \|x_k^i - \xi_i\|_2 \leq \|\alpha^\star\|_1 \kappa_\nabla \tau.$$

As for the penultimate term, remember that $q^\star = -\nabla f(A\mu^\star)$. This implies

$$\langle \nabla f(A\mu^\star), A\tilde{\mu}_k - A\mu^\star \rangle = \langle A^* q^\star, \mu^\star - \tilde{\mu}_k \rangle = \sum_{i=1}^s \alpha_i^\star \left((A^* q^\star)(\xi_i) - A^* q^\star(x_k^i) \right)$$

By making a Taylor expansion of $A^* q^\star$ in each ξ_i, utilizing that the derivative vanishes there, and that $\|(A^* q^\star)''(x)\| \leq \kappa_{\text{hess}} \|q^\star\|_2$ for each $x \in \Omega$, we see that $\left|(A^* q^\star)(x_k^i) - (A^* q^\star)(\xi_i)\right| \leq \frac{\kappa_{\text{hess}}\|q^\star\|_2}{2} \|x_k^i - \xi_i\|_2^2$ for each i. This yields

$$\langle \nabla f(A\mu^\star), A\tilde{\mu}_k - A\mu^\star \rangle \leq \|\alpha^\star\|_1 \frac{\kappa_{\text{hess}} \|q^\star\|_2 \tau^2}{2}.$$

Overall, we obtain

$$J(\widehat{\mu}_k) \leq J(\tilde{\mu}_k) = \|\tilde{\mu}_k\|_{\mathcal{M}} + f(A\tilde{\mu}_k) \leq J(\mu^\star) + \|\alpha^\star\|_1 \frac{\kappa_{\text{hess}} \|q^\star\|_2 \tau^2}{2} + \frac{L}{2} \|\alpha^\star\|_1^2 \kappa_\nabla^2 \tau^2.$$

4 Convergence of continuous methods

In this section, we study an alternative algorithm that consists of using nonlinear programming approaches to minimize the following finite dimensional problem:

$$G(\alpha, X) \overset{\text{def.}}{=} J\left(\sum_{i=1}^p \alpha_i \delta_{x_i} \right) = \|\alpha\|_1 + f\left(A\left(\sum_i \alpha_i \delta_{x_i} \right) \right), \tag{27}$$

where $X = (x_1, \ldots, x_p)$. This principle is similar to continuous methods in semi-infinite programming [25] and was proposed specifically for total variation minimization in [2,10,30,7]. By Proposition 3.9, we know that after a finite number of iterations, X_k will contain exactly s points located in a neighborhood of ξ. This motivates the following hybrid algorithm:

- Launch the proposed exchange method until some criterion is met. This yields a grid $X^{(0)} = X_k$ and we let $p = |X_k|$.
- Find the solution of the finite convex program

$$\alpha^{(0)} = \min_{\alpha \in \mathbb{R}^p} G(\alpha, X^{(0)}).$$

- Use the following gradient descent:

$$(\alpha^{(t+1)}, X^{(t+1)}) = (\alpha^{(t+1)}, X^{(t+1)}) - \tau \nabla G(\alpha^{(t)}, X^{(t)}), \tag{28}$$

  where τ is a suitably defined step-size (e.g. defined using Wolfe conditions).

We tackle the following question: does the gradient descent algorithm converge to the solution if initialized well enough?

4.1 Existence of a basin of attraction

This section is devoted to proving the existence of a basin of attraction of a descent method in G. Under two additional assumptions, we state our result in Proposition 4.1.

Assumption 7 *The function f is twice differentiable and Λ-strongly convex.*

The twice differentiability assumption is mostly due to convenience, but the strong convexity is crucial. The second assumption is related to the structure of the support ξ of the solution $\mu^\star$.

Assumption 8 *For any $x, y \in \Omega$ denote $K(x,y) = \sum_\ell a_\ell(x) a_\ell(y)$. The* transition matrix

$$T(\xi) = \begin{bmatrix} [K(\xi_i,\xi_j)]_{i,j=1}^s & [\nabla_x K(\xi_i,\xi_j)^*]_{i,j=1}^s \\ [\nabla_x K(\xi_i,\xi_j)]_{i,j=1}^s & [\nabla_x \nabla_y K(\xi_i,\xi_j)^*]_{i,j=1}^s \end{bmatrix} \in \mathbb{R}^{s+sd, s+sd}.$$

is assumed to be positive definite, with a smallest eigenvalue larger than $\Gamma > 0$.

It is again possible to prove for many important operators A that this assumption is satisfied if the set ξ is separated. See the references listed in the discussion about Assumption 6. The following proposition describes the links between minimizing G and solving $(\mathcal{P}(\Omega))$.

Proposition 4.1 *Let $\mu^\star = \sum_{i=1}^s \alpha_i^\star \delta_{\xi_i} \neq 0$ be the solution of $(\mathcal{P}(\Omega))$. Under Assumption 7 and 8, $(\alpha^\star, \xi)$ is the global minimum of G. Additionally, G is differentiable with a Lipschitz gradient and strongly convex in a neighborhood of $(\alpha^\star, \xi)$.*

Hence, there exists a basin of attraction around $(\alpha^\star, \xi)$ such that performing a gradient descent on G will yield the solution of $(\mathcal{P}(\Omega))$ at a linear rate.

The rest of this section is devoted to the proof of Proposition 4.1. Let us begin by stating a simple auxiliary result.

Lemma 4.2 *Let U and V be vector spaces and $C : V \to V$ be a linear operator with $C \succcurlyeq \lambda \mathrm{id}_V$ for a $\lambda \geq 0$. Then, for any $B : U \to V$*

$$B^* C B \succcurlyeq \lambda B^* B.$$

Proof If $B^*CB - \lambda B^*B$ is positive semidefinite, the claim holds. Since for $v \in U$ arbitrary

$$\langle (B^*CB - \lambda B^*B)v, v\rangle = \langle C(Bv), Bv\rangle - \lambda \langle Bv, Bv\rangle \geq \lambda \|Bv\|_V^2 - \lambda \|Bv\|_V^2 = 0,$$

the former is the case.

Let us introduce some notation that will be used in this section: for an $X = (x_1, \dots, x_p) \in \Omega^p$ for some p, $A(X)$ denotes the matrix $[a_i(x_j)]$. Analogously, $A'(X)$ and $A''(X)$ denote the operators

$$A'(X) : (\mathbb{R}^d)^p \to \mathbb{R}^m, (v_i)_{i=1}^p \mapsto \left(\sum_{i=1}^p \partial_x a_j(x_i) v_i \right)_j, \quad A''(X) : (\mathbb{R}^d \times \mathbb{R}^d)^p \to \mathbb{R}^m, (v_i, w_i)_{i=1}^p \mapsto \sum_{i=1}^p A''(x_i)[v_i, w_i]$$

respectively. Note that for $q \in \mathbb{R}^m$ and $X \in \Omega^p$,

$$\begin{aligned} A(X)^* q &= ((A^*q)(x_i))_{i=1}^p \overset{\text{def.}}{=} (A^*q)(X) \in \mathbb{R}^p \\ A'(X)^* q &= (\nabla(A^*q)(x_1), \dots, \nabla(A^*q)(x_p)) \in (\mathbb{R}^d)^p \\ A''(X)^* q &= ((A^*q)''(x_1), \dots, (A^*q)''(x_p)) \in (\mathbb{R}^d \times \mathbb{R}^d)^p \end{aligned}$$

We will also use the shorthands $\mu = \sum_i \alpha_i \delta_{x_i}$, $G_f(\alpha, X) = f(A\mu)$, and, for $\alpha \in \mathbb{R}^p$, $D(\alpha)$ denotes the operator

$$D(\alpha) : (\mathbb{R}^d)^p \to (\mathbb{R}^d)^p, (v_i)_{i=1}^p \mapsto (\alpha_i v_i)_{i=1}^p.$$

We have

$$\frac{\partial G_f}{\partial \alpha}(\alpha, X)\beta = \langle \nabla f(A\mu), A(X)\beta \rangle$$
$$\frac{\partial G_f}{\partial X}\delta = \langle \nabla f(A\mu), A'(X)D(\alpha)\delta \rangle ,$$

so that in points (α, X) with $\alpha_i \neq 0$ for all i, and in particular in a neighborhood of $(\alpha^\star, \xi)$, G is differentiable and its gradient is given by :

$$\mathbb{R}^p \times (\mathbb{R}^p)^d \ni \nabla G(\alpha, X) = (\operatorname{sign}(\alpha) - (A^* q)(X), -D(\alpha)(A^* q)'(X)) , \quad \text{with } q = -\nabla f(A\mu). \tag{29}$$

As for the second derivatives, we have

$$\frac{\partial^2 G_f}{\partial^2 \alpha}(\alpha, X)[\beta, \gamma] = f''(A\mu)(A(X)\beta, A(X)\gamma)$$
$$\frac{\partial^2 G_f}{\partial \alpha \partial X}(\alpha, X)[\beta, \delta] = f''(A\mu)(A(X)\beta, A'(X)D(\alpha)\delta) + \langle \nabla f(A\mu), A'(X)D(\beta)\delta \rangle$$
$$\frac{\partial^2 G_f}{\partial^2 X}(\alpha, X)[\delta, \epsilon] = f''(A\mu)(A'(X)D(\alpha)\delta, A'(X)D(\alpha)\epsilon) + \langle \nabla f(A\mu), A''(X)(D(\alpha)\delta, \epsilon) \rangle .$$

We may now prove our claims.

Proof (Proof 4.1) First, let us note that due to the optimality conditions of $\mathcal{P}(\Omega)$, we know that

$$q^\star = -\nabla f(A\mu^\star).$$

Now, $|A^* q^\star|$ has local maxima in the points ξ_i, so that $(A^* q^\star)'(\xi) = 0$. In these points, we furthermore have that $\operatorname{sign}(\alpha_i^\star) = A^* q^\star(\xi_i)$, so that the gradient of G given in (29) vanishes.

To prove the rest, it is enough to show that the Hessian of G_f is positive definite in a neighborhood around $(\alpha^\star, \xi)$. For this, it is fruitful to decompose it into two parts. Letting $q = -\nabla f(A\mu)$, we have $G_f'' = H_1 + H_2$, with

$$H_1(\alpha, X) = \begin{bmatrix} A(X)^* f''(A\mu)A(X) & A(X)^* f''(A\mu)A'(X)D(\alpha) \\ D(\alpha)^* A'(X)^* f''(A\mu)A(X) & D(\alpha)^* A'(X)^* f''(A\mu)A'(X)D(\alpha) \end{bmatrix}$$
$$H_2(\alpha, X)[(\beta, \delta), (\gamma, \epsilon)] = -\sum_{i=1}^{s} \beta_i (A^* q)'(x_i)\epsilon_i + \gamma_i (A^* q)'(x_i)\delta_i + \alpha_i (A^* q)''(x_i)[\delta_i, \epsilon_i],$$

Let (α, X) be arbitrary. H_1 is an operator of the form $M_1^* M_2(X)^* \mathcal{L} M_2(X) M_1$, with $\mathcal{L} = f''(A\mu) : \mathbb{R}^m \to \mathbb{R}^m$ and

$$M_1 = \begin{bmatrix} \mathrm{id} & 0 \\ 0 & D(\alpha) \end{bmatrix} : \mathbb{R}^p \times (\mathbb{R}^d)^s \to \mathbb{R}^s \times (\mathbb{R}^d)^s, \quad M_2(X) = \begin{bmatrix} A(X) & A'(X) \end{bmatrix} : \mathbb{R}^s \times (\mathbb{R}^d)^s \to \mathbb{R}^m.$$

Due to the Λ-strong convexity of f, $\mathcal{L} \succcurlyeq \Lambda \,\mathrm{id}$. We furthermore have

$$M_1^* M_1 = \begin{bmatrix} \mathrm{id} & 0 \\ 0 & D(\alpha)^* D(\alpha) \end{bmatrix} \succcurlyeq \min_{1 \leq i \leq n} |\alpha_i|^2 \cdot \mathrm{id} \succcurlyeq \frac{\min_{1 \leq i \leq n} |\alpha_i^\star|^2}{2} \cdot \mathrm{id}$$

in some neighborhood U of $\alpha^\star \neq 0$.

Let us now turn to $M_2(X)^* M_2(X)$. If we define $M_2(\xi) = \begin{bmatrix} A(\xi) & A'(\xi) \end{bmatrix}$, we have

$$M_2(\xi)^* M_2(\xi) = \begin{bmatrix} A(\xi)^* A(\xi) & A(\xi) A'(\xi)^* \\ A'(\xi)^* A(\xi) & A'(\xi)^* A'(\xi)^* \end{bmatrix} = T(\xi) \succcurlyeq \Gamma \,\mathrm{id}$$

by Assumption (8). Since, by assumption (5), both $A(X)$ and $A'(X)$ are continuously dependent on X, we even have

$$M_2^*(X) M_2(X) \geq \frac{\Gamma}{2}$$

for X in some neigborhood V of ξ. We may now apply Lemma 4.2 twice to conclude

$$H_1(\alpha, X) \succcurlyeq \frac{\Lambda\Gamma \min_{1\leq i\leq n} |\alpha_i^\star|^2}{4} \,\mathrm{id} \tag{30}$$

for $(\alpha, X) \in U \times V$.

It remains to analyze H_2. We again begin by evaluating the expression in $(\alpha^\star, \xi)$. The assumption (6) implies that

$$\begin{aligned}(A^*q)'(x_i) &= 0\\ \alpha_i(A^*q)''(x_i) &\preccurlyeq 0\end{aligned}$$

for each i. We therefore obtain

$$\begin{aligned}H_2(\alpha^\star, \xi)[(\beta, \delta), (\beta, \delta)] &= -\sum_{i=1}^{s} \beta_i(A^*q)'(\xi_i)\delta_i + \beta_i(A^*q)'(\xi_i)\delta_i + \alpha_i(A^*q)''(\xi_i)[\delta_i, \delta_i]\\ &= -\sum_{i=1}^{s} \alpha_i(A^*q)''(x_i)[\delta_i, \delta_i] \geq 0\end{aligned}$$

Hence, the bidual form $H_2(\alpha^\star, \xi)$ is positive semidefinite. Due to the assumptions that the measurement functions a_i are members of $\mathcal{C}_0^2$, and that ∇f is Lipschitz continuous, $\mathcal{H}_2$ depends continuously on α and x. Consequently,

$$\|H_2(\alpha, X)\| \leq \frac{\Lambda\Gamma \min_{1\leq i\leq n} |\alpha_i^\star|^2}{8} \tag{31}$$

for (α, X) in some neighborhood W of $(\alpha^\star, \xi)$.

Combining (30) and (31), we obtain

$$H_1(\alpha, X) + H_2(\alpha, X) \succcurlyeq \frac{\Lambda\Gamma \min_{1\leq i\leq n} |\alpha_i^\star|^2}{8} \,\mathrm{id}$$

for all $(\alpha, X) \in (U \times V) \cap W$, which was to be proven.

4.2 Eventually entering the basin of attraction

The following proposition shows that $(\tilde{\alpha}, X_k)$ defined as the amplitudes and positions of the Dirac-components of the solution $\widehat{\mu}$ of $(\mathcal{P}(X_k))$, $(\tilde{\alpha}, X_k)$ will lie in the basin described by Proposition 4.1. This result is stated in Corollary 4.4, the rest of this section is dedicated to proving it.

Proposition 4.3 *Assume that Assumptions 7 and 8 are true. Consider an s-sparse measure*

$$\tilde{\mu} = \sum_{\ell=1}^{s} \tilde{\alpha}_\ell \delta_{\tilde{x}_\ell}$$

for some $\tilde{\alpha} \in \mathbb{R}^s$ and $(\tilde{x}_\ell)_{\ell=1\ldots s}$ pairwise different points of Ω. We then have

$$\|\tilde{\alpha} - \alpha^\star\|_2 \leq \frac{1}{\sqrt{\Gamma}} \left(\kappa_\nabla \|\tilde{\mu}\|_{\mathcal{M}} \sup_{1\leq \ell\leq s} \|\xi_\ell - \tilde{x}_\ell\|_2 + \sqrt{\frac{2}{\Lambda}\left(J(\tilde{\mu}) - J(\mu^\star)\right)} \right).$$

Proof Let $A(\xi)^\dagger$ be the Moore-Penrose inverse of $A(\xi) = [A(\xi_1), \ldots, A(\xi_s)]$. Due to Assumption 8, $A(\xi)^\dagger$ has full rank and has an operator norm no larger than $\Gamma^{-1/2}$. Since

$$\tilde{\alpha} = \alpha^\star + A(\xi)^\dagger(A(\xi)\tilde{\alpha} - A\tilde{\mu}) + A(\xi)^\dagger(A\tilde{\mu} - A(\xi)\alpha^\star),$$

bounds on $A(\xi)\tilde{\alpha} - A\tilde{\mu}$ and $A\tilde{\mu} - A(\xi)\alpha^\star$ will therefore transform to a bound on $\tilde{\alpha} - \alpha^\star$.

Let us begin with the former. We have

$$\|A(\xi)\tilde{\alpha} - A\tilde{\mu}\|_2 \leq \sum_{\ell=1}^{s} |\tilde{\alpha}_\ell| \, \|A(\xi_\ell) - A(\tilde{x}_\ell)\| \leq \sum_{\ell=1}^{s} \kappa_\nabla \, |\tilde{\alpha}_\ell| \, \|\xi_\ell - \tilde{x}_\ell\|_2 = \kappa_\nabla \|\tilde{\alpha}\|_1 \sup_{\substack{1\leq \ell \leq s \\ \tilde{\alpha}_\ell \neq 0}} \|\xi_\ell - \tilde{x}_\ell\|_2,$$

where we used the Cauchy-Schwarz inequality in the last step.

To bound the latter, recall that Λ-strong convexity of f means that

$$f(A\tilde{\mu}) \geq f(A\mu^\star) + \langle \nabla f(A\mu^\star), A\tilde{\mu} - A\mu^\star \rangle + \frac{\Lambda}{2} \|A\tilde{\mu} - A\mu^\star\|_2^2. \tag{32}$$

The optimality conditions for $(\mathcal{P}(\Omega))$ tell us that $q^\star = -\nabla f(A\mu^\star)$, and hence

$$\langle \nabla f(A\mu^\star), A\tilde{\mu} - A\mu^\star \rangle = \langle A^* q^\star, \mu^\star - \tilde{\mu}) \rangle = \sum_{\ell=1}^{s} \alpha_\ell^\star (A^* q^\star)(\xi_\ell) - \tilde{\alpha}_\ell (A^* q^\star)(\tilde{x}_\ell) \geq \|\alpha^\star\|_1 - \|\tilde{\alpha}\|_1,$$

where we in the last step used that $\|A^* q^\star\|_\infty \leq 1$. Plugging the above inequality in (32) yields

$$\frac{\Lambda}{2} \|A\tilde{\mu} - A\mu^\star\|_2^2 \leq J(\tilde{\mu}) - J(\mu^\star).$$

The claim follows.

Corollary 4.4 *By Proposition 3.9, if k is large enough then X_k contains exactly s points. In this case, let $\widehat{\mu}_k = \sum_{i=1}^{s} \widehat{\alpha}_i \delta_{\hat{x}_i^k}$ be the solution of $(\mathcal{P}(X_k))$. Applying Proposition 4.3, recalling that $\max_i \|\xi_i - \hat{x}_i^k\|_2 \leq \operatorname{dist}(X_k|\xi)$ and using the bound (26), we obtain :*

$$\|\widehat{\alpha} - \alpha^\star\|_2 \leq \frac{\operatorname{dist}(X_k|\xi)}{\sqrt{\Gamma}} \left(\kappa_\nabla \|\widehat{\mu}_k\|_{\mathcal{M}} + \sqrt{\frac{2}{\Lambda} \left(\|\alpha^\star\|_1 \frac{\kappa_{\mathrm{hess}} \|q^\star\|_2}{2} + \frac{L}{2} \|\alpha^\star\|_1^2 \kappa_\nabla^2 \right)} \right).$$

Since $\operatorname{dist}(X_k|\xi)$ is guaranteed to eventually converge to zero by Theorem 3.11 and $\|\widehat{\mu}_k\|_{\mathcal{M}}$ are bounded (e.g. by lower boundedness of f and upper boundedness of $J(\widehat{\mu}_k)$) , $(\widehat{\alpha}, X_k)$ will eventually lie in the basin of attraction of G.

5 Description of the hybrid approach

To conclude this paper, we propose a method alternating between an exchange step and a continuous gradient descent. It is detailed in Algorithm 2. The idea is, after each iteration of an exchange algorithm, to start a gradient descent of G initialized at the solution $\widehat{\mu}_k$ of $(\mathcal{P}(X_k))$. If this gradient descent converges to a measure $\bar{\mu}_k$, we can subsequently test if it is an optimal point by checking if $\bar{q}_k = -\nabla f(A\bar{\mu}_k)$ fulfills the stopping criterion $\|A^* \bar{q}_k\|_\infty \leq 1 + \epsilon$, where ϵ is a user defined stopping criterion (the latter is justified by Proposition 3.3). If so, we may output $\bar{\mu}_k$, and if not, we may instead continue our exchange algorithm, possibly after adding also the support points of $\bar{\mu}_k$. Its behavior is described in the following theorem.

Theorem 5.1 (Convergence guarantees for the alternating method) *Algorithm 2 comes with the following guarantees:*

1. (Theorem 3.1) *Under Assumptions 1, 2 and 3, it is guaranteed to stop after a finite number of iterations for any stopping criterion $\epsilon > 0$.*
2. (Theorem 3.11) *If in addition Assumptions 5 and 6 are satisfied, then the algorithm eventually converges linearly: $k \geq N + k_\tau$ with $k_\tau \lesssim \log(\tau^{-1})$, we have $\operatorname{dist}(\Omega_k|\xi) \leq \tau$.*
3. (Proposition 4.1, Theorem 3.11 and Proposition 4.3) *If in addition Assumptions 7 and 8 are satisfied, then - for large enough k - the low complexity gradient descent (28) method converges linearly : $\|(\alpha^{(t)}, X^{(t)}) - (\alpha^\star, \xi)\|_2 \leq c^t \|(\alpha^{(0)}, X^{(0)}) - (\alpha^\star, \xi)\|_2$ for some $0 \leq c < 1$.*

Overall, this method has many desirable properties: the continuous method should be used whenever the exchange method reaches its basin of attraction since its per iteration cost is much cheaper. However, it is unclear in general that this basin even exists. In that case, the exchange method should be preferred since it eventually converges linearly under quite mild assumptions. The proposed algorithmic scheme somehow captures the best of all methods. Let us notice that it is very similar in spirit to the sliding Frank-Wolfe algorithm proposed in [10], apart from the fact that we suggest adding *all* the points X_k violating the constraints, while the single most violating point is added in [10]. We believe that the proposed analysis sheds some light on the good numerical performance of this method.

Arguably the most complicated step in this algorithm is to evaluate X_k, the set of local maximizers of A^*q_k exceeding 1. This is an impossible task for an arbitrary function A^*q_k. However, a simple heuristic described in the next section provided rather satisfactory results for the measurement functions considered in this paper (trigonometric polynomials and Gaussian convolution).

Apart from this, let us outline that the subproblems in this algorithm are well suited for numerical resolution. In the exchange algorithm, we only solve the dual problems $\mathcal{D}(\Omega_k)$ which are strongly convex. Hence first-order methods for instance come with guarantees of convergence to q_k in ℓ^2-norm. Recovering the masses $\hat{\alpha}_k$, solutions of $\mathcal{P}(X_k)$ is also stable since X_k (the local maximizers of A^*q_k) is typically a well separated set of low cardinality. The gradient descent (or alternative nonlinear programming approach) on $G(\alpha, X)$ is performed over a low dimensional set. If the convergence is not satisfactory (e.g. the norm of ∇G doesn't decay fast enough), it can be stopped, and we can switch back to the exchange algorithm.

Algorithm 2 Alternating method

Input: Operator A, data fitting term f, stopping criterion $\epsilon > 0$.
Set $q_0 = 0$, $k = 0$, $\Omega_0 = \emptyset$
Evaluate X_0 in 5 and $\|A^*q_0\|_\infty$ ▷ Nonconvex - Possibly complicated
while $\|A^*q_k\|_\infty > 1 + \epsilon$ **do**
    $k = k + 1$
    Set $\Omega_k = \Omega_{k-1} \cup X_k$
    Solve $\mathcal{D}(\Omega_k)$ to retrieve q_k ▷ Convex - Stable
    Evaluate X_k in 5 and $\|A^*q_k\|_\infty$ ▷ Nonconvex - Possibly complicated
    Solve $\mathcal{P}(X_k)$ to retrieve $\widehat{\alpha}_k$ ▷ Convex - Low dimensional
    Gradient descent on $G(\alpha, X)$ in (27) starting from $(\widehat{\alpha}_k, X_k)$ ▷ Nonconvex - Low dimensional
    if Gradient descent converged to $(\bar{\alpha}_k, \bar{X}_k)$ **then**
        Define $q_k = -\nabla f(A\bar{\mu}_k)$ with $\bar{\mu}_k = \sum_{i=1}^{|X_k|} \bar{\alpha}_k(i)\delta_{\bar{X}_k(i)}$
        Evaluate X_k in 5 and $\|A^*q_k\|_\infty$ ▷ Nonconvex - Possibly complicated
        (Optional) Define $\Omega_k = \Omega_k \cup \bar{X}_k$.
    end if
end while
Solve $\mathcal{P}(X_k)$ to retrieve α_k ▷ Convex - Low dimensional
Output: $\mu_k = \sum_{i=1}^{|X_k|} \alpha_k(i)\delta_{X_k(i)}$ and $q_k = -\nabla f(A\mu_k)$.

6 Numerical Experiments

To test our theory, we have implemented our algorithm in MATLAB. Before displaying the results of the experiments, let us discuss a few key steps in the implementation. In the entire section, we assume that $\Omega = [0,1]^d$ for $d = 1$ or 2 for simplicity. Note that this is no true restriction: we can always by scaling and translation ensure that $\Omega \subseteq [0,1]^d$, and trivially extend the measurement functions by 0 to the entirety of $[0,1]^d$.

Evaluating X_k Each iteration of the exchange algorithm requires the exact calculation of the local maximizers of A^*q_k exceeding 1. This is, in general, an impossible task. We resort to the following heuristic method: Given a q_k, we first evaluate $|A^*q_k|$ on a fixed rectangular grid $G = ((n)^{-1}[0,\dots,n])^d$, and determine all of the discrete peaks, i.e. points in which $\{A^*q_k\}$ is larger than all of its neighbors in the grid, and where A^*q_k exceeds $1 - \epsilon_1$ for a threshold $\epsilon_1 > 0$. Next, we start a gradient descent in each of these points, stopping them once $\|(A^*q_k)'\|_2$ is lower than another threshold. Since it is possible that several of these gradient descents land in the same point x, we subsequently check if the set

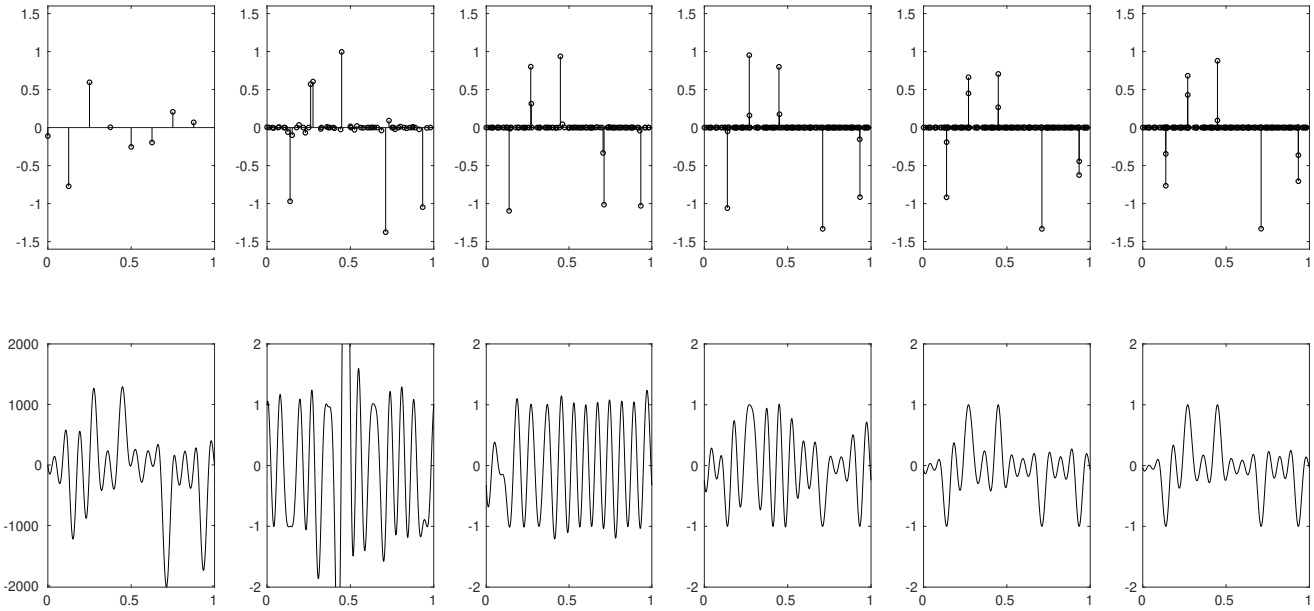


Fig. 1: Above: μ_k for $k = 0, 2, 4, 6, 8, 20$ along one run of the algorithm. Below: A^*q_k for $k = 0, 2, 4, 6, 8, 20$ along the same run. Note that the range of the first plot is different from the others.

contains sets of points which are too close to each other - if this is the case, we discard all but one of them in such a group. We finally remove any point in which $|A^*q_k|$ is not larger than $1 - \epsilon_2$, for a small $\epsilon_2 > 0$.

Solving the Discrete Problems We have chosen to solve the problems $(\mathcal{D}(\Omega_k))$ and $(\mathcal{P}(X_k))$ using an accelerated proximal gradient descent [21].

6.1 Example 1: Super-resolution from Fourier measurements in 1D.

We start by testing our algorithm on a popular instance of problem $(\mathcal{P}(\Omega))$: super-resolution of a measure $\mu \in \mathcal{M}(0, 1)$ from finitely many of its Fourier moments

$$y_k = \langle a_k, \mu \rangle = \int_0^1 \exp(-ikx) d\mu, -m/2 \leq k \leq m/2 - 1.$$

We use a quadratic data fidelity term $f(z) = \frac{L}{2}\|z-y\|_2^2$. This example is well studied by the signal processing community [28, 5, 12, 23].

We chose m to be equal to 30, and a vector y generated as $A\mu_0$, where μ_0 is chosen at random as a 5-sparse atomic measure with amplitudes close to 1 or -1. The positions of the Dirac masses were chosen as a small random perturbation from a uniform grid. The initial grid Ω_0 was chosen as a uniform grid with 8 points, i.e. $[0, \frac{1}{8}, \ldots, \frac{7}{8}]$. We made 100 experiments, with 20 iterations of the exchange algorithm. The evolution of μ_k and q_k for the first iterations for a typical iteration is displayed in Figure 1. We see that after already 8 iterations, A^*q_k appears to be very close to $A^*q^\star$. Before this iteration, the algorithm 'chooses' to add points relatively uniformly to the grid, but after that, new points are only added close to ξ. This is further emphasized by Figure 2, in which X_k is plotted for each iteration, along with size of Ω_k.

To track the success of the algorithm a bit more systematically, we chose to track the evolution of $\mathrm{dist}(\xi|X_k)$, $\mathrm{dist}(\Omega_k|X_k)$ and $\mathrm{dist}(\Omega_k|\xi)$. The median over the 100 iterations, along with confidence intervals covering all experiments but the top and bottom 5% are plotted in Figures 3. We see that all of the quality measures seem to converge linearly to 0.

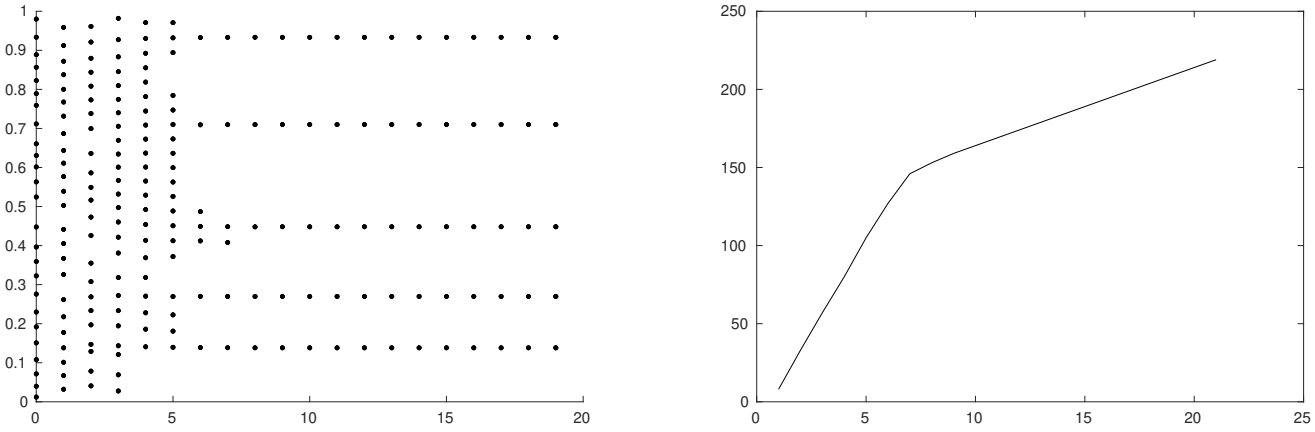


Fig. 2: Left: The set X_k of added points for each iteration along a run of the algorithm. Right: The total number of points in Ω_k along the same run.

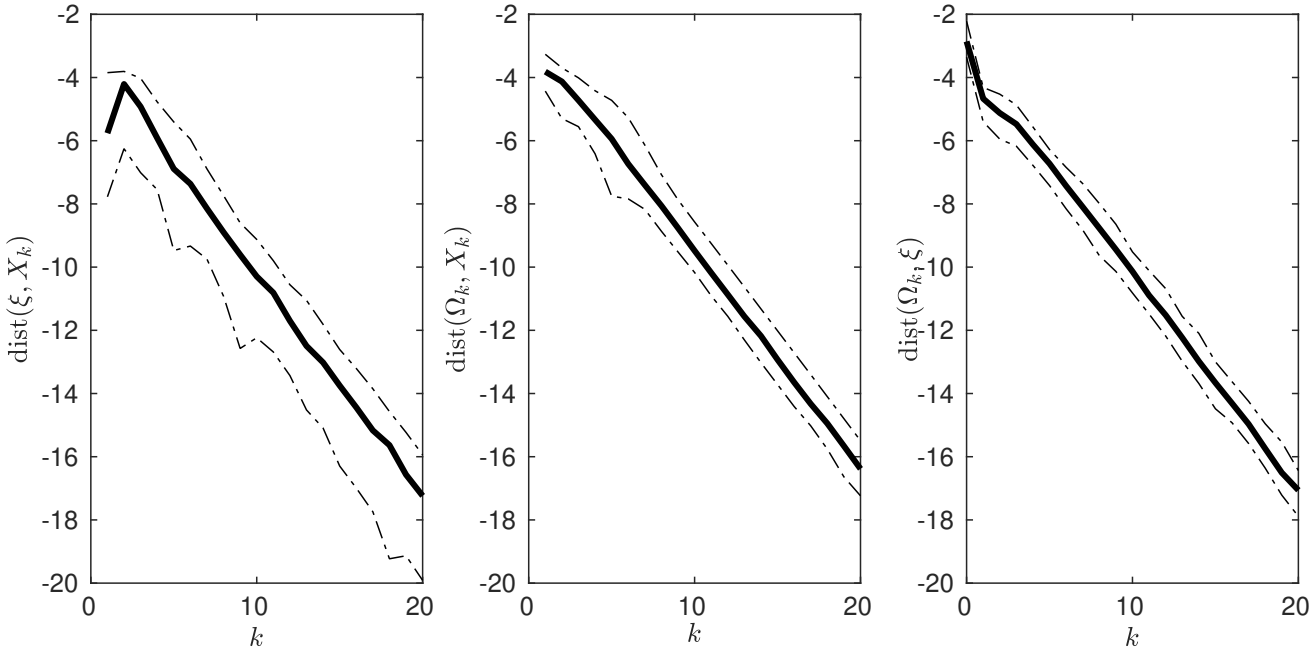


Fig. 3: Logarithmic plot of $\mathrm{dist}(\xi|X_k)$, $\mathrm{dist}(\Omega_k|X_k)$ and $\mathrm{dist}(\Omega_k|\xi)$. Shown is the median value (oblique line) along with confidence intervals(dashed) covering all but the top and lower 5% values.

Finally, we performed the same analysis for the optimum gap $\min(\mathcal{P}(\Omega_k)) - \min(\mathcal{P}(\Omega))$, the error $\|q_k - q^\star\|_2$ and the sizes of the grids Ω_k. ($\min(\mathcal{P}(\Omega))$ was in each case chosen as the lowest value of $\min(\mathcal{P}(\Omega_k))$ over all iterations k, and $q^\star$ as the corresponding dual solution). We see that the optimum gap seems to converge exponentially to 0 right from the first iteration, wheras the error $\|q_k - q^\star\|_2$ initially does not. The 'two-phase'-effect is also easy to spot: After about $5-6$ iterations, the algorithm switches from adding many points to adding only few points close to ξ. Interestingly, the plateau of the q-errors seems to be simultaneuos with the 'phase-transition'.

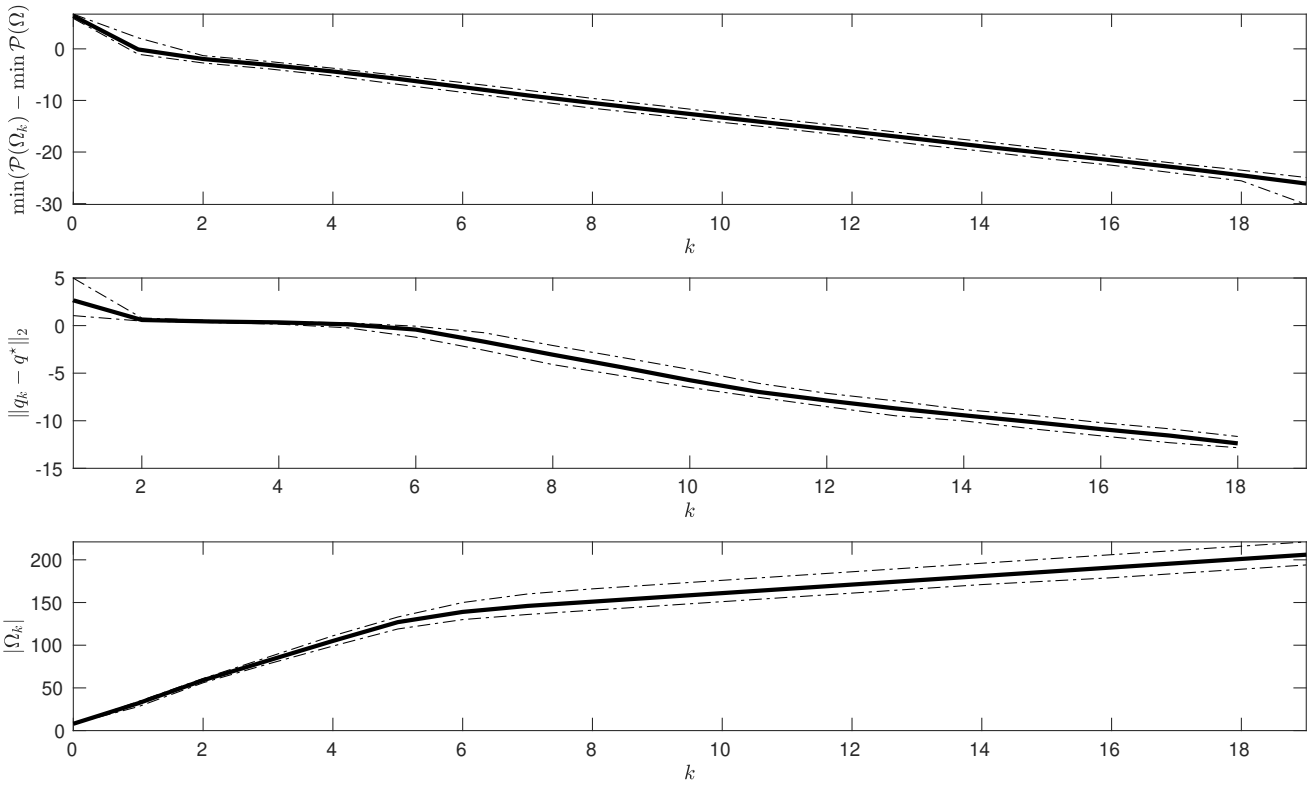


Fig. 4: Plot of the evolution optimum gap, q-error and grid sizes. The top two plots are logarithmic, while the bottom one is not. The oblique lines are represent the median iterations, the dashed ones are confidence intervals covering all but the top and bottom 5% values.

6.2 Example 2: Super-resolution from Gaussian measurements in 2D

Next, we perform a study in a two-dimensional setting. We consider $\Omega = [-1, 1]^2$ and measurement functions of the form

$$a_i(x) = \exp\left(-\frac{\|x - x_i\|^2}{2\sigma^2}\right),$$

where the points x_i live on a Euclidean grid of size 64×64, restricted to the domain $[-0.5, 0.5]^2$. We then add white Gaussian noise to the measurements, leading to pictures of the type shown in Fig. 5. Here, the true underlying measure contains 11 Dirac masses with random positive amplitudes and random locations on $[-0.4, 0.4]^2$.

6.2.1 Exchange algorithm

The evolution of the grids Ω_k and of the dual certificates $|A^* q_k|$ is shown in Figure 6. As can be seen, points are initially added anywhere in the domain, but after a few iterations, they all cluster around the true locations, as expected from the theory. To further stress this phenomenon and illustrate our theorems and lemmata, we display many quantities of interest appearing in our main results in Fig. 7. the distance from X_k to ξ (where ξ is estimated as X_{40}) on Fig. 7c, the distance from Ω_k to ξ on Fig. 7b, the evolution of $J(\widehat{\mu}_k) - J(\widehat{\mu}_{40})$ on Fig. 7a, $\|A^* q_k\|_\infty - 1$ on Fig. 7e. Finally, the number of maxima of $|A^* q_k|$ is shown on Fig. 7f. As can be seen, the number of maxima quickly stabilizes, suggesting that we reached a τ_0-regime. Then all the quantities (cost function, distance from ξ, violation of the constraints) seem to converge to 0 linearly. This is not true after iteration 15, and we suspect that this is solely due to numerical inaccuracies when computing the solution of the discretized problems. Notice however that the accuracy of the Dirac locations

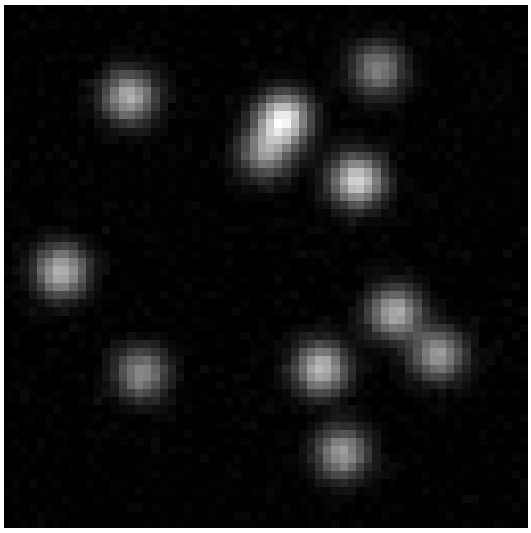


Fig. 5: Measurements y associated to a super-resolution experiment. A sparse measure is convolved with a Gaussian kernel and Gaussian white noise is added.

drops below 10^{-3} after 14 iterations, and that this accuracy is more than enough for the particular super-resolution application. Notice that if we wished to reach this accuracy with a fixed grid, we would need a Euclidean discretization containing 10^6 points, while we here needed only 152 ($|\Omega_{14}| = 152$). In addition, the ℓ^1 resolution is stable since it is accomplished on a grid X_{14} containing only 11 points.

6.2.2 Continuous method

In this experiment, we evaluate the behavior of the gradient descent (28) depending on the initialization $(\alpha^{(0)}, X^{(0)})$ and on the number of iterations. We use the same setting as in the previous section. The left graph of Fig. 8 illustrates that the gradient descent typically converges linearly when initialized close enough to the true minimizer $(\alpha^\star, \xi)$. This was predicted by Theorem 4.1. In this case (and actually all the others related to this experiment), it converges to machine precision in less than 1000 iterations. This is remarkable since the gradient descent is a simple algorithm that can be easily improved by using e.g. Nesterov acceleration (we proved that the function is locally convex) or other optimization schemes such as L-BFGS.

In order to evaluate the size of the basin of attraction around the global minimizer, we start from random points of the form $(\alpha^{(0)}, X^{(0)}) = (\alpha^\star, \xi) + (\Delta_\alpha, \Delta_X)$, where Δ_α and Δ_X are random perturbations with an amplitude set as $\|(\Delta_\alpha, \Delta_X)\|_2 = \gamma \|(\alpha^\star, \xi)\|_2$, with γ in $[0, 1]$. We then run 50 gradient descents with different realizations of $(\alpha^{(0)}, X^{(0)})$ and record the success rate (i.e. the number of times the gradient descent converges to $(\alpha^\star, \xi)$ with an accuracy of at least 10^{-6}). We plot this success rate with respect to γ in Fig. 8b. As can be seen, the success rate is always 1 when the relative error γ is less than 5%, showing that for this particular problem, a rather rough initialization suffices for the gradient descent to converge to the global minimizer.

6.2.3 Alternating method

The alternating method suggested in Algorithm 2 turns out to converge in a single iteration when applied to the setting described above. We therefore apply it to a more challenging scenario with 30 Dirac masses instead of 11 and more noise. The measurements y are shown in Fig. 9. We compare three implementations: a pure exchange method, an alternating method as in Algorithm 2 without line 14 and an alternating method as in in Algorithm 2 with line 14. The conclusions are as follows:

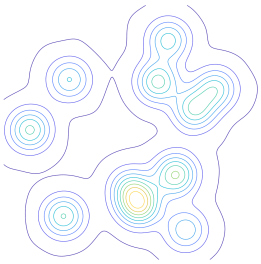
– All methods rapidly conclude that the underlying measure contains 30 Dirac masses. (The pure exchange algorithm after 10 iterations, the alternating method with line 14 already after the first).

– The pure exchange algorithm quickly gets to a point close to the optimum. The positions then slowly converge to the tue locations. It does however eventually find the basin of attraction of G (in this example, it needed 10 iterations).
– Line 14 in the alternating method improves the convergence significantly. In fact, omitting it, we need 10 iterations to find the basin of attraction, whereas the version with the line finds it directly. Investigating this effect more closely is an interesting line of future research.

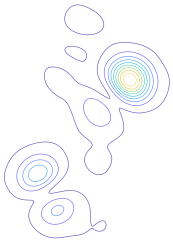
References

1. John M. Borwein and Adrian S. Lewis. Partially finite convex programming, part i: Quasi relative interiors and duality theory. *Mathematical Programming*, 57(1):15–48, May 1992.
2. Nicholas Boyd, Geoffrey Schiebinger, and Benjamin Recht. The alternating descent conditional gradient method for sparse inverse problems. *SIAM Journal on Optimization*, 27(2):616–639, 2017.
3. Claire Boyer, Antonin Chambolle, Yohann De Castro, Vincent Duval, Frédéric De Gournay, and Pierre Weiss. On Representer Theorems and Convex Regularization. *SIAM Journal of Optimization*, 29(2):1260–1281, 2019.
4. Kristian Bredies and Hanna Katriina Pikkarainen. Inverse problems in spaces of measures. *ESAIM: Control, Optimisation and Calculus of Variations*, 19(1):190–218, 2013.
5. Emmanuel J Candès and Carlos Fernandez-Granda. Towards a Mathematical Theory of Super-resolution. *Communications on Pure and Applied Mathematics*, 67(6):906–956, 2014.
6. Scott Shaobing Chen, David L Donoho, and Michael A Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.
7. Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. In *Advances in neural information processing systems*, pages 3036–3046, 2018.
8. Yohann De Castro and Fabrice Gamboa. Exact reconstruction using Beurling minimal extrapolation. *Journal of Mathematical Analysis and applications*, 395(1):336–354, 2012.
9. Yohann De Castro, Fabrice Gamboa, Didier Henrion, and J-B Lasserre. Exact solutions to Super Resolution on semi-algebraic domains in higher dimensions. *IEEE Transactions on Information Theory*, 63(1):621–630, 2017.
10. Quentin Denoyelle, Vincent Duval, Gabriel Peyré, and Emmanuel Soubies. The Sliding Frank-Wolfe Algorithm and its Application to Super-Resolution Microscopy. *Inverse Problems*, 36, 2019.
11. Charles Dossal, Vincent Duval, and Clarice Poon. Sampling the Fourier transform along radial lines. *SIAM Journal on Numerical Analysis*, 55(6):2540–2564, 2017.
12. Vincent Duval and Gabriel Peyré. Exact support recovery for sparse spikes deconvolution. *Foundations of Computational Mathematics*, 15(5):1315–1355, 2015.
13. Armin Eftekhari and Andrew Thompson. Sparse inverse problems over measures: Equivalence of the conditional gradient and exchange methods. *SIAM Journal on Optimization*, 29:1329–1349, 2019.
14. S. D. Fisher and J. W. Jerome. Spline solutions to L1 extremal problems in one and several variables. *Journal of Approximation Theory*, 13(1):73–83, 1975.
15. Axel Flinth and Pierre Weiss. Exact solutions of infinite dimensional total-variation regularized problems. *Information and Inference*, 8:407–443, 2017.
16. Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 3(1-2):95–110, 1956.
17. Rainer Hettich and Kenneth O Kortanek. Semi-infinite programming: theory, methods, and applications. *SIAM review*, 35(3):380–429, 1993.
18. Rainer Hettich and Peter Zencke. *Numerische Methoden der Approximation und semi-infiniten Optimierung*. Vieweg+Teubner, 1982.
19. Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal. *Convex analysis and minimization algorithms I: Fundamentals*, volume 305. Springer science & business media, 2013.
20. Evgenii Solomonovich Levitin and Boris Teodorovich Polyak. Constrained minimization methods. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 6(5):787–823, 1966.
21. Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.
22. Konstantin Pieper and Daniel Walter. Linear convergence of accelerated conditional gradient algorithms in spaces of measures. *arXiv preprint arXiv:1904.09218*, 2019.
23. Clarice Poon, Nicolas Keriven, and Gabriel Peyré. Support Localization and the Fisher Metric for off-the-grid Sparse Regularization. *Proceedings of Machine Learning Research*, 89:1341–1350, 2019.
24. Rembert Reemtsen. Modifications of the first Remez algorithm. *SIAM journal on numerical analysis*, 27(2):507–518, 1990.
25. Rembert Reemtsen and Stephan Görner. Numerical methods for semi-infinite programming: A survey. pages 195–262, 1998.
26. Eugène Remes. Sur un procédé convergent d'approximations successives pour déterminer les polynômes d'approximation. *CR Acad. Sci. Paris*, 198:2063–2065, 1934.
27. Gongguo Tang, Badri Narayan Bhaskar, and Benjamin Recht. Sparse recovery over continuous dictionaries-just discretize. In *Signals, Systems and Computers, 2013 Asilomar Conference on*, pages 1043–1047. IEEE, 2013.
28. Gongguo Tang, Badri Narayan Bhaskar, Parikshit Shah, and Benjamin Recht. Compressed sensing off the grid. *IEEE transactions on information theory*, 59(11):7465–7490, 2013.
29. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

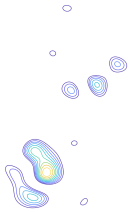
30. Yann Traonmilin and Jean-François Aujol. The basins of attraction of the global minimizers of the non-convex sparse spikes estimation problem. *Inverse Problems*, 36, 2020.
31. Michael Unser, Julien Fageot, and John Paul Ward. Splines are universal solutions of linear inverse problems with generalized tv regularization. *SIAM Review*, 59(4):769–793, 2017.
32. Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.
33. S.I. Zuhovickiĭ. Remarks on problems in approximation theory. *Mat. Zbirnik KDU*, pages 169–183, 1948. (Ukrainian).



(a) $|A^*q_1|$

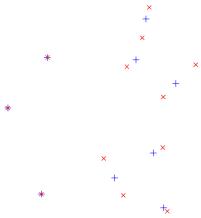


(b) $|A^*q_2|$

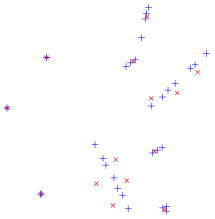


(c) $|A^*q_4|$

(d) Grid 1



(e) Grid 2

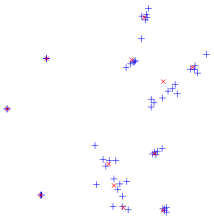


(f) Grid 4

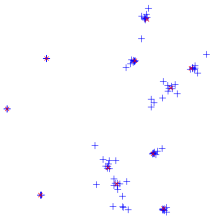
(g) $|A^*q_8|$

(h) $|A^*q_{10}|$

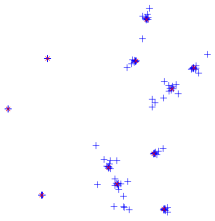
(i) $|A^*q_{12}|$



(j) Grid 8

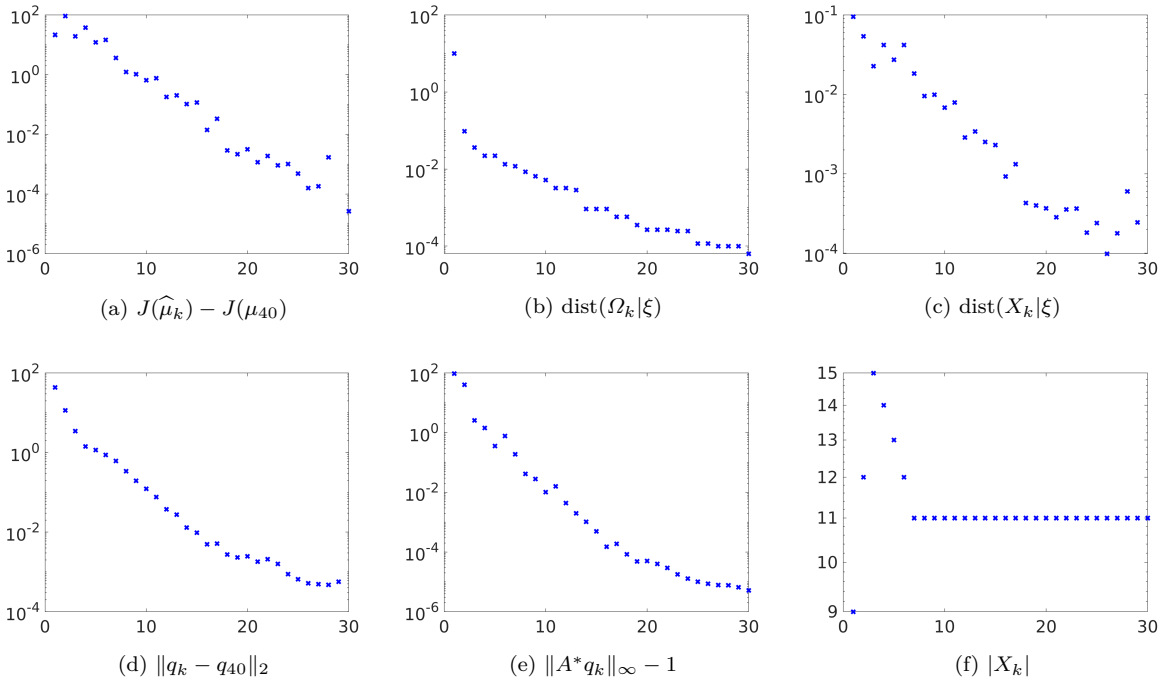


(k) Grid 10



(l) Grid 12

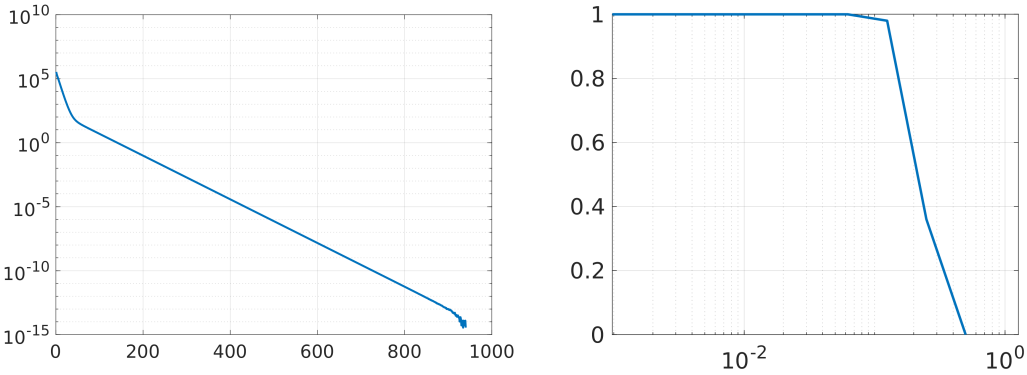
Fig. 6: Evolution of the dual certificate and of the grid through the 12 first iterations. This is a contour plot with the levels from 1 to the maximum of $|A^*q_i|$ indicated.



(a) $J(\widehat{\mu}_k) - J(\mu_{40})$ (b) $\mathrm{dist}(\Omega_k|\xi)$ (c) $\mathrm{dist}(X_k|\xi)$

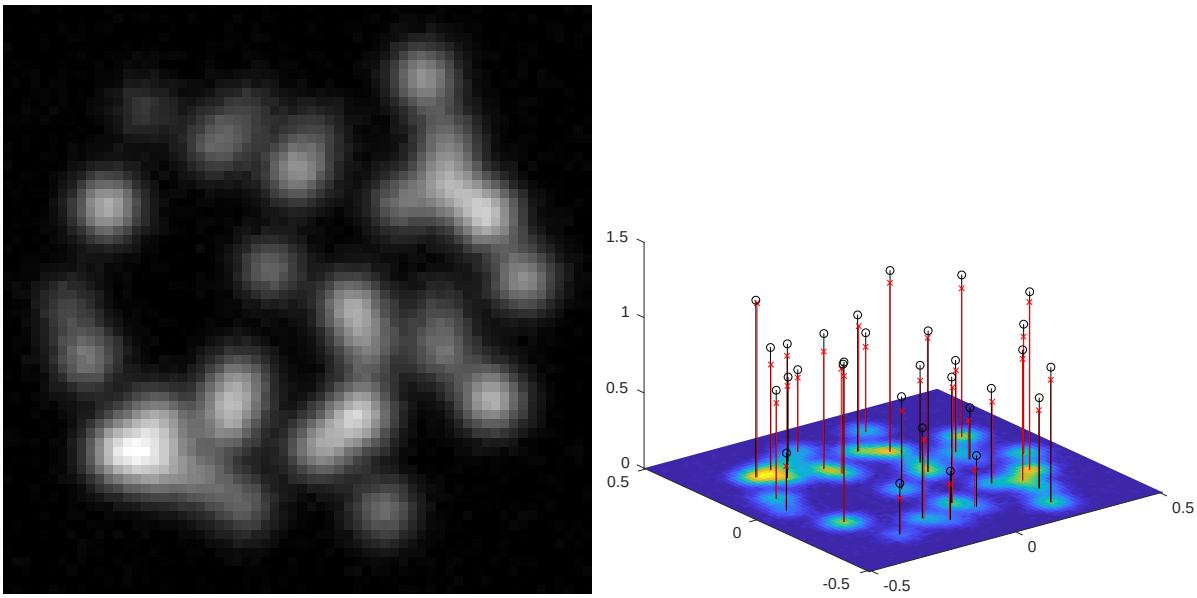
(d) $\|q_k - q_{40}\|_2$ (e) $\|A^* q_k\|_\infty - 1$ (f) $|X_k|$

Fig. 7: Plot of several quantities of interest along the exchange algorithm's iterates.



(a) $G(\alpha^{(t)}, X^{(t)}) - G(\alpha^\star, \xi)$ (b) Success rate VS starting point

Fig. 8: Left: Typical convergence curve in logarithmic scale when the initial guess $(\alpha^{(0)}, X^{(0)})$ is good enough. Right: Success rate of the continuous descent method over 50 runs of the algorithm, depending on the relative amplitude of the perturbation.



(a) Measurements y (dense)

(b) Ground truth and recovered solution

Fig. 9: Left: measurements associated to a denser measure with more noise. Right: 3D illustration of the recovery results. The blue vertical bars with circles indicate the locations and amplitude of the ground truth. The red bars with crosses indicated the recovered measures. Apart from a slight bias in amplitude due to the ℓ^1-norm, the ground truth is near perfectly recovered.