"FISTA" in Banach spaces with adaptive discretisations

Antonin Chambolle, Robert Tovey

Computational Optimization and Applications manuscript No.
(will be inserted by the editor)

"FISTA" in Banach spaces with adaptive discretisations

Antonin Chambolle · **Robert Tovey**

the date of receipt and acceptance should be inserted later

Abstract FISTA is a popular convex optimisation algorithm which is known to converge at an optimal rate whenever a minimiser is contained in a suitable Hilbert space. We propose a modified algorithm where each iteration is performed in a subset which is allowed to change at every iteration. Sufficient conditions are provided for guaranteed convergence, although at a reduced rate depending on the conditioning of the specific problem. These conditions have a natural interpretation when a minimiser exists in an underlying Banach space. Typical examples are L1-penalised reconstructions where we provide detailed theoretical and numerical analysis.

Keywords Convex optimization · Multiscale · Multigrid · Sparsity · Lasso

1 Introduction

The Fast Iterative Shrinkage-Thresholding Algorithm (FISTA) was proposed by Beck and Teboulle [3] as an extension of Nesterov's fast gradient method [26] and is now a very popular algorithm for minimising the sum of two convex functions. We write this as the problem of computing

$$\inf_{u\in\mathbb{H}} \mathrm{E}(u) \qquad \text{such that} \qquad \mathrm{E}(u) \coloneqq \mathrm{f}(u) + \mathrm{g}(u), \tag{1}$$

for a Hilbert space $\mathbb{H}$ where $\mathrm{f}\colon \mathbb{H} \to \mathbb{R}$ is a convex differentiable function with L-Lipschitz gradient and $\mathrm{g}\colon \mathbb{H} \to \overline{\mathbb{R}}$ is a "simple" convex function, whose "proximity operator" is easy to compute. Throughout this work we assume that E is bounded below so that the infimum is finite. The iterates of the FISTA algorithm will be denoted $u_n \in \mathbb{H}$. If, moreover the infimum is achieved, it has been shown that $\mathrm{E}(u_n) - \inf_{u\in\mathbb{H}} \mathrm{E}(u)$ converges at the optimum rate of n^{-2} [3], and later (after a small modification) the convergence of the iterates was also shown in a general Hilbert space setting [10]. Many further works have gone on to demonstrate faster practical convergence rates for slightly modified variants of FISTA [37, 23, 1].

In this work we address the case where the minimiser possibly fails to exist or lies in a larger space where $\mathbb{H}$ is dense. There is much overlap between the techniques used in this work and those used in the literature of inexact optimisation, however, our interpretation is relatively novel. In particular, we emphasise the infinite-dimensional setting where errors come from "discretisation", rather than random or decaying errors in $\mathbb{H}$, which enables two new perspectives:

– Analytically, we prove new rates of convergence for FISTA when the minimum energy is not achieved (at least not in $\mathbb{H}$). The exact rate can be computed by quantifying coercivity and regularity properties of E. If there isn't a minimiser in $\mathbb{H}$, then this rate is strictly slower than n^{-2}.

A. Chambolle
CEREMADE, CNRS & Université Paris Dauphine, PSL Research University, Paris
E-mail: chambolle@ceremade.dauphine.fr, ORCID: 0000-0002-9465-4659

R. Tovey
MOKAPLAN, INRIA Paris, Paris
E-mail: robert.tovey@inria.fr, ORCID: 0000-0001-5411-2268

– Numerically, we allow the optimisation domain to change on every iteration. This enables us to understand how FISTA behaves with adaptive discretisations. Adaptive finite-element methods are known to improve the efficiency of, for example, approximating the solutions of PDEs. Our analytical results show how to combine such tools with FISTA without reducing the guaranteed rate of convergence, and our numerical results confirm much improved time and computer memory efficiency in the Lasso example (Section 6).

All the examples in this work, discussed from Section 5 onward, consider $\{u \in \mathbb{H} \text{ s.t. } \mathrm{E}(u) < \infty\}$ to be contained in some ambient Banach space $\mathbb{U}$. The idea is that FISTA provides a minimising sequence in $\mathbb{H} \cap \mathbb{U}$, but further properties like rate of convergence (of E or the iterates) must come from the topology of $\mathbb{U}$. It will not be necessary for $\mathbb{H} \hookrightarrow \mathbb{U}$ to be a continuous embedding, nor in fact the full inclusion $\mathbb{H} \subset \mathbb{U}$.

Some other works for FISTA-like algorithms include [21,39]. Of particular note, our stability estimate for FISTA in Theorem 1 is very similar to [35, Prop 2] and [2, Prop 3.3]. This is then used to analyse the convergence properties in our more general Banach space setting, but where all sources of inexactness come from subspace approximations. The ideas in [28] are similar although in application to the proximal gradient method with an additional smoothing on the functional g. The permitted refinement steps are also more broad in our work. Very recent work in [40] proposes a "Multilevel FISTA" algorithm which allows similar coarse-to-fine refinement strategies, although only a finite number. We also allow for non-uniform refinement with a posteriori strategies.

1.1 Outline

This work is organised as follows. Section 2 defines notation and the generic form of our proposed refining FISTA algorithm, Algorithm 1. The main theoretical contribution of this work is the convergence analysis of Algorithm 1 which is split into two parts: first we outline the proof structure in Section 3, then we state the specific results in the case of FISTA in Section 4. The main results are Theorems 2/3 which extend the convergence of FISTA to cases with un-attained minima with uniform/adaptively chosen subspaces $\mathbb{U}^n$ respectively.

Section 5 presents some general results for the application of Algorithm 1 in Banach spaces and Section 6 gives a much more detailed discussion of adaptive refinement for Lasso minimisation. In particular, we describe how to choose efficient refining discretisations to approximate $\inf_{u\in\mathbb{H}} \mathrm{E}(u)$, estimate the convergence of E, and identify the support of the minimiser. The numerical results in Section 7 demonstrate these techniques in four different models demonstrating the apparent sharpness of our convergence rates and the computational efficiency of adaptive discretisations.

2 Definitions and notation

We consider optimisation of (1) over a Hilbert space $(\mathbb{H}, \langle\cdot,\cdot\rangle, \|\cdot\|)$. In the more analytical section (Sections 3 and 4) it will be more convenient to use the translated energy

$$\mathrm{E}_0 \colon \mathbb{H} \to \mathbb{R}, \qquad \mathrm{E}_0(u) \coloneqq \mathrm{E}(u) - \inf_{\widetilde{u}\in\mathbb{H}} \mathrm{E}(\widetilde{u}) \tag{2}$$

so that $\inf_{u\in\mathbb{H}} \mathrm{E}_0(u) = 0$, although access to this function is not assumed for numerical examples.

The proposed generalised FISTA algorithm is stated in Algorithm 1 for an arbitrary choice of closed convex subsets $\mathbb{U}^n \subset \mathbb{H}$ for $n \in \mathbb{N}$. The only difference from standard FISTA is that on iteration n, all computations are performed in the subset $\mathbb{U}^n$. If $\mathbb{U}^n = \mathbb{H}$, then we recover the original algorithm. More generally, the idea is that $\mathbb{U}^n$ are "growing", for example $\mathbb{U}^n \subset \mathbb{U}^{n+1}$, but this assumption is not necessary in most of the results.

Without loss of generality we will assume $L = 1$, i.e. $\nabla \mathrm{f}$ is 1-Lipschitz. To get the general statement of any of the results which follow, replace E with $\frac{\mathrm{E}}{L}$. In particular,

$$\|\nabla \mathrm{f}(u) - \nabla \mathrm{f}(v)\| \leq \|u - v\| \tag{3}$$

for all $u, v \in \mathbb{H}$ and g is called "simple" if it is proper, convex, weakly lower-semicontinuous, and

$$\operatorname*{argmin}_{u\in\widetilde{\mathbb{U}}} \tfrac{1}{2}\|u - v\|^2 + \mathrm{g}(u) \tag{4}$$

is exactly computable for all $v \in \mathbb{H}$ and all $\widetilde{\mathbb{U}} \in \{\mathbb{U}^n\}_{n=0}^{\infty}$. Closed subsets of $\mathbb{H}$ are locally weakly compact, therefore this argmin is always non-empty.

One defining property of the FISTA algorithm is an appropriate choice of inertia, dictated by t_n. In particular, we will say that $(t_n)_{n=0}^{\infty}$ is a *FISTA stepsize* if

$$t_0 = 1, \qquad t_n \geq 1, \qquad \text{and} \qquad \rho_n := t_n^2 - t_{n+1}^2 + t_{n+1} \geq 0 \qquad \text{for all } n = 0, 1, \ldots. \tag{5}$$

The precise constants associated to a given rate are given in the statements of the theorems but, for convenience, are otherwise omitted from the text. For sequences $(a_n)_{n=0}^{\infty}$,$(b_n)_{n=0}^{\infty}$ we will use the notation:

$$\begin{aligned} a_n \lesssim b_n \qquad &\iff \qquad \exists C, N > 0 \text{ s.t. } a_n \leq C b_n \text{ for all } n > N, \\ a_n \simeq b_n \qquad &\iff \qquad a_n \lesssim b_n \lesssim a_n. \end{aligned}$$

For $n \in \mathbb{N}$ we use the abbreviation $[n] = \{1, 2, \ldots, n\}$. When the subdifferential of E is set-valued, we will use the short-hand

$$|||\partial \mathrm{E}(u)||| := \inf_{v \in \partial \mathrm{E}(u)} |||v||| \tag{6}$$

for any specified norm $|||\cdot|||$.

Algorithm 1 Refining subset FISTA

1: Choose $(\mathbb{U}^n)_{n\in\mathbb{N}}$, $u_0 \in \mathbb{U}^0$ and some FISTA stepsize choice $(t_n)_{n\in\mathbb{N}}$
2: $v_0 \leftarrow u_0, n \leftarrow 0$
3: **repeat**
4: $\quad \overline{u}_n \leftarrow (1 - \frac{1}{t_n})u_n + \frac{1}{t_n} v_n$
5: $\quad u_{n+1} \leftarrow \operatorname{argmin}_{u \in \mathbb{U}^{n+1}} \frac{1}{2}\|u - \overline{u}_n + \nabla \mathrm{f}(\overline{u}_n)\|^2 + \mathrm{g}(u)$ ▷ Only modification, $\mathbb{U}^{n+1} \subset \mathbb{U}$
6: $\quad v_{n+1} \leftarrow (1 - t_n)u_n + t_n u_{n+1}$
7: $\quad n \leftarrow n + 1$
8: **until** some stopping criterion is met

3 General proof recipe

In this section we give an intuitive outline of the full proof for convergence of Algorithm 1 before giving formal theorems and proofs in the next section. First we recall the classical FISTA convergence guarantee given by [10, Thm 3.1]; if there exists $u^* \in \operatorname{argmin}_{u\in\mathbb{H}} \mathrm{E}_0(u)$, then

$$t_N^2 \mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n \mathrm{E}_0(u_n) + \tfrac{1}{2}\|v_N - u^*\|^2 \leq \tfrac{1}{2}\|u_0 - u^*\|^2 \tag{7}$$

for any FISTA stepsize choice $t_N \simeq N$ such that $\rho_n \geq 0$.

Step 1: Quantifying the stability The first step is to generalise (7) to account for the adapting subsets $\mathbb{U}^n$. In the notation of Algorithm 1, Theorem 1 shows that

$$t_N^2 \mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n \mathrm{E}_0(u_n) + \tfrac{1}{2}\|v_N - w_N\|^2 \leq \tfrac{1}{2}\|u_0 - w_0\|^2 + \tfrac{\|w_N\|^2 - \|w_0\|^2}{2} + \sum_{n=1}^{N} t_n \mathrm{E}_0(w_n) + \langle v_{n-1}, w_{n-1} - w_n \rangle \tag{8}$$

for any $w_n \in \mathbb{U}^n$. The similarities to (7) are clear. If $\mathbb{U}^n = \mathbb{H}$, then we can choose $w_n = u^*$ and the two estimates agree. These extra terms in (8) quantify the robustness to changing of discretisation.

Step 2: Quantifying the scaling properties To show that the extra terms in (8) are small, we need to quantify the approximation properties of $\mathbb{U}^n$. The idea is that there is a sequence $w_n \in \mathbb{U}^n$, $n \in \mathbb{N}$ such that $\|w_n\|$ grows slowly and $\mathrm{E}_0(w_n)$ decreases quickly. To quantify this balance, we introduce a secondary sequence $n_0 < n_1 < \dots$ and constants $a_\mathrm{U}, a_\mathrm{E} \geq 1$ such that for each $k \in \mathbb{N}$

$$n \leq n_k \implies \|w_n\| \lesssim a_\mathrm{U}^k, \qquad n \geq n_k \implies \mathrm{E}_0(w_n) \lesssim a_\mathrm{E}^{-k}. \tag{9}$$

A canonical example would be $\mathbb{U}^n = \{u \in \mathbb{H} \text{ s.t. } \|u\| \leq a_\mathrm{U}^k\}$ for $n \in [n_k, n_{k+1})$, then a_E reflects the smoothness of E_0. The choice of exponential scaling is introduced to improve stability of Algorithm 1. It is natural if we consider the $\mathbb{U}^n$ to be the subspace of functions discretised on a uniform mesh. If that mesh is sequentially refined, then the resolution of the mesh will be of order h^k after k refinements and for some $h < 1$. The integer n_k is then the time at which the mesh has refined k times. The trade-off between a_E and a_U dictates the final convergence rate of the algorithm. If $a_\mathrm{U} > 1$, then we cannot guarantee the original n^{-2} rate of convergence.

Step 3: Generalising the convergence bound In this step we combine the FISTA stability estimate with the subset approximation guarantees to provide a sharper estimate of stability with respect to the parameters a_E and a_U. For example, if for each $k \in \mathbb{N}$

$$w_n = w_{n_k} \text{ for each } n = n_k, n_k + 1, \dots, n_{k+1} - 1,$$

then many terms on the right-hand side of (8) telescope to 0. The result of this is presented in Lemma 3. The key idea is that the stability error in (8) has $K \ll N$ terms, rather than N.

Step 4: Sufficiently fast growth In Step 3 we develop a convergence bound, now we wish to show that it is only worse than the classical (7) by a constant factor. In particular, it is equivalent to either run Algorithm 1 for N iterations, or the classical FISTA algorithm for N iterations on the fixed subset $\mathbb{U}^N$. The estimate from (7) provides the estimate $N^2 \mathrm{E}_0(u_N) \lesssim \|u_0 - w_N\|^2 = O(a_\mathrm{U}^{2K})$ for $N \leq n_K$. Lemma 4 shows that Algorithm 1 can achieve the same order of approximation, so long as $\mathbb{U}^n$ grow sufficiently quickly (in particular $n_k^2 \lesssim a_\mathrm{E}^k a_\mathrm{U}^{2k}$).

Step 5: Sufficiently slow growth The result of Step 4 is sufficient to prove convergence, but not yet a rate. If the subsets grow too quickly, then the influence of $\|u_n\| \to \infty$ will slow the rate of convergence. If n_k is too large, then we overfit to the discrete problem, but if n_k is too small, then FISTA converges slowly. Lemma 5 balances these two factors in an optimal way ($n_k^2 \simeq a_\mathrm{E}^k a_\mathrm{U}^{2k}$) for Algorithm 1 resulting in a convergence rate of

$$\mathrm{E}_0(u_N) \lesssim \frac{a_\mathrm{U}^{2K}}{N^2} \lesssim \frac{N^{2\kappa}}{N^2}$$

for all $N \in \mathbb{N}$ and $\kappa = \frac{2 \log a_\mathrm{U}}{\log a_\mathrm{E} + 2 \log a_\mathrm{U}} \in [0, 1)$. In particular, if the minimum is attained in $\mathbb{H}$, then we recover the classical rate with $\kappa = 0$.

Step 6: Adaptivity Up to this point we have implicitly focused on the case where $\mathbb{U}^n$ (and n_k) are chosen a priori. The main challenge for adaptive choice of $\mathbb{U}^n$ is to guarantee (9) from Step 2 using a posteriori estimates. Combined with the partial telescoping requirement in Step 3, a natural choice is $w_n = u_{n_k - 1}$ for $n \in [n_k, n_{k+1})$, i.e. the value of n_k is chosen to be $n+1$ once the iterate u_n is observed. Theorem 3 shows that a sufficient condition is

$$u_{n_k - 1} \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k + 1} \cap \dots \cap \mathbb{U}^{n_{k+1} - 1}, \qquad \left\|u_{n_k - 1}\right\| \lesssim a_\mathrm{U}^k, \quad \text{and} \quad \mathrm{E}_0(u_{n_k - 1}) \lesssim a_\mathrm{E}^{-k}.$$

Convergence is most stable if the approximation spaces $\mathbb{U}^n$ satisfy a monotone inclusion, breaking the monotonicity requires more care. The only non-trivial property to verify is the energy gap $\mathrm{E}_0(u_n) = \mathrm{E}(u_n) - \inf_{u \in \mathbb{H}} \mathrm{E}(u)$. Lemma 6 proposes some sufficient conditions to guarantee the same overall rate of convergence as in Step 5,

$$\min_{n \leq N} \mathrm{E}_0(u_n) \lesssim \frac{N^{2\kappa}}{N^2}$$

for all $N \in \mathbb{N}$, with the same $\kappa \in [0, 1)$ from Step 5. The penalty for accelerating the change of discretisation is a potential loss of stability or monotonicity in $\mathrm{E}_0(u_n)$, although this behaviour has not been seen in numerical experiments.

4 Proof of convergence

In this section we follow the recipe motivated in Section 3 to prove convergence of two variants of Algorithm 1. Each of the main theorems and lemmas will be stated with a sketch proof in this section. The details of the proofs are either trivial or very technical and are therefore placed in Section A to preserve the flow of the argument.

4.1 Computing the convergence bound

For Step 1 of Section 3 we look to replicate the classical bound of the form in (7) for Algorithm 1. The proofs in this step follow the classical arguments [3,10] very closely. Throughout this section we consider a sequence $(\mathbb{U}^n)_{n\in\mathbb{N}}$ which generate the iterates $(u_n)_{n\in\mathbb{N}}$ in Algorithm 1 such that

$$u^n \in \mathbb{U}^{n+1} \subset \mathbb{H} \quad \text{where } \mathbb{U}^n \text{ is a closed, convex subset for all } n \in \mathbb{N}. \tag{10}$$

4.1.1 Single iterations

We first wish to understand a single iteration of Algorithm 1. This is done through the following two lemmas.

Lemma 1 (equivalent to [10, Lemma 3.1]) *Suppose $\nabla \mathrm{f}$ is 1-Lipschitz, for any $\overline{u} \in \mathbb{H}$ define*

$$u := \operatorname*{argmin}_{u\in\mathbb{U}^n} \tfrac{1}{2}\|u - \overline{u} + \nabla \mathrm{f}(\overline{u})\|^2 + \mathrm{g}(u).$$

Then, for all $w \in \mathbb{U}^n$, we have

$$\mathrm{E}_0(u) + \tfrac{1}{2}\|u - w\|^2 \leq \mathrm{E}_0(w) + \tfrac{1}{2}\|\overline{u} - w\|^2.$$

The proof is exactly the same as in [10] on the subset $\mathbb{U}^n$. Applying Lemma 1 to the iterates from Algorithm 1 gives a more explicit inequality.

Lemma 2 ([10, (17)], [3, Lemma 4.1]) *Let $w_n \in \mathbb{U}^n$ be chosen arbitrarily and u_n/v_n be generated by Algorithm 1 for all $n \in \mathbb{N}$. For all $n > 0$, it holds that*

$$t_n^2(\mathrm{E}_0(u_n) - \mathrm{E}_0(w_n)) - (t_n^2 - t_n)(\mathrm{E}_0(u_{n-1}) - \mathrm{E}_0(w_n)) \leq \tfrac{1}{2}\left[\|v_{n-1}\|^2 - \|v_n\|^2\right] + \langle v_n - v_{n-1}, w_n\rangle. \tag{11}$$

The proof is given in Theorem 5 and is a result of the convexity of E_0 and $\mathbb{U}_n$ for a well chosen w in Lemma 1.

4.1.2 Generic convergence bound

Lemma 2 gives us an understanding of a single iteration of Algorithm 1, summing over n then gives our generic convergence bound for any variant of Algorithm 1.

Theorem 1 (analogous to [10, Thm 3.2], [3, Thm 4.1]) *Fix a sequence of subsets $(\mathbb{U}^n)_{n\in\mathbb{N}}$ satisfying (10), arbitrary $u_0 \in \mathbb{U}^0$, and FISTA stepsize choice $(t_n)_{n\in\mathbb{N}}$. Let u_n and v_n be generated by Algorithm 1, then, for any choice of $w_n \in \mathbb{U}^n$ and $N \in \mathbb{N}$ we have*

$$t_N^2\,\mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n\,\mathrm{E}_0(u_n) + \frac{\|v_N - w_N\|^2}{2} \leq \frac{\|u_0 - w_0\|^2 - \|w_0\|^2 + \|w_N\|^2}{2} + \sum_{n=1}^{N} t_n\,\mathrm{E}_0(w_n) + \langle v_{n-1}, w_{n-1} - w_n\rangle. \tag{12}$$

The proof is given in Theorem 6. This result is the key approximation for showing convergence of FISTA with changing subsets. In the classical setting, we have $\mathbb{U}^n = \mathbb{H}$, $w_n = w_0 \in \operatorname{argmin}_{u\in\mathbb{H}} \mathrm{E}_0(u)$ and the extra terms on the right-hand side collapse to 0.

If there exists a minimiser $u^* \in \operatorname{argmin}_{u\in\mathbb{H}} \mathrm{E}_0(u)$, then the natural choice in (12) is $w_n = \Pi_n u^*$ for some projection $\Pi_n \colon \mathbb{H} \to \mathbb{U}^n$, however, there are simple counter-examples which give $\mathrm{E}_0(\Pi_n u^*) = \infty$ and so this inequality becomes useless. For example, if $\mathrm{f}(u) = \|u\|^2_{L^2([0,1])}$, g is the indicator on the set $\mathbb{D} = \{u \in L^1([0,1]) \text{ s.t. } u(x) \geq x\}$, and Π_n is the L^2 projection onto a set of piecewise constant functions, then $u^* = x \mapsto x$. On the other hand, suppose one of the pixels of the discretisation is $[x_0 - h, x_0 + h]$, then

$$\Pi_n u^*\left(x_0 + \tfrac{h}{2}\right) = \operatorname*{argmin}_{c\in\mathbb{R}} \int_{x_0-h}^{x_0+h} (u^*(x) - c)^2\,\mathrm{d}x = \operatorname*{argmin}_{c\in\mathbb{R}} \int_{x_0-h}^{x_0+h} (x - c)^2\,\mathrm{d}x = x_0 < x_0 + \tfrac{h}{2}.$$

In particular $\Pi_n u^* \notin \mathbb{D}$ therefore $\mathrm{E}_0(\Pi_n u^*) = \infty$. The choice $w_n = \operatorname{argmin}_{u\in\mathbb{U}^n} \mathrm{E}_0(u)$ is much more robust and allows us to apply Algorithm 1 more broadly. The penalty for this flexibility is a more complicated analysis; each time the subset changes, because $v_n \in \mathbb{U}_n$, the system receives a "shock" proportional to $\|v_n\|\|w_n - \Pi_n w_{n+1}\|$.

4.2 Convergence bound with milestones

In standard FISTA, the right-hand side of (12) is a constant. The following lemma minimises the growth of the "constant" as a function of N by partially telescoping the sum on the right-hand side. Before progressing to the content of Step 3, we will first formalise the definition of the constants a_{U} and a_{E} introduced in Step 2.

Definition 1 Fix $a_{\mathrm{U}}, a_{\mathrm{E}} \geq 1$ and a sequence $\widetilde{w}_k \in \mathbb{H}$. We say that $(\widetilde{w}_k)_{k\in\mathbb{N}}$ is an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-*minimising sequence of* E if

$$\|\widetilde{w}_k\| \lesssim a_{\mathrm{U}}^k \qquad \text{and} \qquad \mathrm{E}_0(\widetilde{w}_k) \lesssim a_{\mathrm{E}}^{-k}$$

for all $k \in \mathbb{N}$.

In this section we will simply assume that such sequences exist and in Section 5 we will give some more general examples.

Lemma 3 *Let u_n, v_n be generated by Algorithm 1 with $(\mathbb{U}^n)_{n\in\mathbb{N}}$ satisfying (10), $(n_k \in \mathbb{N})_{k\in\mathbb{N}}$ be a monotone increasing sequence, and choose*

$$\widetilde{w}_k \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$

for each $k \in \mathbb{N}$. If such a sequence exists, then for all $K \in \mathbb{N}$, $n_K \leq N < n_{K+1}$ we have

$$t_N^2 \mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n \mathrm{E}_0(u_n) + \frac{\|v_N - \widetilde{w}_K\|^2}{2} \leq C + \frac{\|\widetilde{w}_K\|^2}{2} + \frac{(N+1)^2 - n_K^2}{2} \mathrm{E}_0(\widetilde{w}_K) + \sum_{k=1}^{K} \frac{n_k^2 - n_{k-1}^2}{2} \mathrm{E}_0(\widetilde{w}_{k-1}) + \left\langle v_{n_k-1}, \widetilde{w}_{k-1} - \widetilde{w}_k \right\rangle$$

where $C = \frac{\|u_0 - \widetilde{w}_0\|^2 - \|\widetilde{w}_0\|^2}{2}$.

The proof is given in Lemma 11. The introduction of n_k has greatly compressed the expression of Theorem 1. On the right-hand side, we now only consider E_0 evaluated on the sequence $\widetilde{w}_k$ and there are K elements to the sum rather than N.

4.3 Refinement without overfitting

The aim of Step 4 is to show that n iterations of Algorithm 1 is no slower (up to a constant factor) than n iterations of classical FISTA on the space $\mathbb{U}^n$. In other words, we would like to ensure that

$$\mathrm{E}_0(u_n) = \mathrm{E}(u_n) - \min_{u\in\mathbb{U}^n} \mathrm{E}(u) \lesssim \frac{\|u_0 - \widetilde{w}_k\|^2}{n^2} \tag{13}$$

uniformly for $n \in [n_k, n_{k+1})$. If this condition is not satisfied, then it indicates that computational effort has been wasted by a poor choice of subsets. This can be interpreted as an overfitting to the discretisation of $\mathrm{E}_0|_{\mathbb{U}^n}$ rather than the desired function $\mathrm{E}_0|_{\mathbb{H}}$. Combining the assumptions given by Definition 1 and the result of Lemma 3, the following lemma proves the convergence of Algorithm 1 provided that the refinement times n_k are sufficiently small (i.e. $\mathbb{U}^n$ grows sufficiently quickly).

Lemma 4 *Suppose $\mathbb{U}^n$, u_n, v_n and n_k satisfy the conditions of Lemma 3 and $(\widetilde{w}_k)_{k\in\mathbb{N}}$ forms an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of* E *with*

$$\widetilde{w}_k \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}.$$

If either:

- $a_{\mathrm{U}} > 1$ *and* $n_k^2 \lesssim a_{\mathrm{E}}^k a_{\mathrm{U}}^{2k}$,
- *or* $a_{\mathrm{U}} = 1$, $\sum_{k=1}^{\infty} n_k^2 a_{\mathrm{E}}^{-k} < \infty$, *and* $\sum_{k=1}^{\infty} \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$,

then

$$\mathrm{E}_0(u_N) \lesssim \frac{a_{\mathrm{U}}^{2K}}{N^2} \qquad \text{for all} \qquad n_K \leq N < n_{K+1}.$$

The proof is given in Lemma 12. We make two observations of the optimality of Lemma 4:

- The convergence guarantee for $N \in [n_K, n_{K+1})$ iterations of classical FISTA in the space $\mathbb{U}^N$ is
$$\mathrm{E}_0(u_N) \lesssim \frac{\|u_0 - \widetilde{w}_K\|^2}{N^2} + \min_{u \in \mathbb{U}^N} \mathrm{E}_0(u) \lesssim \frac{a_{\mathrm{U}}^{2K}}{N^2} + a_{\mathrm{E}}^{-K}.$$
This is equivalent to Lemma 4 after the assumptions on n_k.
- If $\mathbb{H}$ is finite dimensional, then the condition $a_{\mathrm{U}} = 1$ is almost trivially satisfied. Norms in finite dimensions are equivalent and any discretisation can be achieved with a finite number of refinements (i.e. the sums over k are finite).

4.4 Convergence rate

In Lemma 4 we show that $\mathrm{E}_0(u_n)$ converges at a rate depending on k and n, so long as k grows sufficiently quickly. On the other hand, as k grows, the rate becomes worse and so we need to also put a lower limit on the growth of n_k. The following lemma completes Step 5 by computing the global convergence rate of $\mathrm{E}_0(u_n)$ when k grows at the minimum rate which is consistent with Lemma 4.

As a special case, note that if $a_{\mathrm{U}} = 1$ then Lemma 4 already gives the optimal $O(N^{-2})$ convergence rate. This is in fact a special case of that shown in [2, Prop 3.3]. If the minimum is achieved in $\mathbb{H}$, then it is not possible to refine "too quickly" and the following lemma is not needed.

Lemma 5 *Suppose u_n and n_k are sequences satisfying*
$$\forall N \in [n_K, n_{K+1}),\ \mathrm{E}_0(u_N) \lesssim \frac{a_{\mathrm{U}}^{2K}}{N^2} \qquad \text{where} \qquad n_K^2 \gtrsim a_{\mathrm{E}}^K a_{\mathrm{U}}^{2K},$$
then
$$\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\kappa)}} \qquad \text{where} \qquad \kappa = \frac{\log a_{\mathrm{U}}^2}{\log a_{\mathrm{E}} + \log a_{\mathrm{U}}^2}.$$

The proof is given in Lemma 13.

4.4.1 FISTA convergence with a priori discretisation

We can summarise Lemmas 3 to 5 into a single theorem stating the convergence guarantees when $\mathbb{U}^n$ and n_k are chosen a priori.

Theorem 2 *Let $(\widetilde{w}_k)_{k \in \mathbb{N}}$ be an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of* E *and choose any $\mathbb{U}^n$ satisfying* (10) *such that*
$$\widetilde{w}_k \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$
for all $k \in \mathbb{N}$. Compute u_n and v_n by Algorithm 1.

Suppose that either:

- *$a_{\mathrm{U}} > 1$ and $n_k^2 \simeq a_{\mathrm{E}}^k a_{\mathrm{U}}^{2k}$, or*
- *$a_{\mathrm{U}} = 1$, $\sum_{k=1}^{\infty} n_k^2 a_{\mathrm{E}}^{-k} < \infty$ and $\sum_{k=1}^{\infty} \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$,*

then
$$\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\kappa)}} \qquad \text{where} \qquad \kappa = \frac{\log a_{\mathrm{U}}^2}{\log a_{\mathrm{E}} + \log a_{\mathrm{U}}^2} \qquad \text{uniformly for } N \in \mathbb{N}.$$

Analytically, this theorem gives new rates of convergence for FISTA when the minimiser is not achieved in $\mathbb{H}$. Indeed for the original algorithm ($\mathbb{U}^n = \mathbb{H}$), if $u_0 = 0$ for simplicity and $(\widetilde{w}_k)_{k \in \mathbb{N}}$ is *any* $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of E exists, the result of Lemma 3 is
$$\mathrm{E}_0(u_N) \leq \inf_{w \in \mathbb{H}} \frac{\|w\|^2 + N^2 \mathrm{E}_0(w)}{2t_N^2} \leq \min_{k \in \mathbb{N}} \frac{\|\widetilde{w}_k\|^2 + N^2 \mathrm{E}_0(\widetilde{w}_k)}{2t_N^2} \lesssim \min_{k \in \mathbb{N}} \frac{a_{\mathrm{U}}^{2k} + N^2 a_{\mathrm{E}}^{-k}}{N^2} \lesssim N^{-2(1-\kappa)}. \tag{14}$$
In this sense, we could say that E_0 converges at the rate $N^{-2(1-\kappa)}$ if and only if such a sequence exists. Nothing is lost (or gained) analytically by choosing $\mathbb{U}_n \subsetneq \mathbb{H}$.

Numerically, it is easy to implement the strategy of Theorem 2 and requires very little knowledge of how to estimate $\mathrm{E}_0(u_n)$. So long as a_{U} and a_{E} can be computed analytically, one can choose $\widetilde{w}_k$ implicitly to be the discrete minimisers of some "uniform" discretisations (e.g. $\mathbb{U}^n = \{\|u\| \leq k\}$ or finite element spaces with uniform mesh) to achieve the stated convergence rate.

4.4.2 FISTA convergence with adaptivity

There are two properties of the sequence $(\mathbb{U}^n)_{n\in\mathbb{N}}$ which we may wish to decide adaptively: the refinement times n_k and the discretising spaces $\{\mathbb{U}^n \text{ s.t. } n_k \le n < n_{k+1}\}$. We will refer to these as temporal and spatial adaptivity respectively.

Lemma 4 gives a sufficient condition on n_k for converging at the rate $O(N^{2(\kappa-1)})$, but it is not necessary. Indeed for $n \le n_k$ we have

$$\mathrm{E}_0(u_n) \ge \min_{u\in\mathbb{U}^n} \mathrm{E}_0(u) = O(a_{\mathrm{E}}^{-k}) = O(n^{2(\kappa-1)}),$$

which suggests that to converge faster than $n^{2(\kappa-1)}$ requires choosing smaller n_k. As an example, in Section 7.2 we will see Algorithm 1 can converge at a near-linear rate, although this is not possible without adaptive refinement times. On the other hand, choice of spatial adaptivity has no impact on rate but can impact computational efficiency. It will be permitted to use greedy discretisation techniques so long as it is sufficient to estimate $\mathrm{E}_0(u_n)$ accurately.

Theorem 2 already allows for spatial adaptivity, so we focus on temporal adaptivity. Lemma 4 suggests that a good refinement time strategy is to choose n_k to be the minimal integer such that $\mathrm{E}_0(u_{n_k-1}) \lesssim a_{\mathrm{E}}^{-k}$. However, the value of E_0 may be hard to estimate and so we retain a "backstop" condition which guarantees that convergence is no slower than the rate given by Theorem 2. In the non-classical case of $a_{\mathrm{U}} > 1$, we provide the following theorem.

Theorem 3 *Let $(\mathbb{U}^n \subset \mathbb{H})_{n\in\mathbb{N}}$ be a sequence of subsets satisfying (10), compute u_n and v_n by Algorithm 1. Suppose that there exists a monotone increasing sequence $n_k \in \mathbb{N}$ such that*

$$\widetilde{w}_k \coloneqq u_{n_k-1} \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$

for all $k \in \mathbb{N}$.

If $(\widetilde{w}_k)_{k\in\mathbb{N}}$ is an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of E with $a_{\mathrm{U}} > 1$ and $n_k^2 \lesssim a_{\mathrm{E}}^k a_{\mathrm{U}}^{2k}$, then

$$\min_{n\le N} \mathrm{E}_0(u_n) = \min_{n\le N} \mathrm{E}(u_n) - \inf_{u\in\mathbb{H}} \mathrm{E}(u) \lesssim \frac{1}{N^{2(1-\kappa)}} \qquad \textit{where} \qquad \kappa = \frac{\log a_{\mathrm{U}}^2}{\log a_{\mathrm{E}} + \log a_{\mathrm{U}}^2}$$

uniformly for $N \in \mathbb{N}$.

The proof is given in Theorem 7. If we directly compare Theorems 2 and 3, both are a direct result of Lemma 4 assuming a specific choice of n_k or $\widetilde{w}_k$ respectively. We note that the convergence rate is the same in both theorems but the price for better adaptivity (i.e. only an upper bound on n_k) is a slightly weaker stability guarantee (now convergence of $\min_{n\le N} \mathrm{E}_0(u_n)$). In Theorem 2, as in the original FISTA algorithm, the sequence $\mathrm{E}_0(u_n)$ is not monotone but the magnitude of oscillation is guaranteed to decay in time. This behaviour is lost in Theorem 3. Although we do not prove it here, it can be shown that the stronger condition

$$\widetilde{w}_k \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1} \cap \ldots \cap \mathbb{U}^N \tag{15}$$

is sufficient to restore the stronger last-iterate guarantee on $\mathrm{E}_0(u_N)$. Again, monotonicity of $\mathbb{U}^n$ corresponds with improved stability of Algorithm 1.

To enable a more practical implementation of Theorem 3, the following lemma describes several refinement strategies which provide sufficient condition for $\mathrm{E}_0(\widetilde{w}_k) \lesssim a_{\mathrm{E}}^{-k}$.

Lemma 6 *Let $(\widetilde{w}_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{H}$ with $\|\widetilde{w}_k\| \lesssim a_{\mathrm{U}}^k$. Suppose $\widetilde{w}_k \in \widetilde{\mathbb{U}}^k \coloneqq \mathbb{U}^{n_k}$ and denote $\mathrm{E}_0(\widetilde{\mathbb{U}}^k) \coloneqq \inf_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}_0(u)$. Any of the following conditions are sufficient to show that $\widetilde{w}_k$ is an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of E:*

1. *Small continuous gap refinement: $\mathrm{E}_0(\widetilde{w}_k) \le \beta a_{\mathrm{E}}^{-k}$ for all $k \in \mathbb{N}$, some $\beta > 0$.*
2. *Small discrete gap refinement: $\mathrm{E}_0(\widetilde{\mathbb{U}}^k) \le \beta a_{\mathrm{E}}^{-k}$ and $\mathrm{E}_0(\widetilde{w}_k) - \mathrm{E}_0(\widetilde{\mathbb{U}}^{k-1}) \le \beta a_{\mathrm{E}}^{-k}$ for all $k > 0$, some $\beta > 0$.*

Otherwise, suppose there exists a Banach space $(\mathbb{U}, |||\cdot|||)$ which contains each $\widetilde{\mathbb{U}}^k$, $\sup_{k\in\mathbb{N}} |||\widetilde{w}_k||| < \infty$, and the sublevel sets of E are $|||\cdot|||$-bounded. With the subdifferential $\partial\,\mathrm{E}\colon \mathbb{U} \rightrightarrows \mathbb{U}^*$, it is also sufficient if either:

3. *Small continuous gradient refinement: $\sup_{u\in\mathbb{U}} \inf_{v\in\partial\,\mathrm{E}(\widetilde{w}_k)} \frac{|\langle v,u\rangle|}{|||u|||} \le \beta a_{\mathrm{E}}^{-k}$ for all $k \in \mathbb{N}$, some $\beta > 0$.*
4. *Small discrete gradient refinement: $\mathrm{E}_0(\widetilde{\mathbb{U}}^k) \le \beta a_{\mathrm{E}}^{-k}$ and $\sup_{u,\widetilde{w}\in\widetilde{\mathbb{U}}^k} \inf_{v\in\mathbb{V}^k} \frac{|\langle v,u-\widetilde{w}\rangle|}{|||u-\widetilde{w}|||} \le \beta a_{\mathrm{E}}^{-k}$ for all $k \in \mathbb{N}$, some $\beta > 0$, where $\mathbb{V}^k \coloneqq \partial(\mathrm{E}|_{\widetilde{\mathbb{U}}^k})(\widetilde{w}_k)$.*

The proof is given in Lemma 14. The refinement criteria described by Lemma 6 can be split into two groups. Cases (1) and (3) justify that any choice of $\mathbb{U}^{n_k}$ satisfies the required conditions, so long as $\widetilde{w}_k \in \mathbb{U}^{n_k}$. In cases (2) and (4), $\widetilde{w}_k$ is sufficient to choose the refinement time n_k, but an apriori bound is required on $\mathrm{E}_0(\widetilde{\mathbb{U}}^k)$. In these cases one could, for example, choose $\widetilde{\mathbb{U}}^k$ to be a uniform discretisation with a priori estimates.

Another splitting of the criteria is into gap and gradient computations. Typically, gradient norms (in (4) and (5)) should be easier to estimate than function gaps because they only require local knowledge rather than global, i.e. $\partial \mathrm{E}(u_n)$ rather than an estimate of $\inf_{u\in\mathbb{H}} \mathrm{E}(u)$. Implicitly, the global information comes from an extra condition on E to assert that sublevel sets are bounded.

5 General examples

We consider the main use of Algorithm 1 to be where there exists a Banach space $(\mathbb{U}, \|\!|\cdot|\!\|)$ such that $\mathbb{U} \supset \{u \in \mathbb{H} \text{ s.t. } \mathrm{E}(u) < \infty\}$ and

$$\inf_{u\in\mathbb{H}} \mathrm{E}(u) = \min_{u\in\mathbb{U}} \mathrm{E}(u) = \mathrm{E}(u^*)$$

for some $u^* \in \mathbb{U}$. The cases where $\mathbb{H}$ has finite dimension or is separable are more straightforward; if the total number of refinements is finite (i.e. $\mathbb{U}^n = \mathbb{U}^N$ for all $n \geq N$, some $N \in \mathbb{N}$), then $a_{\mathrm{U}} = 1$. This holds for most finite dimensional problems as well as the countable example discussed in detail in Section 6. In this section we give explicit computations of a_{U} and a_{E} in the setting where $\mathbb{H} = L^2(\Omega)$ for some domain $\Omega \subset \mathbb{R}^d$ and the subsets $\mathbb{U}^n$ will be finite dimensional finite-element–like spaces, as defined below.

Definition 2 Suppose $\|\cdot\|_q \lesssim \|\!|\cdot|\!\|$ (i.e. $\mathbb{U} \subset L^q(\Omega)$) for some $q \in [1,\infty]$ and connected, bounded, measurable domain $\Omega \subset \mathbb{R}^d$. We say that a collection $\mathbb{M}$ is a *mesh* if

$$\bigcup_{\omega\in\mathbb{M}} \omega \supset \Omega \qquad \text{and} \qquad |\omega \cap \omega'| = 0 \qquad \text{for all } \omega, \omega' \in \mathbb{M},\ \omega \neq \omega'.$$

Furthermore, we say a sequence of meshes $(\mathbb{M}^k)_{k\in\mathbb{N}}$ is *consistent* if there exists $\omega_0 \subset \Omega$ such that

$$\forall \omega \in \mathbb{M}^k \quad \exists(\alpha_\omega, \vec{\beta}_\omega) \in \mathbb{R}^{d\times d} \times \mathbb{R}^d \quad \text{such that} \quad \vec{x} \in \omega_0 \iff \alpha_\omega \vec{x} + \vec{\beta}_\omega \in \omega.$$

Fix $h \in (0,1)$, linear subspaces $\widetilde{\mathbb{U}}^k \subset \mathbb{H}$, and consistent meshes $\mathbb{M}^k$. We say that the sequence $(\widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ is an *h-refining sequence of finite element spaces* if there exists $c_\alpha > 0$ such that:

$$\forall(\widetilde{u}, \omega) \in \widetilde{\mathbb{U}}^k \times \mathbb{M}^k, \quad \det(\alpha_\omega) \geq c_\alpha h^{kd} \quad \text{and} \quad \exists u \in \widetilde{\mathbb{U}}^0 \quad \text{such that} \quad \forall \vec{x} \in \omega_0,\ u(\vec{x}) = \widetilde{u}(\alpha_\omega \vec{x} + \vec{\beta}_\omega).$$

We say that $(\widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ is *of order p* if for any $u^* \in \operatorname{argmin}_{u\in\mathbb{U}} \mathrm{E}(u)$ there exists a sequence $(\widetilde{w}_k)_{k\in\mathbb{N}}$ such that

$$\forall k \in \mathbb{N}, \qquad \widetilde{w}_k \in \widetilde{\mathbb{U}}^k \quad \text{and} \quad \|\!|\widetilde{w}_k - u^*|\!\| \lesssim_{u^*} h^{kp}. \tag{16}$$

We allow the implicit constant to have any dependence on u^* so long as it is finite. For example, in the case of Sobolev spaces we would expect an inequality of the form $\|\widetilde{w}_k - u^*\|_{W^{0,2}} \lesssim h^{kp} \|u^*\|_{W^{p,2}}$ [36].

Remark 1 To clarify this definition with an example, suppose we wish to approximate $L^q(\Omega)$ with piecewise linear finite elements with a triangulated mesh. Then, $\omega_0 \subset \Omega$ is a single triangle of diameter $O(h)$ and all meshes $\mathbb{M}^k$ must be triangulations of Ω with cell volumes scaling no faster than $O(h^{kd})$. The function u from the h-refining property is an arbitrary linear element, so that each $u \in \widetilde{\mathbb{U}}^k$ is linear on each $\omega \in \mathbb{M}^k$, which leads to an order $p = 2$ if $u^* \in W^{1,2}(\Omega)$.

We note that any piecewise polynomial finite element (or spline) space can be used to form a h-refining sequence of subspaces. Wavelets with a compactly supported basis behave like a multi-resolution finite element space as there is always overlap in the supports of basis vectors. Similarly, a Fourier basis does satisfy the scaling properties, but each basis vector has global support. Both of these exceptions are important and could be accounted for with further analysis but we focus on the more standard finite element case. In order to align these discretisation properties with the assumptions of Theorems 2 and 3, we make the following observation.

Lemma 7 *Fix $u^* \in \operatorname{argmin}_{u\in\mathbb{U}} \mathrm{E}(u)$ and $p', q' > 0$. If a sequence $\widetilde{w}_k \in \mathbb{H}$ satisfies*

$$\|\widetilde{w}_k\| \lesssim h^{-kq'} \quad \text{and} \quad \mathrm{E}(\widetilde{w}_k) - \mathrm{E}(u^*) \lesssim h^{kp'},$$

then $(\widetilde{w}_k)_{k\in\mathbb{N}}$ is an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of E for $a_{\mathrm{U}} = h^{-q'}$ and $a_{\mathrm{E}} = h^{-p'}$.

This is precisely rewriting the statement of Definition 1 into terms of resolution h. The following theorem links p and q from Definition 2 with p' and q' from Lemma 7.

Theorem 4 *Suppose $\mathbb{H} = L^2(\Omega)$ for some connected, bounded domain $\Omega \subset \mathbb{R}^d$ and $\|\cdot\|_q \lesssim |||\cdot|||$ for some $q \in [1,\infty]$. For $p \geq 0$ and $h \in (0,1)$, if $(\widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ is an h-refining sequence of finite element spaces of order p, then $(\widetilde{w}_k)_{k\in\mathbb{N}}$ is an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of E for*

$$a_{\mathrm{U}} \leq \begin{cases} 1 & \text{if } q \geq 2, \\ \sqrt{h^{-d}} & q < 2 \text{ and } \sup_{u\in\widetilde{\mathbb{U}}^0} \frac{\|u\|_{L^\infty(\omega_0)}}{\|u\|_{L^2(\omega_0)}} < \infty \end{cases}, \qquad a_{\mathrm{E}} \geq \begin{cases} h^{-2p} & \text{if } \nabla\mathrm{E} \text{ is } |||\cdot|||\text{-Lipschitz at } u^*, \\ h^{-p} & \text{if } \mathrm{E} \text{ is } |||\cdot|||\text{-Lipschitz at } u^*, \\ 1 & \text{otherwise.} \end{cases}$$

The proof of this theorem is in Appendix B. Note that $\sup_{u\in\widetilde{\mathbb{U}}^0} \|u\|_{L^\infty(\omega_0)} \|u\|^{-1}_{L^2(\omega_0)}$ is finite whenever $\widetilde{\mathbb{U}}^0 \subset L^\infty(\Omega)$ is finite dimensional, so this is not a very strong assumption. The main take-home for this theorem is that the computation of a_{U} and a_{E} is typically very simple and clear given a particular choice of $|||\cdot|||$ and E. We also briefly remark that the Lipschitz constants in this lemma do not need to be valid globally, only on the sequence $\widetilde{w}_k$. The same result holds under a local-Lipschitz assumption, for example on the ball of radius $\sup_{k\in\mathbb{N}} |||\widetilde{w}_k|||$ which is finite whenever $p \geq 0$.

6 L1 penalised reconstruction

The canonical example for FISTA is the LASSO problem with a quadratic data fidelity and L1 regularisation. In this section we develop the necessary analytical tools for the variant with general smooth fidelity term which will be used for numerical results in Section 7. We consider three forms which will be referred to as the continuous, countable, and discrete problem depending on whether the space $\mathbb{U}$ is $\mathscr{M}([0,1]^d)$, $\ell^1(\mathbb{R})$, or $\mathbb{R}^M$ respectively. We choose $\mathbb{H}$ to be $L^2([0,1]^d)$, $\ell^2(\mathbb{R})$, or $\mathbb{R}^M$ correspondingly. Let $\mathsf{A}\colon \mathbb{U}\cap\mathbb{H} \to \mathbb{R}^m$ be a linear operator represented by the kernels $\psi_j \in \mathbb{H}$ such that

$$\forall u \in \mathbb{U}\cap\mathbb{H},\ j = 1,\dots,m, \qquad (\mathsf{A}u)_j = \langle \psi_j, u \rangle. \tag{17}$$

In the continuous case we will assume the additional smoothness $\psi_j \in C^1([0,1]^d)$. In Section 6.5 we will formally define and estimate several operator semi-norms for A of this form, for example Lemma 8 confirms that A is continuous on $\mathbb{H}$ (without loss of generality $\|\mathsf{A}\| \leq 1$). In each case, the energy we consider is written as

$$\mathrm{E}(u) = \mathrm{f}(\mathsf{A}u - \eta) + \mu|||u||| \tag{18}$$

for some $\mu > 0$ where $|||\cdot||| = \|\cdot\|_1$. We assume $\mathrm{f} \in C^1(\mathbb{R}^m)$ is convex, bounded from below, and $\nabla\mathrm{f}$ is 1-Lipschitz. Let $u^* \in \operatorname{argmin}_{u\in\mathbb{U}} \mathrm{E}(u)$, which is non-empty so long as $\psi_j \in C([0,1]^d)$, see the proof of [7, Prop. 3.1] when f is quadratic.

The aim of this section is to develop all of the necessary tools for implementing Algorithm 1 on the energy (18) using the convergence guarantees of either Theorem 2 or Theorem 3. This includes computing the rates a_{U} and a_{E}, estimating the continuous gap $\mathrm{E}_0(u_n)$, and developing an efficient refinement choice for $\mathbb{U}^n$. Below we will just describe the form of $\widetilde{\mathbb{U}}^k$ under the assumption that $\mathbb{U}^n \subset \widetilde{\mathbb{U}}^k$ is chosen adaptively for $n = n_{k-1}+1,\dots,n_k$. The index k refers to the scale or resolution and n refers to the iteration number of the reconstruction algorithm.

6.1 Continuous case

We start by estimating rates in the case $\mathbb{U} = \mathscr{M}(\Omega)$ where $\Omega = [0,1]^d$. In this case we choose $\widetilde{\mathbb{U}}^k$ to be the span of all piecewise constant functions on a mesh of squares with maximum side length 2^{-k} (i.e. $h = \frac{1}{2}$) and

$$\widetilde{w}_k \coloneqq \sum_{\omega \in \mathbb{M}^k} \frac{u^*(\omega)}{|\omega|} \mathbb{1}_\omega \quad \text{where} \quad \mathbb{1}_\omega(\vec{x}) = \begin{cases} 1 & \vec{x} \in \omega \\ 0 & \text{else} \end{cases}.$$

By construction $\widetilde{w}_k \in \widetilde{\mathbb{U}}^k$, however note that for any $u \in L^1(\Omega)$ and Dirac mass δ supported in $(0,1)^d$,

$$|||u - \delta||| = \sup_{\varphi \in C(\Omega), \|\varphi\|_{L^\infty} \leq 1} \langle \varphi, u - \delta \rangle = |||u||| + |||\delta||| \geq 1 = h^0. \tag{19}$$

Because of this, application of Theorem 4 with $p = 0$ gives $a_{\mathrm{U}} = 2^{\frac{d}{2}}$ but only $a_{\mathrm{E}} \geq 1$. To improve our estimate of a_{E} requires additional assumptions on A. Note that $|||\widetilde{w}_k||| = \sum_{\omega \in \mathbb{M}^k} |u^*(\omega)| \leq |||u^*|||$, therefore we have

$$\begin{aligned} \mathrm{E}(\widetilde{w}_k) - \mathrm{E}(u^*) &= \mathrm{f}(\mathsf{A}\widetilde{w}_k - \eta) - \mathrm{f}(\mathsf{A}u^* - \eta) + \mu\left(|||\widetilde{w}_k||| - |||u^*|||\right) && (20) \\ &\leq \nabla \mathrm{f}(\mathsf{A}\widetilde{w}_k - \eta) \bullet \mathsf{A}(\widetilde{w}_k - u^*) && (21) \\ &\leq \left[\|\nabla \mathrm{f}(\mathsf{A}u^* - \eta)\|_{\ell^2} + \|\mathsf{A}(\widetilde{w}_k - u^*)\|_{\ell^2}\right] \|\mathsf{A}(\widetilde{w}_k - u^*)\|_{\ell^2}. && (22) \end{aligned}$$

as f is convex with 1-Lipschitz gradient. Clearly $\|\nabla \mathrm{f}(\mathsf{A}u^* - \eta)\|_{\ell^2}$ is a constant. For the other term, for all $\vec{r} \in \mathbb{R}^m$ denote $\varphi \coloneqq \mathsf{A}^*\vec{r}$, then note that

$$\vec{r} \bullet \mathsf{A}(\widetilde{w}_k - u^*) = \langle \varphi, \widetilde{w}_k - u^* \rangle = \sum_{\omega \in \mathbb{M}^k} \int_\omega \varphi(\vec{x}) \, \mathrm{d}[\widetilde{w}_k - u^*] = \sum_{\omega \in \mathbb{M}^k} |\omega|^{-1} \iint_{\omega^2} [\varphi(\vec{x}) - \varphi(\vec{y})] \, \mathrm{d}\vec{x} \mathrm{d}u^*(\vec{y}). \tag{23}$$

With the pointwise bound $|\varphi(\vec{x}) - \varphi(\vec{y})| \leq \mathrm{diam}(\omega) \|\nabla\varphi\|_{L^\infty} = \sqrt{d} 2^{-k} \|\nabla[\mathsf{A}^*\vec{r}]\|_{L^\infty}$, we deduce the estimate

$$\|\mathsf{A}(\widetilde{w}_k - u^*)\|_{\ell^2} = \sup_{\vec{r} \in \mathbb{R}^m} \|\vec{r}\|_{\ell^2}^{-1} \langle \mathsf{A}^*\vec{r}, \widetilde{w}_k - u^* \rangle \leq \sqrt{d} 2^{-k} |||u^*||| \sup_{\vec{r} \in \mathbb{R}^m} \|\vec{r}\|_{\ell^2}^{-1} \|\nabla[\mathsf{A}^*\vec{r}]\|_{L^\infty}. \tag{24}$$

In Lemma 9 we will show that this last term, which we denote the semi-norm $|\mathsf{A}^*|_{\ell^2 \to C^1}$, is bounded by $\sqrt{m} \max_{j \in [m]} \left\|\nabla \Psi_j\right\|_\infty$. We conclude that $\mathrm{E}(\widetilde{w}_k) - \mathrm{E}(u^*) \lesssim 2^{-k}$. In particular, this computation confirms two things. Firstly that the scaling constant is $a_{\mathrm{E}} = 2$, and secondly that the required smoothness to achieve a good rate with Algorithm 1 is that $\mathsf{A}^* \colon \mathbb{R}^m \to C^1(\Omega)$ is a bounded operator. This accounts for using the weaker topology of $\mathscr{M}(\Omega)$ rather than $L^1(\Omega)$.

Inserting the computed rates into Theorem 2 or Theorem 3 gives the guaranteed convergence rate

$$\kappa = \frac{\log a_{\mathrm{U}}^2}{\log a_{\mathrm{E}} + \log a_{\mathrm{U}}^2} = \frac{d}{1+d} \quad \Longrightarrow \quad \mathrm{E}(u_n) - \inf_{u \in \mathbb{H}} \mathrm{E}(u) \lesssim n^{-2(1-\kappa)} = n^{-\frac{2}{1+d}}. \tag{25}$$

This rate can be used to infer the required resolution at each iteration, in particular on iteration n with $n^2 \simeq (a_{\mathrm{E}} a_{\mathrm{U}}^2)^k$ we expect the resolution to be

$$2^{-k} = \left(a_{\mathrm{E}} a_{\mathrm{U}}^2\right)^{\frac{k}{1+d}} \simeq n^{-\frac{2}{1+d}}. \tag{26}$$

6.2 Countable and discrete case

We now extend the rate computations to the case when $\mathbb{U} = \ell^1(\mathbb{R})$, or a finite dimensional subspace. The key fact here is that, even when $\mathbb{U}$ is infinite dimensional, it is known (e.g. [38, Thm 6] and [6, Cor 3.8]) that there exists $u^* \in \mathrm{argmin}_{u \in \mathbb{U}} \mathrm{E}(u)$ with at most m non-zeros. If this is the case, then $u^* \in \ell^2(\mathbb{R})$, indeed $\|u^*\|_{\ell^2} \leq \sqrt{m} \|u^*\|_{\ell^1}$. This makes the estimates of $a_{\mathrm{E}}/a_{\mathrm{U}}$ much simpler than in the continuous case as we can stay in the finite-dimensional Hilbert-space setting.

For countable dimensions we consider discretisation subspaces of the form

$$\widetilde{\mathbb{U}}^k = \{u \in \ell^1(\mathbb{R}) \text{ s.t. } i \notin J_k \implies u_i = 0\}$$

for some sets $J_k \subset \mathbb{N}$, i.e. infinite vectors with finitely many non-zeros. The key change in analysis from the continuous case is $\|u^*\| < \infty$, so $a_{\mathrm{U}} = 1$ and the expected rate of n^{-2}, independent of a_{E} or any additional properties of A. The number of refinements will also be finite, therefore $n_k = \infty$ for some k, the remaining conditions of Theorems 2 and 3 hold trivially.

6.3 Refinement metrics

Lemma 6 shows that adaptive refinement can be performed based on estimates of the function gap or the subdifferential. In this subsection we provide estimates for the forth case of Lemma 6 which can be easily computed. In this case we consider $\partial \mathrm{E}\colon \mathbb{H} \rightrightarrows \mathbb{H}$ so that subdifferentials are well behaved, for example for explicit computation assuming validity of the chain/sum rules for differentiation.

6.3.1 Bounds for discretised functionals

We start by computing estimates for discretised energies. This covers the cases when either the continuous/countable energy is projected onto $\mathbb{U}^n$, or $\mathbb{U}$ is finite dimensional. For notation we will use the continuous case, to recover the other cases just replace continuous indexing with discrete (i.e. $u(\vec{x}) \rightsquigarrow u_i$).

Let $\Pi_n\colon \mathbb{H} \to \mathbb{U}^n$ denote the orthogonal projection. We consider the discretised function $\mathrm{E}|_{\mathbb{U}^n}\colon \mathbb{U}^n \to \mathbb{R}$ and its subdifferential $\partial_n \mathrm{E}(\cdot) = \Pi_n \partial \mathrm{E}(\cdot)$ on $\mathbb{U}^n$. In our case, the behaviour of $\mathrm{E}|_{\mathbb{U}^n}$ is equivalent to replacing u with $\Pi_n u$, and A^* with $\Pi_n \mathsf{A}^*$.

Discrete gradient We can use Π_n to compute the discrete subdifferential at $u_n \in \mathbb{U}^n$:

$$\partial_n \mathrm{E}(u_n)(\vec{x}) = [\Pi_n \mathsf{A}^* \nabla \mathrm{f}(\mathsf{A}u_n - \eta)](\vec{x}) + \begin{cases} \{+\mu\} & u_n(\vec{x}) > 0 \\ [-\mu, \mu] & u_n(\vec{x}) = 0 \\ \{-\mu\} & u_n(\vec{x}) < 0 \end{cases} \tag{27}$$

$$=: [\Pi_n \mathsf{A}^* \nabla \mathrm{f}(\mathsf{A}u_n - \eta)](\vec{x}) + \mu \Pi_n \operatorname{sign}(u_n(\vec{x})) \tag{28}$$

where we define $s + \mu[-1,1] = [s-\mu, s+\mu]$ for all $s \in \mathbb{R}$, $\mu \geq 0$.

As $|||\cdot||| = \|\cdot\|_1$, the natural metric for $\partial_n \mathrm{E}$ is $|||\cdot|||_* = \|\cdot\|_\infty$ which we can estimate

$$|||\partial_n \mathrm{E}(u_n)|||_* = \max_{\vec{x}\in\Omega} \min_v \{|v| \text{ s.t. } v \in \Pi_n \mathsf{A}^* \nabla \mathrm{f}(\mathsf{A}u_n - \eta)(\vec{x}) + \mu \Pi_n \operatorname{sign}(u_n(\vec{x}))\} \tag{29}$$

$$= \max_{\vec{x}\in\Omega} \begin{cases} |[\Pi_n \mathsf{A}^* \nabla \mathrm{f}(\mathsf{A}u_n - \eta)(\vec{x}) + \mu| & u_n(\vec{x}) > 0 \\ |[\Pi_n \mathsf{A}^* \nabla \mathrm{f}(\mathsf{A}u_n - \eta)(\vec{x}) - \mu| & u_n(\vec{x}) < 0 \\ \max\left(|\Pi_n \mathsf{A}^* \nabla \mathrm{f}(\mathsf{A}u_n - \eta)(\vec{x})| - \mu, 0\right) & u_n(\vec{x}) = 0 \end{cases} \tag{30}$$

which can be used directly in Lemma 6.

Discrete gap We now move on to the discrete gap, $\mathrm{E}(u_n) - \min_{u\in\mathbb{U}^n} \mathrm{E}(u)$. This can be computed with a dual representation (e.g. [15]),

$$\min_{u\in\mathbb{U}^n} \mathrm{f}(\mathsf{A}u - \eta) + \mu |||u||| = \min_{u\in\mathbb{H}} \max_{\vec{\varphi}\in\mathbb{R}^m} (\mathsf{A}\Pi_n u - \eta) \bullet \vec{\varphi} + \mu |||\Pi_n u||| - \mathrm{f}^*(\vec{\varphi}) \tag{31}$$

$$= \max_{\vec{\varphi}\in\mathbb{R}^m} \min_{u\in\mathbb{H}} (\mathsf{A}\Pi_n u - \eta) \bullet \vec{\varphi} + \mu |||\Pi_n u||| - \mathrm{f}^*(\vec{\varphi}) \tag{32}$$

$$= \max_{\vec{\varphi}\in\mathbb{R}^m} \begin{cases} -\eta \bullet \vec{\varphi} - \mathrm{f}^*(\vec{\varphi}) & |||\Pi_n \mathsf{A}^* \vec{\varphi}|||_* \leq \mu \\ -\infty & \text{else} \end{cases} \tag{33}$$

$$= - \min_{\vec{\varphi}\in\mathbb{R}^m} \underbrace{\mathrm{f}^*(\vec{\varphi}) + \eta \bullet \vec{\varphi}}_{=:\mathrm{E}^\dagger(\vec{\varphi})} + \chi(|||\Pi_n \mathsf{A}^* \vec{\varphi}|||_* \leq \mu). \tag{34}$$

In particular,

$$\mathrm{E}(u) - \min_{u\in\mathbb{U}^n} \mathrm{E}(u) = \mathrm{E}(u) + \min_{\vec{\varphi}\in\mathbb{R}^m \text{ s.t. } |||\Pi_n \mathsf{A}^* \vec{\varphi}|||_* \leq \mu} \mathrm{E}^\dagger(\vec{\varphi}) \leq \mathrm{E}(u) + \mathrm{E}^\dagger(\vec{\varphi}) \tag{35}$$

for any feasible $\vec{\varphi} \in \mathbb{R}^m$. We further derive the criticality condition, if $(u^*, \vec{\varphi}^*)$ is a saddle point, then

$$\mathsf{A}u^* - \eta \in \partial \mathrm{f}^*(\vec{\varphi}^*), \qquad \text{or equivalently} \qquad \vec{\varphi}^* = \nabla \mathrm{f}(\mathsf{A}u^* - \eta). \tag{36}$$

We remark briefly that $\mathrm{E}^\dagger$ should be thought of as the dual of E but without the constraint. We choose to omit it here to highlight that it is only the constraint which changes between the discrete and continuous cases; the value of $\mathrm{E}^\dagger$ will remain the same.

Given $u_n \in \mathbb{U}^n$, the optimality condition motivates a simple rule for choosing $\vec{\varphi}$:

$$\vec{\varphi}_n := \nabla \mathrm{f}(\mathsf{A}u_n - \eta), \qquad \mathrm{E}(u) - \min_{u' \in \mathbb{U}^n} \mathrm{E}(u') \leq \mathrm{E}(u) + \mathrm{E}^\dagger(\gamma \vec{\varphi}_n) \tag{37}$$

for some $0 \leq \gamma \leq \frac{\mu}{\|\Pi_n \mathsf{A}^* \vec{\varphi}_n\|_*}$. In the case $\mathrm{f}(\cdot) = \frac{1}{2}\|\cdot\|_{\ell^2}^2$, one can use the optimal choice

$$\gamma = \max\left(0, \min\left(\frac{-\eta \bullet \vec{\varphi}_n}{\|\vec{\varphi}_n\|_{\ell^2}^2}, \frac{\mu}{\|\Pi_n \mathsf{A}^* \vec{\varphi}_n\|_*}\right)\right). \tag{38}$$

To apply Algorithm 1, we are assuming that both $\mathrm{f}(\mathsf{A}u_n - \eta)$ and $\Pi_n \nabla \mathrm{f}(\mathsf{A}u_n - \eta)$ are easily computable, therefore γ and $\mathrm{E}(u_n) + \mathrm{E}^\dagger(\gamma \vec{\varphi}_n)$ are also easy to compute.

6.3.2 Bounds for countable functionals

Extending the results of Section 6.3.1 to $\mathbb{U} = \ell^1(\mathbb{R})$ is analytically very simple but computationally relies heavily on the specific choice of A. The computations of subdifferentials and gaps carry straight over replacing Π_n with the identity and adding the sets $J_n \subset \mathbb{N}$ which define $\mathbb{U}^n = \{u \in \ell^1 \text{ s.t. } i \notin J_n \implies u_i = 0\}$. Recall that $\|\partial \mathrm{E}(u_n)\|_* := \inf_{s \in \mathrm{sign}(u_n)} \|\mathsf{A}^* \vec{\varphi}_n + \mu s\|_*$ where the sign function has the pointwise set-valued definition as indicated in (27)-(28). Where $[u_n]_i = 0$, the choice $s_i = \min(1, \max(-1, -\mu^{-1}[\mathsf{A}^* \vec{\varphi}_n]_i))$ achieves the minimal value

$$\|\partial \mathrm{E}(u_n)\|_* = \max_{i \in \mathbb{N}} \begin{cases} |[\mathsf{A}^* \vec{\varphi}_n]_i + \mu| & [u_n]_i > 0 \\ |[\mathsf{A}^* \vec{\varphi}_n]_i - \mu| & [u_n]_i < 0 \\ \max(|[\mathsf{A}^* \vec{\varphi}_n]_i| - \mu, 0) & [u_n]_i = 0 \end{cases} \tag{39}$$

$$\mathrm{E}(u_n) - \inf_{u \in \mathbb{H}} \mathrm{E}(u) \leq \mathrm{E}(u_n) + \mathrm{E}^\dagger(\gamma_0 \vec{\varphi}_n), \qquad \gamma_0 \in \left[0, \frac{\mu}{\|\mathsf{A}^* \vec{\varphi}_n\|_*}\right] \tag{40}$$

where $\vec{\varphi}_n = \nabla \mathrm{f}(\mathsf{A}u_n - \eta) \in \mathbb{R}^m$ is always exactly computable.

In the countable case, the sets J_n give a clear partition into known/unknown values in these definitions. For $i \in J_n$ the computation is the same as in Section 6.3.1, then for $i \notin J_n$ we know $[u_n]_i = 0$ which simplifies the remaining computations. This leads to:

$$\|\partial \mathrm{E}(u_n)\|_* = \max\left(\max_{i \in J_n} |[\partial \mathrm{E}(u_n)]_i|, \sup_{i \notin J_n} |[\partial \mathrm{E}(u_n)]_i|\right) = \max\left(\|\partial_n \mathrm{E}(u_n)\|_*, \sup_{i \notin J_n} |[\mathsf{A}^* \vec{\varphi}_n]_i| - \mu\right) \tag{41}$$

$$\|\mathsf{A}^* \vec{\varphi}_n\|_* = \max\left(\max_{i \in J_n} |[\mathsf{A}^* \vec{\varphi}_n]_i|, \sup_{i \notin J_n} |[\mathsf{A}^* \vec{\varphi}_n]_i|\right) \qquad = \max\left(\|\Pi_n \mathsf{A}^* \vec{\varphi}_n\|_*, \sup_{i \notin J_n} |[\mathsf{A}^* \vec{\varphi}_n]_i|\right). \tag{42}$$

Both estimates only rely on an upper bound of $\max_{i \notin J_n} |[\mathsf{A}^* \vec{\varphi}_n]_i|$. One example computing this value is seen in Section 7.2.

6.3.3 Bounds for continuous functionals

Finally we extend the results of Section 6.3.1 to continuous problems. Similar to the countable case (39)-(40), the exact formulae can be written down immediately:

$$\|\partial \mathrm{E}(u_n)\|_* = \max_{\vec{x} \in \Omega} \begin{cases} |[\mathsf{A}^* \vec{\varphi}_n](\vec{x}) + \mu| & u_n(\vec{x}) > 0 \\ |[\mathsf{A}^* \vec{\varphi}_n](\vec{x}) - \mu| & u_n(\vec{x}) < 0 \\ \max(|[\mathsf{A}^* \vec{\varphi}_n](\vec{x})| - \mu, 0) & u_n(\vec{x}) = 0 \end{cases} \tag{43}$$

$$\mathrm{E}(u_n) - \inf_{u \in \mathbb{H}} \mathrm{E}(u) \leq \mathrm{E}(u_n) + \mathrm{E}^\dagger(\gamma_0 \vec{\varphi}_n), \qquad \gamma_0 \in \left[0, \frac{\mu}{\|\mathsf{A}^* \vec{\varphi}_n\|_*}\right] \tag{44}$$

with $\mathrm{E}^\dagger$ as defined in (34). Recall that there is a mesh $\mathbb{M}^n$ corresponding to $\mathbb{U}^n$ such that u_n is constant on each $\omega \in \mathbb{M}^n$, so we can rewrite these bounds:

$$\|\!|\partial \mathrm{E}(u_n)|\!\|_* = \max_{\omega\in\mathbb{M}^n} \begin{cases} \|\mathsf{A}^*\vec{\varphi}_n + \mu\|_{L^\infty(\omega)} & u_n|_\omega > 0 \\ \|\mathsf{A}^*\vec{\varphi}_n - \mu\|_{L^\infty(\omega)} & u_n|_\omega < 0 \\ \max(0, \|\mathsf{A}^*\vec{\varphi}_n\|_{L^\infty(\omega)} - \mu) & u_n|_\omega = 0 \end{cases} \tag{45}$$

$$\|\!|\mathsf{A}^*\vec{\varphi}_n|\!\|_* = \max_{\omega\in\mathbb{M}^n} \|\mathsf{A}^*\vec{\varphi}_n\|_{L^\infty(\omega)}. \tag{46}$$

Now, both values can be estimated relying on pixel-wise supremum norms of $\mathsf{A}^*\vec{\varphi}_n$ which we have assumed is sufficiently smooth. We will therefore use a pixel-wise Taylor expansion to provide a simple and accurate estimate. For instance, let $\vec{x}_i$ be the midpoint of the pixel ω, then

$$\|\mathsf{A}^*\vec{\varphi}_n\|_{L^\infty(\omega)} \leq |[\mathsf{A}^*\vec{\varphi}_n](\vec{x}_i)| + \frac{\operatorname{diam}(\omega)}{2}|[\nabla\mathsf{A}^*\vec{\varphi}_n](\vec{x}_i)| + \frac{\operatorname{diam}(\omega)^2}{8}|\mathsf{A}^*\vec{\varphi}_n|_{C^2}. \tag{47}$$

In this work we chose a first order expansion because we are looking for extrema of $\mathsf{A}^*\vec{\varphi}_n$, i.e. we are most interested in the squares ω such that

$$|[\mathsf{A}^*\vec{\varphi}_n](\vec{x}_i)| \approx \mu, \qquad |[\nabla\mathsf{A}^*\vec{\varphi}_n](\vec{x}_i)| \approx 0, \qquad [\nabla^2\mathsf{A}^*\vec{\varphi}_n](\vec{x}_i) \preceq 0. \tag{48}$$

A zeroth order expansion would be optimally inefficient (approximating $|[\nabla\mathsf{A}^*\vec{\varphi}_n](\vec{x}_i)|$ with $|\mathsf{A}^*\vec{\varphi}_n|_{C^1}$) and a second order expansion would possibly be more elegant but harder to implement. We found that a first order expansion was simple and efficient.

The bounds presented here for continuous problems emphasise the twinned properties required for adaptive mesh optimisation. The mesh should be refined greedily to the structures of u^*, but also must be sufficiently uniform to provide a good estimate for $\mathrm{E}(u^*)$. This is a classical exploitation/exploration trade-off; exploiting visible structure whilst searching for other structures which are not yet visible.

6.4 Support detection

The main motivation for using L1 penalties in applications is because it recovers sparse signals, in the case of compressed sensing the support of u^* is also provably close to the "true" support [15,29]. If $u_n \approx u^*$ in the appropriate sense, then we should also be able to quantify the statement $\operatorname{supp}(u_n) \approx \operatorname{supp}(u^*)$. Such methods are referred to as *safe screening* rules [18] which gradually identify the support and allow the optimisation algorithm to constrain parts of the reconstruction to 0. In this subsection we propose a new simple screening rule which is capable of generalising to our continuous subspace approximation setting. It is likely that more advanced methods [4,25] can also be adapted, although that is beyond the scope of this work. The key difference is the allowance of inexact computations resulting from estimates such as (47).

The support of u^* has already been characterised very precisely [15,29]. In particular, the support is at most m distinct points and are a subset of $\{\vec{x} \in \Omega \text{ s.t. } |\mathsf{A}^*\vec{\varphi}^*|(\vec{x}) = \mu\}$ (an equivalent statement holds for the countable case). Less formally, this can also be seen from the the subdifferential computations in Section 6.3, for all $\vec{x} \in \operatorname{supp}(u^*)$ we have

$$0 \in \partial \mathrm{E}(u^*)(\vec{x}) = [\mathsf{A}^*\vec{\varphi}^*](\vec{x}) + \mu \operatorname{sign}(u^*(\vec{x})). \tag{49}$$

Heuristically, we will use strong convexity of $\mathrm{E}^\dagger$ from (34) and smoothness of A^* to quantify the statement:

$$\text{if} \quad \mathrm{E}(u_n) + \mathrm{E}^\dagger(\gamma_0\vec{\varphi}_n) \approx 0 \quad \text{then} \quad \{\vec{x} \text{ s.t. } |[\mathsf{A}^*\vec{\varphi}_n](\vec{x})| \ll \mu\} \subset \{\vec{x} \text{ s.t. } u^*(\vec{x}) = 0\}.$$

Recall that $\nabla \mathrm{f}$ is 1-Lipschitz if and only if f^* is 1-strongly convex [19, Chapter 10, Thm. 4.2.2]. Therefore, if $\gamma_0\vec{\varphi}_n$ and $\vec{\varphi}^*$ are both dual-feasible, then

$$\tfrac{1}{2}\|\gamma_0\vec{\varphi}_n - \vec{\varphi}^*\|_{\ell^2}^2 \leq \mathrm{E}^\dagger(\gamma_0\vec{\varphi}_n) - \mathrm{E}^\dagger(\vec{\varphi}^*) = \mathrm{E}^\dagger(\gamma_0\vec{\varphi}_n) + \mathrm{E}(u^*) \leq \mathrm{E}^\dagger(\gamma_0\vec{\varphi}_n) + \mathrm{E}(u_n), \tag{50}$$

which gives an easily computable bound on $\|\gamma_0\vec{\varphi}_n - \vec{\varphi}^*\|_{\ell^2}$. Now we estimate $\mathsf{A}^*\vec{\varphi}_n$ on the support of u^*:

$$\min_{\vec{x}\in\operatorname{supp}(u^*)} |[\Pi_n\mathsf{A}^*\vec{\varphi}_n](\vec{x})| \geq \min_{\vec{x}\in\operatorname{supp}(u^*)} |[\mathsf{A}^*\vec{\varphi}_n](\vec{x})| \tag{51}$$

$$= \gamma_0^{-1} \min_{\vec{x}\in\operatorname{supp}(u^*)} |[\mathsf{A}^*\gamma_0\vec{\varphi}_n](\vec{x})| \tag{52}$$

$$\geq \gamma_0^{-1} \min_{\vec{x}\in\operatorname{supp}(u^*)} |[\mathsf{A}^*\vec{\varphi}^*](\vec{x})| - |[\mathsf{A}^*\gamma_0\vec{\varphi}_n - \mathsf{A}^*\vec{\varphi}^*](\vec{x})| \tag{53}$$

$$= \gamma_0^{-1} \min_{\vec{x}\in\operatorname{supp}(u^*)} \mu - |[\mathsf{A}^*\gamma_0\vec{\varphi}_n - \mathsf{A}^*\vec{\varphi}^*](\vec{x})| \tag{54}$$

$$\geq \gamma_0^{-1} \left(\mu - |\mathsf{A}^*|_{\ell^2\to L^\infty}\|\gamma_0\vec{\varphi}_n - \vec{\varphi}^*\|_{\ell^2}\right). \tag{55}$$

Therefore,

$$|[\Pi_n\mathsf{A}^*\vec{\varphi}_n](\vec{x})| < \gamma_0^{-1}\left(\mu - \sqrt{2(\mathrm{E}(u_n)+\mathrm{E}^\dagger(\gamma_0\vec{\varphi}_n))}|\mathsf{A}^*|_{\ell^2\to L^\infty}\right) \qquad \Longrightarrow \qquad u^*(\vec{x}) = 0. \tag{56}$$

This equation is valid when $\vec{x}$ is either a continuous or countable index, the only distinction is to switch to ℓ^∞ in the norm of A^*. To make the equivalent statement on the discretised problem, simply replace γ_0 with γ and A^* with $\Pi_n\mathsf{A}^*$. There are two short observations on this formula:

- The convergence guarantee from Theorem 2 is for the primal gap $\mathrm{E}(u_n) - \mathrm{E}(u^*)$, rather than the primal-dual gap $\mathrm{E}(u_n) + \mathrm{E}^\dagger(\gamma_0\vec{\varphi}_n)$ used here. Although there is no guaranteed rate for the primal-dual gap, it is much more easily computable than the primal gap.
- In Section 6.1, $|\mathsf{A}^*|_{\ell^2\to C^1} < \infty$ was required to compute a rate of convergence for $\mathrm{E}(u_n)$, but only $|\mathsf{A}^*|_{\ell^2\to L^\infty} < \infty$ is needed to estimate the support.

6.5 Operator norms

For numerical implementation of (18), we are required to accurately estimate several operator norms of A of the form in (17). In particular, there are kernels $\psi_j \in \mathbb{H}$ such that $(\mathsf{A}u)_j = \langle \psi_j, u\rangle$ for each $j \in [m]$. Verifying that $\|\mathsf{A}\| \leq 1$ can be performed by computing $|\mathsf{A}\mathsf{A}^*|_{\ell^2\to\ell^2}$, and the adaptivity described in Sections 6.1, 6.3.3, and 6.4 requires the values of $|\mathsf{A}^*|_{\ell^2\to L^\infty}$, $|\mathsf{A}^*|_{\ell^2\to C^1}$, and $|\mathsf{A}^*|_{\ell^2\to C^2}$. The aim for this section is to provide estimates of these norms and seminorms for the numerical examples presented in Section 7.

The following lemma allows for exact computation of the operator norm of A.

Lemma 8 *If $\mathsf{A}\colon \mathbb{H}\to\mathbb{R}^m$ has kernels $\psi_j \in \mathbb{H}$ for $j\in[m]$, then $\mathsf{A}\mathsf{A}^* \in \mathbb{R}^{m\times m}$ has entries $(\mathsf{A}\mathsf{A}^*)_{i,j} = \langle\psi_i,\psi_j\rangle$, so the spectral norm $\|\mathsf{A}^*\mathsf{A}\| = \|\mathsf{A}\mathsf{A}^*\|$ can be computed efficiently.*

Proof To compute the entries of $\mathsf{A}\mathsf{A}^*\colon \mathbb{R}^m\to\mathbb{R}^m$, observe that for any $\vec{r}\in\mathbb{R}^m$

$$(\mathsf{A}\mathsf{A}^*\vec{r})_i = \langle\psi_i, \mathsf{A}^*\vec{r}\rangle = \left\langle \psi_i, \sum_{j=1}^m r_j\psi_j\right\rangle = \sum_{j=1}^m \langle\psi_i,\psi_j\rangle r_j \tag{57}$$

as required. □

If $\|\mathsf{A}^*\mathsf{A}\|$ is not analytically tractable, then Lemma 8 enables it to be computed using standard finite dimensional methods. The operator $\mathsf{A}\mathsf{A}^*$ is always finite dimensional, and can be computed without discretisation error.

In the continuous case, when $\mathbb{H} = L^2(\Omega)$ we also need to estimate the smoothness properties of A^*. A generic result for this is given in the following lemma.

Lemma 9 *If $\mathsf{A}\colon L^2([0,1]^d)\to\mathbb{R}^m$ has kernels $\psi_j\in L^2(\Omega)\cap C^k(\Omega)$ for $j\in[m]$, then for all $\frac{1}{q}+\frac{1}{q^*}=1$, $q\in[1,\infty]$, we have*

$$|\mathsf{A}^*\vec{r}|_{C^k} := \sup_{\vec{x}\in\Omega}|\nabla^k[\mathsf{A}^*\vec{r}]|(\vec{x}) \leq \sup_{\vec{x}\in\Omega}\left\|(\nabla^k\psi_j(\vec{x}))_{j=1}^m\right\|_{\ell q^*}\|\vec{r}\|_{\ell^q}, \tag{58}$$

$$|\mathsf{A}^*|_{\ell^2\to C^k} := \sup_{\|\vec{r}\|_{\ell^2}\leq 1}|\mathsf{A}^*\vec{r}|_{C^k} \leq \sup_{\vec{x}\in\Omega}\left\|(\nabla^k\psi_j(\vec{x}))_{j=1}^m\right\|_{\ell q^*}\times\begin{cases}1 & q\geq 2\\ \sqrt{m^{2-q}} & q<2\end{cases}. \tag{59}$$

Proof For the first inequality, we apply the Hölder inequality on $\mathbb{R}^m$:

$$|\nabla^k[\mathsf{A}^*\vec{r}]|(\vec{x}) = \left|\sum_{j=1}^m \nabla^k \psi_j(\vec{x}) r_j\right| \leq \left(\sum_{j=1}^m |\nabla^k \psi_j(\vec{x})|^{q^*}\right)^{\frac{1}{q^*}} \|\vec{r}\|_{\ell q} = \left\|(\nabla^k \psi_j(\vec{x}))_j\right\|_{\ell q^*} \|\vec{r}\|_{\ell q} .$$

For the second inequality, if $q \geq 2$ and $\sum_{j=1}^m r_j^2 \leq 1$, then $|r_j| \leq 1$ for all j and $\|\vec{r}\|_{\ell q}^q \leq \|\vec{r}\|_{\ell^2}^2 \leq 1$. If $q < 2$ and $\|\vec{r}\|_{\ell^2} \leq 1$, then we again use Hölder's inequality:

$$\sum_{j=1}^m r_j^q \leq \left(\sum_{j=1}^m 1^{Q^*}\right)^{\frac{1}{Q^*}} \left(\sum_{j=1}^m r_j^{qQ}\right)^{\frac{1}{Q}} \leq m^{\frac{2-q}{2}}$$

for $Q = \frac{2}{q}$. □

The examples in Section 7 require explicit computations of the expressions in Lemmas 8 and 9. These computations are provided in the appendix, Theorem 8.

7 Numerical examples

We present four numerical examples. The first two are in 1D to demonstrate the performance of different variants of Algorithm 1, both with and without adaptivity. In particular, we explore sparse Gaussian deconvolution and sparse signal recovery from Fourier data. We compare with the *continuous basis pursuit* (CBP) discretisation [17, 16] which is also designed to achieve super-resolution accuracy within a convex framework. More details of this method will be provided in Section 7.1.

The next example is 2D reconstruction from Radon or X-ray data with wavelet-sparsity and a robust data fidelity. As the forward operator is not sufficiently smooth, we must optimise in $\ell^1(\mathbb{R})$, which naturally leads to the choice of a wavelet basis.

Finally, we process a dataset which represents a realistic application in biological microscopy, referred to as STORM microscopy. In essence, the task is to perform 2D Gaussian de-blurring/super-resolution and denoising to find the location of sparse spikes of signal.

In this section, the main aim is to minimise $\mathrm{E}_0(u_n) = \mathrm{E}(u_n) - \mathrm{E}(u^*)$, and so this will be our main metric for the success of an algorithm, referred to as the "continuous gap". Lemma 6 only provides guarantees on the values of $\min_{n\leq N} \mathrm{E}_0(u_n)$ so it is this monotone estimate which is plotted. As $\mathrm{E}(u^*)$ is not known exactly, we always use the estimate $\min_{n\leq N} \mathrm{E}_0(u_n) \approx \min_{n\leq N} \mathrm{E}(u_n) + \min_{n'\leq n} \mathrm{E}^\dagger(\gamma_0 \vec{\varphi}_{n'})$. Another quantity of interest is minimisation of the discrete energy $\min_{n\leq N} \mathrm{E}(u_n) + \min_{n'\leq n} \mathrm{E}^\dagger(\gamma \vec{\varphi}_{n'})$ which will be referred to as the "discrete gap". Note that for the adaptive schemes the discrete gap may not be monotonic as the discrete dual problem changes with N.

The code to reproduce these examples can be found online[1].

7.1 1D continuous LASSO

In this example we choose $\mathbb{U} = \mathscr{M}([0,1])$, $\mathbb{H} = L^2([0,1])$, $\mathrm{f}(\cdot) = \frac{1}{2}\|\cdot\|_{\ell^2}^2$ and $\mathsf{A}\colon \mathbb{U} \to \mathbb{R}^{30}$ with either random Fourier kernels:

$$(\mathsf{A}u)_j = \int_0^1 \cos(a_j x)\,\mathrm{d}u(x), \qquad a_j \sim \mathrm{Uniform}[-100,100],\ j = 1,2,\ldots,30,\ \mu = 0.02, \tag{60}$$

or Gaussian kernels on a regular grid:

$$(\mathsf{A}u)_j = (2\pi\sigma^2)^{-\frac{1}{2}} \int_0^1 \exp\left(-\frac{(x-(j-1)\Delta)^2}{2\sigma^2}\right)\mathrm{d}u(x), \quad \sigma = 0.12,\ \Delta = \tfrac{1}{29},\ j = 1,2,\ldots,30,\ \mu = 0.06. \tag{61}$$

Several variants of FISTA are compared for these examples but the key alternative shown here is the CBP discretisation. For this choice of f, we call (18) the continuous LASSO problem, for which there are many numerical methods (c.f. [7,12,5,9]) however, most require the solution of a non-convex problem. We have focused on CBP because it approximates u^* through a convex discrete optimisation problem which is asymptotically exact in

[1] https://github.com/robtovey/2020SpatiallyAdaptiveFISTA

the limit $h \to 0$. It can also be optimised with FISTA which allows for direct comparison with the uniform and adaptive mesh approaches. The idea is that for a fixed mesh, the kernels of A are expanded to first order on each pixel and a particular first order basis is also chosen [17, 16]. If u^* has only one Dirac spike in each pixel, then the zeroth order information should correspond to the mass of the spike, and additional first order information should determine the location.

As shown in Section 6, in 1D we have $a_{\mathrm{U}} = a_{\mathrm{E}} = 2$. The estimates given in (25) and (26) in dimension $d = 1$ predict that the adaptive energy will decay at a rate of $\mathrm{E}(u_n) - \mathrm{E}(u^*) \lesssim \frac{1}{n}$ so long as the pixel size also decreases at a rate of $h \sim \frac{1}{n}$. To achieve these rates, we implement a refinement criterion from Lemma 6 with guarantee of $\mathrm{E}(u_{n_k-1}) - \mathrm{E}(u^*) \lesssim 2^{-k}$ using the estimates made in Section 6.3. We choose subspaces $\mathbb{U}^n$ to approximately enforce

$$\mathrm{E}(u_n) + \mathrm{E}^\dagger(\gamma_0 \vec{\varphi}_n) \leq 2(\mathrm{E}(u_n) + \mathrm{E}^\dagger(\gamma \vec{\varphi}_n)), \tag{62}$$

i.e. the continuous gap is bounded by twice the discrete gap. In particular, note that for $\gamma_0 \approx \gamma$,

$$\mathrm{E}^\dagger(\gamma_0 \vec{\varphi}_n) = \tfrac{1}{2}\|\gamma_0 \vec{\varphi}_n\|^2 + \gamma_0 \eta \bullet \vec{\varphi}_n = \frac{\gamma_0}{\gamma}\left(\frac{\gamma_0}{\gamma}\tfrac{1}{2}\|\gamma \vec{\varphi}_n\|^2 + \gamma \eta \bullet \vec{\varphi}_n\right) \approx \frac{\gamma_0}{\gamma}\mathrm{E}^\dagger(\gamma \vec{\varphi}_n). \tag{63}$$

Converting this into a spatial refinement criteria, recall

$$\frac{\gamma_0}{\gamma} \approx \frac{\|\!|\mathsf{A}^* \vec{\varphi}_n|\!\|_*}{\|\!|\Pi_n \mathsf{A}^* \vec{\varphi}_n|\!\|_*} = \frac{\max_{\omega \in \mathbb{M}^n} \|\mathsf{A}^* \vec{\varphi}_n\|_{L^\infty(\omega)}}{\max_{\omega \in \mathbb{M}^n} |\Pi_n \mathsf{A}^* \vec{\varphi}_n(\omega)|} \approx \max_{\omega \in \mathbb{M}^n} \frac{\|\mathsf{A}^* \vec{\varphi}_n\|_{L^\infty(\omega)}}{|\Pi_n \mathsf{A}^* \vec{\varphi}_n(\omega)|} \tag{64}$$

is the maximum ratio of second vs. zeroth order Taylor approximations of $\mathsf{A}^* \vec{\varphi}_n$ on pixel ω. This was found to be an efficient method of selecting pixels for refinement using quantities which had already been computed. Note briefly that this greedy strategy directly targets uncertainty, refinements also happen outside of the support of u_n to guarantee that this is representative of u^*. Such refinement is necessary to avoid discrete minimisers of E which are not global minimisers.

Comparison of discretisation methods In Fig. 1 we compare the three core approaches: fixed uniform discretisation, adaptive discretisation, and CBP. In particular, we wish to observe their convergence properties as the number of pixels is allowed to grow. In each case we use a FISTA stepsize of $t_n = \frac{n+19}{20}$. The adaptive discretisation is started with one pixel and limited to 128, 256, or 512 pixels while the fixed and CBP discretisations have uniform discretisations with the maximum number of pixels. The main observations are:

- The adaptive scheme is much more efficient, in both examples the adaptive scheme with 128 pixels is at least as good as both fixed discretisations with 512 pixels. In fact, only a maximum of 214 pixels were needed by the adaptive method in either example.
- With Fourier kernels the uniform piecewise constant discretisation is more efficient than CBP but in the Gaussian case this is reversed. This suggests that the performance of CBP depends on the smoothness of A.
- The discrete gaps for non-adaptive optimisation behave as is common for FISTA, initial convergence is polynomial until a locally linear regime activates [37]. CBP is always slower to converge than the piecewise constant discretisation.
- The adaptive refinement criterion succeeds in keeping the continuous/discrete gaps close for all n, i.e. (62).

It is not completely fair to judge CBP with the continuous gap because, although it generates a continuous representation, this continuous representation is not necessarily consistent with the discrete gap being optimised, unlike when discretised with finite element methods. On the other hand, this is still the intended interpretation of the algorithm and we have no more appropriate metric for success in this case.

Comparison of FISTA variants Fig. 2 compares many methods with either fixed or adaptive discretisations. Each adaptive scheme is allowed up to 1024 pixels and each uniform discretisation uses exactly 1024. An example of each reconstruction method is shown in Fig. 3. The adaptive method better identifies the support of u^* and clearly localises pixels on that support. The reconstruction with uniform grid fails to provably identify the support of u^*, despite having found a qualitatively accurate discrete minimiser. The "Greedy FISTA" implementation was proposed by in [24] and we include the adaptive variant despite a lack of convergence proof. The remaining FISTA algorithms use a FISTA time step of $t_n = \frac{n+a-1}{a}$ for the given value of a, as proposed in [10]. In this example CBP used the greedy FISTA implementation which gave faster observed convergence. Fig. 2 compares the discrete gaps because it is the accurate metric for fixed discretisations, and for the adaptive discretisation it should also be an accurate predictor of the continuous gap. The main observations are:

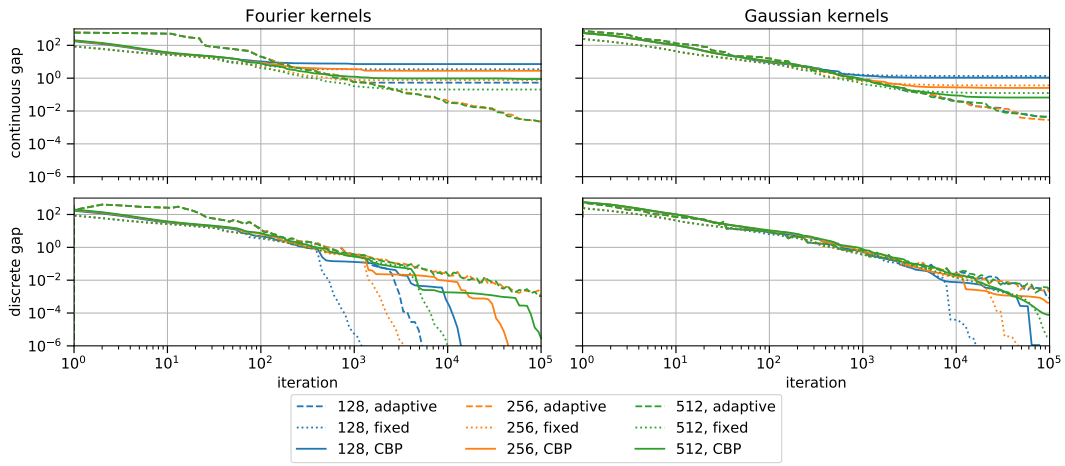

Fig. 1 Rates of continuous/discrete gap convergence for different LASSO algorithms with 128, 256, or 512 pixels. The "adaptive" method uses the proposed algorithm. Both "fixed" and "CBP" use standard FISTA with a uniform discretisation.

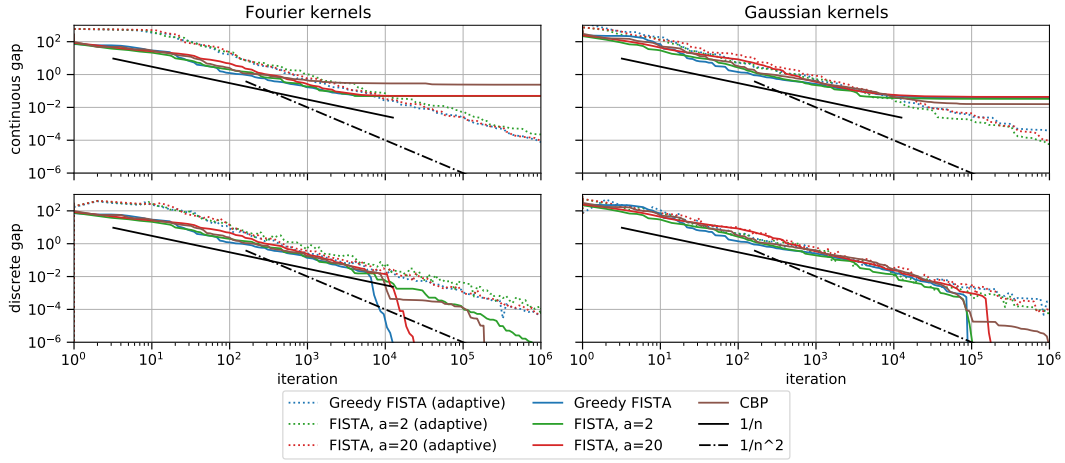

Fig. 2 Convergence plots for solving 1D problems with different algorithms. "Adaptive" methods use Algorithm 1 with fewer than 1024 pixels and the remaining methods use a uniform discretisation of 1024 pixels.

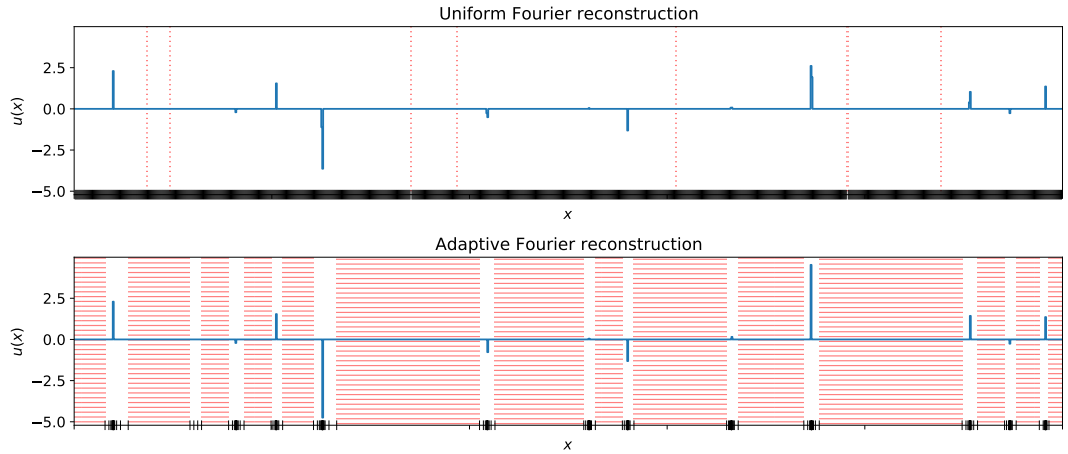

Fig. 3 Example reconstruction from the algorithms considered in Fig. 2. Pixel boundaries are indicated on the x-axis and the filtering method of Section 6.4 allows us to exclude the red shaded regions from $\mathrm{supp}(u^*)$. Values on the y-axis are normalised to units of mass, i.e. a Dirac mass would have height 1.

– Each algorithm displays very similar convergence properties. The main difference is that the reconstructions with fixed discretisations accelerate after 10^4-10^5 iterations.
– During the initial "slow" phase, adaptive and fixed discretisations appear to achieve very similar (discrete) convergence rates. The coarse-to-fine adaptivity is not slower than fixed discretisations in this regime.
– Lemma 6 accurately predicts the $\frac{1}{n}$ rate of the adaptive methods, mirrored in the fixed discretisations. This suggests that high-resolution discretisations are also initially limited by this $\frac{1}{n}$ rate before entering the asymptotic regime, consistent with (14).
– The fastest FISTA stepsize choice is consistently the greedy variant, although $a = 20$ is very comparable.
– While each adaptive algorithm is allowed to use up to 1024 pixels, in Fig. 2 the most used was 235.

Comparison of fixed and adaptive discretisation Motivated by the findings in Fig. 2, we now look more closely at the performance of the $a = 20$ and the greedy FISTA schemes. We have convergence results for the former, but the latter typically performs the best for non-adaptive optimisation and is never worse than $a = 20$ in the adaptive setting. The question is whether it is faster/more efficient to use the proposed adaptive scheme, or to use a classical scheme at sufficiently high uniform resolution. The fixed discretisations use 1024 pixels (i.e. constant pixel size of 2^{-10} in Fig. 4) and the adaptive discretisation starts with two pixels with an upper limit of 1024. As expected, the fixed discretisation starts with a smaller continuous gap before plateauing to a sub-optimal gap around $\mathrm{E} = \mathrm{E}(u^*) + 0.1$.

Fig. 4 shows convergence of pixel size and continuous gap with respect to number of iterations. Fig. 5 shows the more practical attributes of continuous gap and number of pixels against execution time. We see that the adaptive discretisation is consistently capable of computing lower energies with fewer pixels and in less time than the uniform discretisation. The convergence behaviour is very consistent with respect to number of iterations.

Suppose that the numerical aim is to find a function u_n with $\mathrm{E}(u_n) - \mathrm{E}(u^*) \leq 0.1$, all methods would converge after $O(10^3)$ iterations, demonstrating some equivalence between the two FISTA algorithms. For $n \in [10^3, 10^4]$, in both problems, the adaptive schemes coincide with the fixed schemes in both energy and minimum pixel size. On the other hand, we also see that the adaptive scheme achieves this energy in almost an order of magnitude less time and fewer pixels.

7.2 2D robust sparse wavelet reconstruction

In this example we consider A to be a 2D Radon transform. In particular, the rows of A correspond to integrals over the sets $\mathbb{X}_i^I$ where

$$\mathbb{X}_i^I = \left\{ \vec{x} \in [-\tfrac{1}{2}, \tfrac{1}{2}]^2 \text{ s.t. } \vec{x} \bullet \begin{pmatrix} \cos\theta_I \\ \sin\theta_I \end{pmatrix} \in \left[-\tfrac{1}{2} + \tfrac{i-1}{100}, -\tfrac{1}{2} + \tfrac{i}{100} \right) \right\}, \quad \theta_I = \frac{180^\circ}{51} I \tag{65}$$

for $i = \in [100]$, $I \in [50]$. This is not exactly in the form analysed by Theorem 8, only the sets $\{\mathbb{X}_i^I \text{ s.t. } i \in [100]\}$ for each I are disjoint, therefore we apply Theorem 8 block-wise to estimate

$$\|\mathsf{A}\|_{L^2 \to \ell^2} \leq \sqrt{\sum_{I \in [50]} \max_{i \in [100]} |\mathbb{X}_i^I|} = \sqrt{\sum_{I \in [50]} \max_{i \in [100]} \int_{\mathbb{X}_i^I} 1 \, \mathrm{d}\vec{x}} = \sqrt{\sum_{I \in [50]} \max_{i \in [100]} (\mathsf{A}\mathbb{1})_{i,I}} \,. \tag{66}$$

A is not smooth, therefore we can't bound $|\mathsf{A}^*|_{C^k}$ for $k > 0$, and so we must look to minimise over ℓ^1 rather than L^1. The natural choice is to promote sparsity in a wavelet basis which can be rearranged into the form of (18):

$$\min_{u \in \mathbb{U}} \mathrm{f}(\mathsf{A}u - \eta) + \mu \left\| \mathsf{W}^{-1} u \right\|_{\ell^1} = \min_{\widehat{u} \in \ell^1(\mathbb{R})} \mathrm{f}(\mathsf{A}\mathsf{W}\widehat{u} - \eta) + \mu \|\widehat{u}\|_{\ell^1} \,. \tag{67}$$

The minimisers are related by $u^* = \mathsf{W}\widehat{u}^*$ and, for wavelet bases, W is orthonormal so $\|\mathsf{A}\mathsf{W}\|_{\ell^2 \to \ell^2} = \|\mathsf{A}\|_{L^2 \to \ell^2}$. In this example we consider the smoothed robust fidelity [30]

$$\mathrm{f}(\vec{\varphi}) = \sum_{i=1}^{m} \begin{cases} 10^{-4} |\varphi_i| & |\varphi_i| \geq 10^{-4} \\ \frac{1}{2} |\varphi_i|^2 + \frac{1}{2} 10^{-8} & \text{else} \end{cases} \approx 10^{-4} \|\vec{\varphi}\|_{\ell^1} \,. \tag{68}$$

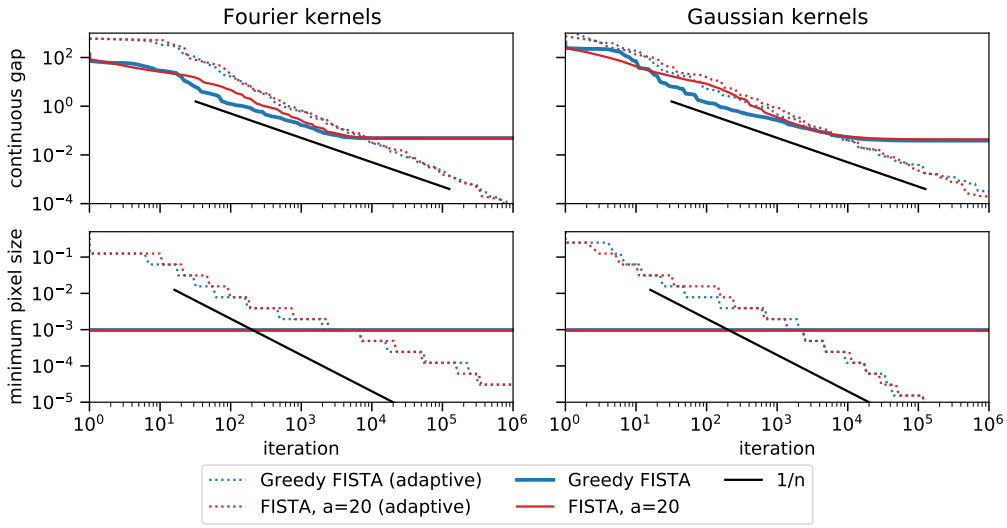

Fig. 4 Continuous convergence of adaptive (coarse-to-fine pixel size) compared with uniform discretisation (constant pixel size) with respect to number of iterations.

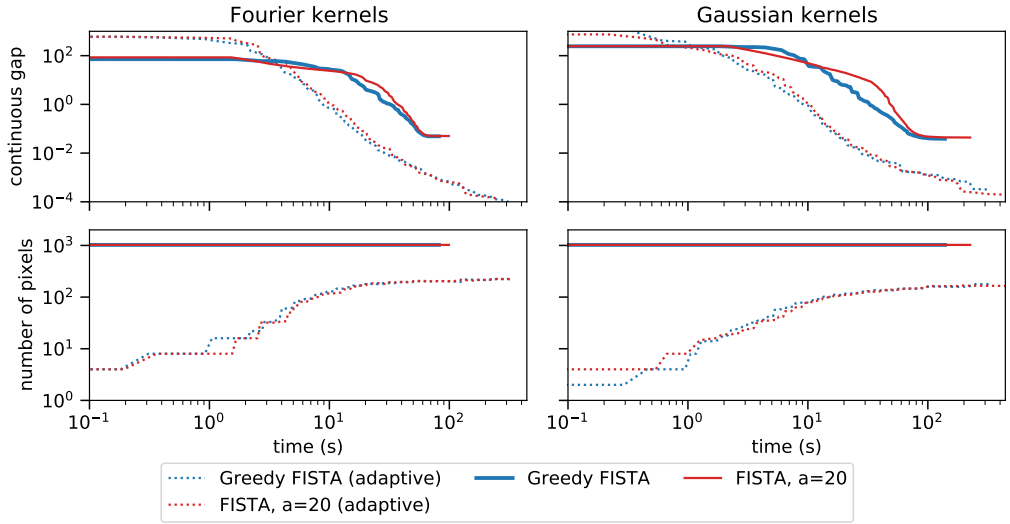

Fig. 5 Continuous convergence of adaptive compared with uniform discretisation with respect to wall-clock time and total number of pixels (memory requirement).

$$J_n = \{(0,0),(0,1),(0,2),(1,2),(1,1)\}$$
$$\operatorname{leaf}(J_n) = \{(0,2),(1,2),(1,1)\}$$
$$\mathbb{M}^n = \{[0,\tfrac{1}{4}),[\tfrac{1}{4},\tfrac{1}{2}),[\tfrac{1}{2},1)\}$$

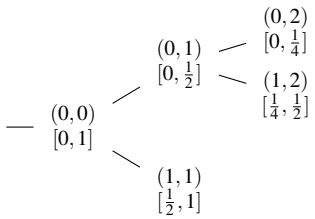

Fig. 6 Example tree representation of 1D wavelets. Left: nodes, leaves, and mesh of discretisation. Right: arrangement into a tree with index (j,k) and corresponding support of wavelet $w_{j,k}$ underneath.

From Section 6.3 we know that to track convergence and perform adaptive refinement, it is sufficient to accurately bound $|[\mathsf{W}^\top \mathsf{A}^* \vec{\varphi}_n]_j|$ for all $j \notin J_n$. If W is a wavelet transformation then its columns, $w_j \in L^2$, are simply the wavelets themselves and we can use the bound

$$|\langle w_j, \mathsf{A}^* \vec{\varphi}_n \rangle| = \left| \left\langle w_j, \mathbb{1}_{\operatorname{supp}(w_j)} \mathsf{A}^* \vec{\varphi}_n \right\rangle \right| \leq \left\| \mathbb{1}_{\operatorname{supp}(w_j)} \mathsf{A}^* \vec{\varphi}_n \right\|_{L^2} \leq \| \mathbb{1}_{\mathbb{X}} \mathsf{A}^* \vec{\varphi}_n \|_{L^2} \tag{69}$$

for all $\mathbb{X} \supset \operatorname{supp}(w_j)$. In the case of the Radon transform, we can compute the left-hand side explicitly for the finitely many $j \in J_n$, but we wish to use the right-hand side in a structured way to avoid computing the infinitely many $j \notin J_n$. To do this, we will take a geometrical perspective on the construction of wavelets to view them in a tree format.

Tree structure of wavelets Finite elements are constructed with a mesh which provided a useful tool for adaptive refinement in Section 6.3.3. For wavelets, we will associate a tree with every discretisation and the leaves of the tree correspond to a mesh. This perspective comes from the multi-resolution interpretation of wavelets. An example is seen in Fig. 6 for 1D Haar wavelets, $w_{j,k}(x) = \sqrt{2}^k \psi(2^k x - j)$ where $\psi = \mathbb{1}_{[0,1)} - \mathbb{1}_{[-1,0)}$.

In higher dimensions, the only two things which change are the number of children (2^d for non-leaves) and at each node you store the coefficients of $2^d - 1$ wavelets. The support on each node is still a disjoint partition of unity consisting of regular cubes of side length 2^{-k} at level k. The only change in our own implementation is to translate the support to $[-\frac{1}{2}, \frac{1}{2}]^2$. We briefly remark that the tree structuring of wavelets is not novel and appears more frequently in the Bayesian inverse problems literature [8, 22].

Continuous gradient estimate In Section 7.1 we used the continuous gap as a measure for convergence, for wavelets we will use the continuous subdifferential. With the tree structure we can easily adapt the results of Section 6.3 to estimate subdifferentials (or function gaps). In particular,

$$\begin{aligned} \| \partial \mathrm{E}(u_n) \|_* &= \max \left(\| \partial_n \mathrm{E}(u_n) \|_*, \max_{j \notin J_n} |\langle w_j, \mathsf{A}^* \vec{\varphi}_n \rangle| - \mu \right) && (70) \\ &\leq \max \left(\| \partial_n \mathrm{E}(u_n) \|_*, \max_{j \in \operatorname{leaf}(J_n)} \left\| \mathbb{1}_{\operatorname{supp}(w_j)} \mathsf{A}^* \vec{\varphi}_n \right\|_{L^2} - \mu \right). && (71) \end{aligned}$$

Numerical results We consider two phantoms where the ground-truth is either a binary disc or the Shepp-Logan phantom. Both examples are corrupted with 2 % Laplace distributed noise. This is visualised in Fig. 7. All optimisations shown are spatially adaptive using Haar wavelets and initialised with $\mathbb{U}^0 = \{x \mapsto c \text{ s.t. } c \in \mathbb{R}\}$. The gradient metric shown throughout is the ℓ^∞ norm. Motivated by (71), the spatial adaptivity is chosen to refine nodes $j \in \operatorname{leaf}(J_n)$ to ensure that

$$\left\| \mathbb{1}_{\operatorname{supp}(w_j)} \mathsf{A}^* \vec{\varphi}_n \right\|_{L^2} - \mu \leq 10 \| \partial_n \mathrm{E}(u_n) \|_*$$

for all j and n (i.e. so that the continuous gradient is less than 10 times the discrete gradient). We do not expect wavelet regularisation to have state-of-the-art performance in the examples of Fig. 7. What they demonstrate is the preference Haar wavelets have to align large discontinuities with a coarse grid, even when the discretisation is allowed to be as fine as necessary. There is an average of $2 \cdot 10^6$ wavelet coefficients in each discretised reconstruction, although the higher frequencies have much smaller intensities. In limited data scenarios, wavelet regularisation automatically selects a local "resolution" which reflects the quality of data. Particularly in the Shepp-Logan reconstruction, we see that the outer ring is detected with a finer precision than the dark interior ellipses.

The first numerical results shown in Fig. 8 compare the same adaptive FISTA variants as shown in Fig. 2. In these examples we see that the greedy FISTA and the $a = 20$ algorithms achieve almost linear convergence while $a = 2$ is significantly slower. Interestingly, in both examples the $a = 20$ variant uses half as many wavelets as the Greedy variant, and therefore converges slightly faster in time.

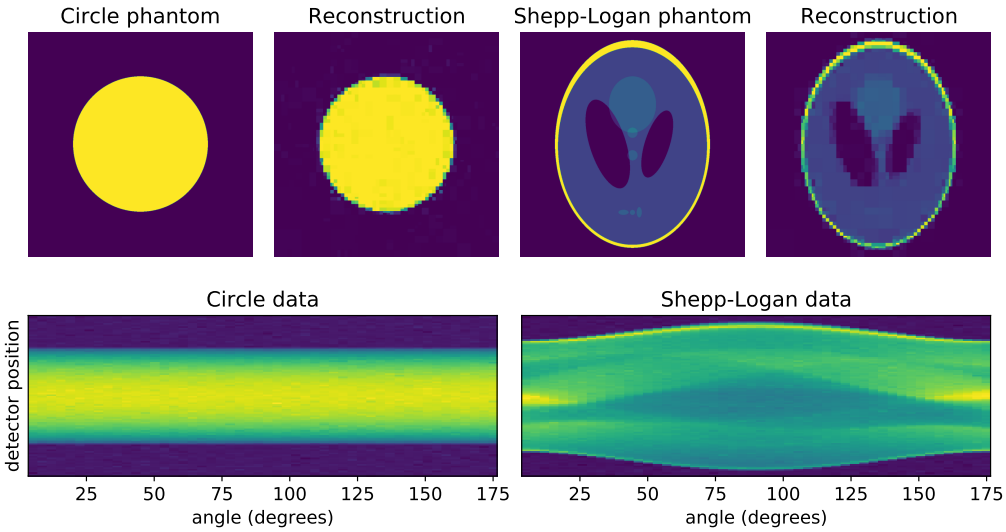

Fig. 7 Phantoms, data and reconstructions for wavelet-sparse tomography optimisation. Both examples are corrupted with 2 % Laplace distributed noise.

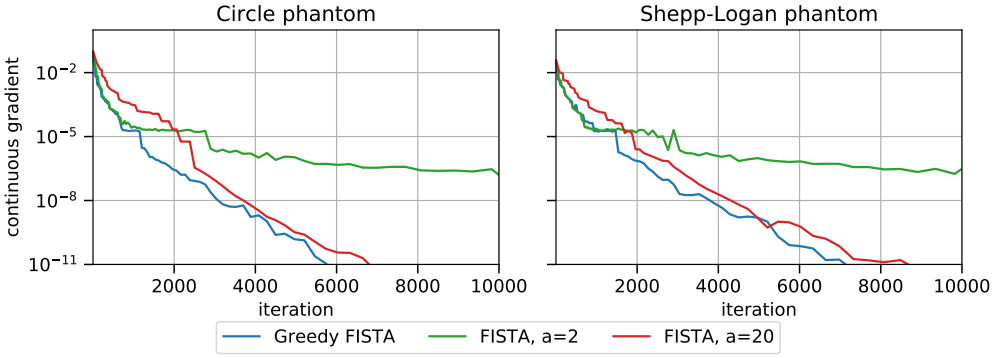

Fig. 8 Convergence of different implementations of Algorithm 1 with an unlimited number of pixels for sparse wavelet optimisation.

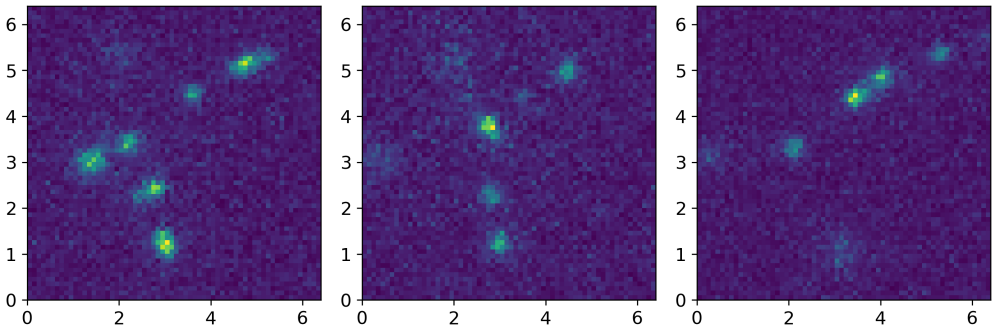

Fig. 9 Example images from STORM dataset.

7.3 2D continuous LASSO

Our final application is a super-resolution/de-blurring inverse problem from biological microscopy. In mathematical terms, the observed data is a large number of sparse images which are corrupted by blurring and a large amount of noise, examples are seen in Fig. 9. The task is to compute the centres of the spikes of signal in each image and then re-combine into a single super-resolved image, as in Fig. 11. This technique is referred to as *Single Molecule Localisation Microscopy* (SMLM), of which we consider the specific example of *Stochastic Optical Reconstruction Microscopy* (STORM). Readers are directed to the references [31,32,33] for further details. The LASSO formulation ($\mathrm{f}(\cdot) = \frac{1}{2}\|\cdot\|^2_{\ell^2}$) has previously been shown to be effective in the context of STORM [20,13].

Here we use a simulated dataset provided as part of the 2016 SMLM challenge[2] for benchmarking software in this application. The corresponding LASSO formulation is

$$(\mathrm{A}u)_i = (2\pi\sigma^2)^{-1} \int_{[0,6.4]^2} \exp\left(-\frac{1}{2\sigma^2}\left|\vec{x} - \Delta\left(i_1 + \tfrac{1}{2}\ i_2 + \tfrac{1}{2}\right)^\top\right|^2\right) u(\vec{x})\,\mathrm{d}\vec{x}, \qquad \sigma = 0.2,\ \Delta = 0.1 \tag{72}$$

for $i_1, i_2 = 1, 2, \ldots, 64$, $\mathbb{U} = \mathscr{M}([0,6.4]^2)$ and $\mathbb{H} = L^2([0,6.4]^2)$ with lengths in µm. 3020 frames are provided, examples of which are shown in Fig. 9. To process this dataset, image intensities were normalised to $[0,1]$ then a constant was subtracted to approximate 0-mean noise. The greedy FISTA algorithm was used for optimisation with $\mu = 0.15$, 10^3 iterations, and a maximum of 10^5 pixels per image.

Finally, all the reconstructions were summed and the result shown in Fig. 11. The adaptive scheme used fewer than 10^4 pixels per frame, a fixed discretisation with equivalent resolution of 1.3 nm would have required more than $3 \cdot 10^6$ per frame. LASSO is compared with ThunderSTORM [27], a popular ImageJ plugin [34] which finds the location of signal using Fourier filtering. The performance of ThunderSTORM was rated very highly in the initial SMLM challenge [31]. Both methods compared here demonstrate the key structures of the reconstruction, however, both are sensitive to tuning parameters. In this examples, LASSO has possibly recovered too little signal and ThunderSTORM contains spurious signal.

Fig. 10 shows various convergence metrics for the adaptive reconstructions. The magenta line in the first panel shows that the continuous gap converges slightly faster than the $n^{-2/3}$ predicted by (26) in dimension $d = 2$. In this example we also implement the suggestion of Section 6.4 to remove pixels outside of the support of u^*. From (56), any pixel $\omega \in \mathbb{M}^n$ satisfying

$$\gamma_0 \|\Pi_n \mathrm{A}^* \vec{\varphi}_n\|_{L^\infty(\omega)} \leq (1 - \mathrm{threshold}_n)\mu \tag{73}$$

guarantees that $\omega \cap \mathrm{supp}(u^*) = \emptyset$. This threshold is plotted in red in the first panel of Fig. 10. Once the value becomes less than 1, we can start reducing the number of pixels instead of continual refinement. We see that the resolution decreases steadily (second panel), but the total number of pixels (final panel) stops increasing after around 30 iterations.

8 Conclusions and outlook

In this work we have proposed a new adaptive variant of FISTA and provided convergence analysis. This algorithm allows FISTA to be applied outside of the classical Hilbert space setting, still with a guaranteed rate of convergence. We have presented several numerical examples where convergence with the refining discretisation is at least as fast as a uniform discretisation, although more efficient with regards to both memory and computation time.

In 1D we see good agreement with the theoretical rate. This rate also seems to be a good predictor for all variants of FISTA tested, although this is yet to be proven. Even the classical methods with a fixed discretisation are initially limited to the slower adaptive rate for small n.

The results in 2D are similar, all tested FISTA methods converge at least at the guaranteed rate. The wavelet example was most impressive, achieving nearly linear convergence in energy. This is similar to the behaviour for classical FISTA although it is also yet to be formally proven.

An interesting observation over all of the adaptive LASSO examples is that the standard oscillatory behaviour of FISTA has not occurred. With the monotone gaps plotted, oscillatory convergence should correspond to a piecewise constant descending gap. Either this behaviour only emerges for larger n, or the adaptivity provides a dampening effect for this oscillation.

[2] http://bigwww.epfl.ch/smlm/challenge2016/datasets/MT4.N2.HD/Data/data.html

Moving forward, it would be interesting to see how far the analysis extends to other optimisation algorithms. Other variants of FISTA, such as the "greedy" implementation used here or the traditional Forward-Backward algorithm, should also be receptive to the analysis performed here. Furthermore, it would also interesting to attempt to replicate this refinement argument to extend the primal-dual algorithm [11] or the Douglas-Rachford algorithm [14].

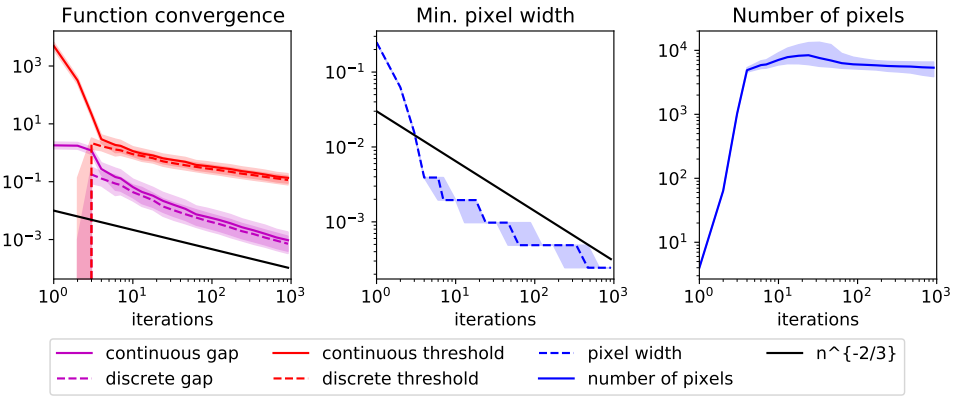

Fig. 10 Convergence of adaptive FISTA for STORM dataset. Lines indicate the median value over 3020 STORM frames. Shaded regions indicate the 25 % to 75 % interquartile range. Pixel width is scaled $[0, 1]$ rather than $[0, 6.4\,\mu\text{m}]$.

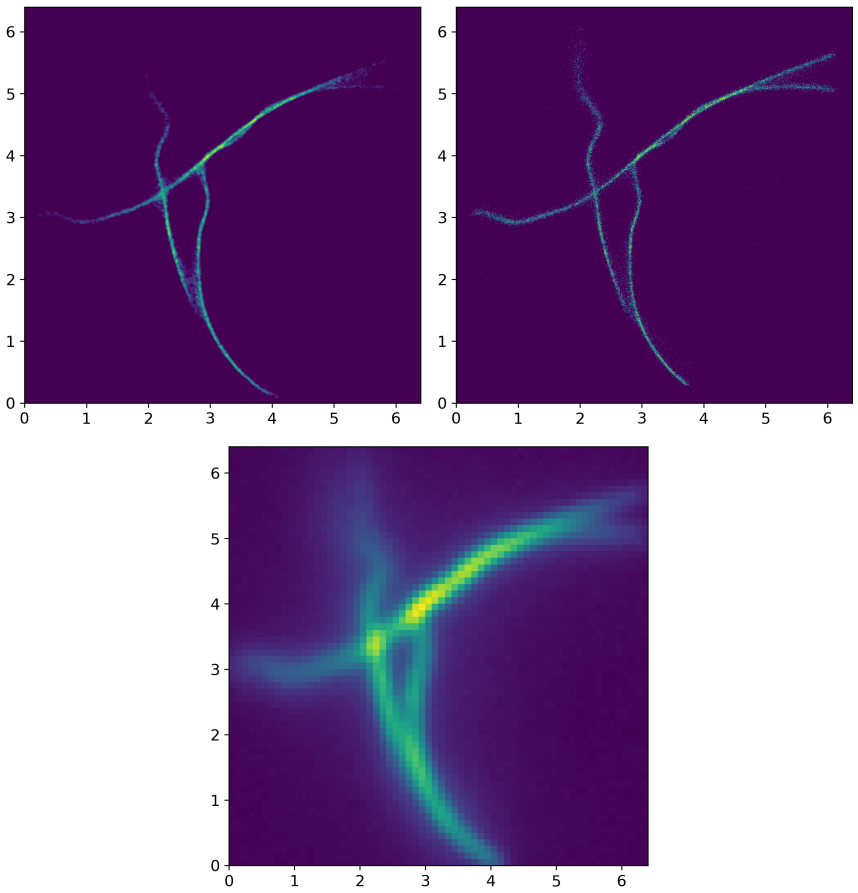

Fig. 11 Processed results of the STORM dataset. Top left: LASSO optimisation with Algorithm 1. Top right: Comparison with ThunderSTORM plugin. Bottom: Average data, no super-resolution or de-blurring.

Acknowledgements R.T. acknowledges funding from EPSRC grant EP/L016516/1 for the Cambridge Centre for Analysis, and the ANR CIPRESSI project grant ANR-19-CE48-0017-01 of the French Agence Nationale de la Recherche. Most of this work was done while A.C. was still in CMAP, CNRS and Ecole Polytechnique, Institut Polytechnique de Paris, Palaiseau, France. Both authors would like to thank the anonymous reviewers who put in so much effort to improving this work.

Data availability statement The synthetic STORM dataset was provided as part of the 2016 SMLM challenge, http://bigwww.epfl.ch/smlm/challenge2016/datasets/MT4.N2.HD/Data/data.html. The remaining examples used in this work can be generated with the supplementary code, https://github.com/robtovey/2020SpatiallyAdaptiveFISTA.

Conflict of interest The authors have no conflicts of interest to declare which are relevant to the content of this article.

References

1. Alamo, T., Limon, D., Krupa, P.: Restart fista with global linear convergence. In: 2019 18th European Control Conference (ECC), pp. 1969–1974. IEEE (2019)
2. Aujol, J.F., Dossal, C.: Stability of over-relaxations for the forward-backward algorithm, application to fista. SIAM Journal on Optimization **25**(4), 2408–2433 (2015)
3. Beck, A., Teboulle, M.: A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM Journal on Imaging Sciences **2**(1), 183–202 (2009)
4. Bonnefoy, A., Emiya, V., Ralaivola, L., Gribonval, R.: Dynamic screening: Accelerating first-order algorithms for the lasso and group-lasso. IEEE Transactions on Signal Processing **63**(19), 5121–5132 (2015)
5. Boyd, N., Schiebinger, G., Recht, B.: The alternating descent conditional gradient method for sparse inverse problems. SIAM Journal on Optimization **27**(2), 616–639 (2017)
6. Boyer, C., Chambolle, A., Castro, Y.D., Duval, V., De Gournay, F., Weiss, P.: On representer theorems and convex regularization. SIAM Journal on Optimization **29**(2), 1260–1281 (2019)
7. Bredies, K., Pikkarainen, H.K.: Inverse problems in spaces of measures. ESAIM: Control, Optimisation and Calculus of Variations **19**(1), 190–218 (2013)
8. Castillo, I., Rockova, V.: Multiscale analysis of bayesian cart. University of Chicago, Becker Friedman Institute for Economics Working Paper (2019-127) (2019)
9. Catala, P., Duval, V., Peyré, G.: A low-rank approach to off-the-grid sparse superresolution. SIAM Journal on Imaging Sciences **12**(3), 1464–1500 (2019)
10. Chambolle, A., Dossal, C.: On the convergence of the iterates of the "fast iterative shrinkage/thresholding algorithm". Journal of Optimization Theory and Applications **166**(3), 968–982 (2015)
11. Chambolle, A., Pock, T.: A First-Order Primal-Dual Algorithm for Convex Problems with Applications to Imaging. Journal of Mathematical Imaging and Vision **40**(1), 120–145 (2011). DOI 10.1007/s10851-010-0251-1
12. De Castro, Y., Gamboa, F., Henrion, D., Lasserre, J.B.: Exact solutions to super resolution on semi-algebraic domains in higher dimensions. IEEE Transactions on Information Theory **63**(1), 621–630 (2016)
13. Denoyelle, Q., Duval, V., Peyré, G., Soubies, E.: The sliding frank–wolfe algorithm and its application to super-resolution microscopy. Inverse Problems **36**(1), 014001 (2019)
14. Douglas, J., Rachford, H.H.: On the numerical solution of heat conduction problems in two and three space variables. Transactions of the American Mathematical Society **82**(2), 421–439 (1956)
15. Duval, V., Peyré, G.: Sparse spikes super-resolution on thin grids i: the lasso. Inverse Problems **33**(5), 055008 (2017)
16. Duval, V., Peyré, G.: Sparse spikes super-resolution on thin grids ii: the continuous basis pursuit. Inverse Problems **33**(9), 095008 (2017)
17. Ekanadham, C., Tranchina, D., Simoncelli, E.P.: Recovery of sparse translation-invariant signals with continuous basis pursuit. IEEE transactions on signal processing **59**(10), 4735–4744 (2011)
18. El Ghaoui, L., Viallon, V., Rabbani, T.: Safe feature elimination in sparse supervised learning. Tech. Rep. UCB/EECS-2010–126, EECS Department, University of California, Berkeley (2010)
19. Hiriart-Urruty, J.B., Lemaréchal, C.: Convex Analysis and Minimization Algorithms II: Advanced Theory and Bundle Methods, vol. 305. Springer-Verlag, Berlin, Heidelberg (1993)
20. Huang, J., Sun, M., Ma, J., Chi, Y.: Super-resolution image reconstruction for high-density three-dimensional single-molecule microscopy. IEEE Transactions on Computational Imaging **3**(4), 763–773 (2017)
21. Jiang, K., Sun, D., Toh, K.C.: An inexact accelerated proximal gradient method for large scale linearly constrained convex sdp. SIAM Journal on Optimization **22**(3), 1042–1064 (2012)
22. Kekkonen, H., Lassas, M., Saksman, E., Siltanen, S.: Random tree besov priors–towards fractal imaging. arXiv preprint arXiv:2103.00574 (2021)
23. Liang, J., Fadili, J., Peyré, G.: Activity identification and local linear convergence of forward–backward-type methods. SIAM Journal on Optimization **27**(1), 408–437 (2017)
24. Liang, J., Schönlieb, C.B.: Improving fista: Faster, smarter and greedier. arXiv preprint arXiv:1811.01430 (2018)
25. Ndiaye, E., Fercoq, O., Gramfort, A., Salmon, J.: Gap safe screening rules for sparsity enforcing penalties. The Journal of Machine Learning Research **18**(1), 4671–4703 (2017)
26. Nesterov, Y.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers Boston, Dordrecht, London (2004)
27. Ovesnỳ, M., Křížek, P., Borkovec, J., Švindrych, Z., Hagen, G.M.: Thunderstorm: a comprehensive imagej plug-in for palm and storm data analysis and super-resolution imaging. Bioinformatics **30**(16), 2389–2390 (2014)
28. Parpas, P.: A multilevel proximal gradient algorithm for a class of composite optimization problems. SIAM Journal on Scientific Computing **39**(5), S681–S701 (2017)
29. Poon, C., Keriven, N., Peyré, G.: The geometry of off-the-grid compressed sensing. arXiv preprint arXiv:1802.08464 (2018)

30. Rosset, S., Zhu, J.: Piecewise linear regularized solution paths. The Annals of Statistics pp. 1012–1030 (2007)
31. Sage, D., Kirshner, H., Pengo, T., Stuurman, N., Min, J., Manley, S., Unser, M.: Quantitative evaluation of software packages for single-molecule localization microscopy. Nature Methods **12**(8), 717–724 (2015)
32. Sage, D., Pham, T.A., Babcock, H., Lukes, T., Pengo, T., Chao, J., Velmurugan, R., Herbert, A., Agrawal, A., Colabrese, S., et al.: Super-resolution fight club: assessment of 2d and 3d single-molecule localization microscopy software. Nature Methods **16**(5), 387–395 (2019)
33. Schermelleh, L., Ferrand, A., Huser, T., Eggeling, C., Sauer, M., Biehlmaier, O., Drummen, G.P.: Super-resolution microscopy demystified. Nature cell biology **21**(1), 72–84 (2019)
34. Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., Preibisch, S., Rueden, C., Saalfeld, S., Schmid, B., et al.: Fiji: an open-source platform for biological-image analysis. Nature Methods **9**(7), 676–682 (2012)
35. Schmidt, M., Roux, N.L., Bach, F.R.: Convergence rates of inexact proximal-gradient methods for convex optimization. In: Advances in Neural Information Processing Systems, pp. 1458–1466 (2011)
36. Strang, G.: Approximation in the finite element method. Numerische Mathematik **19**(1), 81–98 (1972)
37. Tao, S., Boley, D., Zhang, S.: Local linear convergence of ista and fista on the lasso problem. SIAM Journal on Optimization **26**(1), 313–336 (2016)
38. Unser, M., Fageot, J., Gupta, H.: Representer theorems for sparsity-promoting ℓ^1 regularization. IEEE Transactions on Information Theory **62**(9), 5167–5180 (2016)
39. Villa, S., Salzo, S., Baldassarre, L., Verri, A.: Accelerated and inexact forward-backward algorithms. SIAM Journal on Optimization **23**(3), 1607–1633 (2013)
40. Yu, J., Lai, R., Li, W., Osher, S.: A fast proximal gradient method and convergence analysis for dynamic mean field planning. arXiv preprint arXiv:2102.13260 (2021)

A Proofs for FISTA convergence

This section contains all of the statements and proofs of the results contained in Section 4. Recall that the subsets $\mathbb{U}^n \subset \mathbb{H}$ satisfy (10).

A.1 Proofs for Step 3

Theorem 5 (Lemma 2) *Let $w_n \in \mathbb{U}^n$ be chosen arbitrarily and u_n/v_n be generated by Algorithm 1 for all $n \in \mathbb{N}$. For all $n > 0$, it holds that*

$$t_n^2(\mathrm{E}(u_n) - \mathrm{E}(w_n)) - (t_n^2 - t_n)(\mathrm{E}(u_{n-1}) - \mathrm{E}(w_n)) \leq \tfrac{1}{2}\left[\|v_{n-1}\|^2 - \|v_n\|^2\right] + \langle v_n - v_{n-1}, w_n\rangle . \tag{74}$$

Proof Modifying [10, Thm 3.2], for $n \geq 1$ we apply Lemma 1 with $\overline{u} = \overline{u}_{n-1}$ and $w = (1 - \frac{1}{t_n})u_{n-1} + \frac{1}{t_n}w_n$. By (10), $u_{n-1} \in \mathbb{U}^n$ is convex so $w \in \mathbb{U}^n$. This gives

$$\mathrm{E}(u_n) + \tfrac{1}{2}\left\|\tfrac{1}{t_n}v_n - \tfrac{1}{t_n}w_n\right\|^2 \leq \mathrm{E}\left((1 - \tfrac{1}{t_n})u_{n-1} + \tfrac{1}{t_n}w_n\right) + \tfrac{1}{2}\left\|\tfrac{1}{t_n}v_{n-1} - \tfrac{1}{t_n}w_n\right\|^2 . \tag{75}$$

By the convexity of E, this reduces to

$$\mathrm{E}(u_n) - \mathrm{E}(w_n) - (1 - \tfrac{1}{t_n})[\mathrm{E}(u_{n-1}) - \mathrm{E}(w_n)] \leq \tfrac{1}{2t_n^2}\|v_{n-1} - w_n\|^2 - \tfrac{1}{2t_n^2}\|v_n - w_n\|^2 = \tfrac{1}{2t_n^2}\left[\|v_{n-1}\|^2 - \|v_n\|^2\right] + \tfrac{1}{t_n^2}\langle v_n - v_{n-1}, w_n\rangle . \tag{76}$$

Multiplying through by t_n^2 gives the desired inequality. □

Theorem 6 (Theorem 1) *Fix a sequence of subsets $(\mathbb{U}^n)_{n\in\mathbb{N}}$ satisfying (10), arbitrary $u_0 \in \mathbb{U}^0$, and FISTA stepsize choice $(t_n)_{n\in\mathbb{N}}$. Let u_n and v_n be generated by Algorithm 1, then, for any choice of $w_n \in \mathbb{U}^n$ and $N \in \mathbb{N}$ we have*

$$t_N^2\,\mathrm{E}_0(u_N) + \sum_{n=1}^{N-1}\rho_n\,\mathrm{E}_0(u_n) + \frac{\|v_N - w_N\|^2}{2} \leq \frac{\|u_0 - w_0\|^2 - \|w_0\|^2 + \|w_N\|^2}{2} + \sum_{n=1}^{N} t_n\,\mathrm{E}_0(w_n) + \langle v_{n-1}, w_{n-1} - w_n\rangle . \tag{77}$$

Proof Theorem 6 is just a summation of (74) over all $n = 1, \ldots, N$. To see this: first add and subtract $\inf_{u\in\mathbb{H}}\mathrm{E}(u)$ to each term on the left-hand side to convert E to E_0, then move $\mathrm{E}_0(w_n)$ to the right-hand side. Now (74) becomes

$$t_n^2\,\mathrm{E}_0(u_n) - (t_n^2 - t_n)\,\mathrm{E}_0(u_{n-1}) \leq t_n\,\mathrm{E}_0(w_n) + \tfrac{1}{2}\left[\|v_{n-1}\|^2 - \|v_n\|^2\right] + \langle v_n - v_{n-1}, w_n\rangle . \tag{78}$$

Summing this inequality from $n = 1$ to $n = N$ gives

$$t_N^2\,\mathrm{E}_0(u_N) + \sum_{n=1}^{N-1}\underbrace{(t_n^2 - t_{n+1}^2 + t_{n+1})}_{=\rho_n}\mathrm{E}_0(u_n) \leq \frac{\|v_0\|^2 - \|v_N\|^2}{2} + \sum_{n=1}^{N} t_n\,\mathrm{E}_0(w_n) + \langle v_n - v_{n-1}, w_n\rangle . \tag{79}$$

The final step is to flip the roles of v_n/w_n in the final inner product term. Re-writing the right-hand side gives

$$\sum_{n=1}^{N}\langle v_n - v_{n-1}, w_n\rangle = \langle v_N, w_N\rangle - \langle v_0, w_0\rangle + \sum_{n=1}^{N}\langle v_{n-1}, w_{n-1} - w_n\rangle . \tag{80}$$

Noting that $v_0 = u_0$, the previous two equations combine to prove the statement of Theorem 6. □

The following lemma is used to produce a sharper estimate on sequences t_n.

Lemma 10 *If $\rho_n = t_n^2 - t_{n+1}^2 + t_{n+1} \geq 0$, $t_n \geq 1$ for all $n \in \mathbb{N}$ then $t_n \leq n - 1 + t_1$.*

Proof This is trivially true for $n = 1$. Suppose true for $n-1$, the condition on ρ_{n-1} gives

$$t_n^2 - t_n \leq t_{n-1}^2 \leq (n-2+t_1)^2 = (n-1+t_1)^2 - 2(n-1+t_1) + 1. \tag{81}$$

Assuming the contradiction, if $t_n > n-1+t_1$ then the above equation simplifies to $n-1+t_1 < 1$. However, $t_1 \geq 1$ implying that $n < 1$ which completes the contradiction. □

Lemma 11 (Lemma 3) *Let u_n, v_n be generated by Algorithm 1 with $(\mathbb{U}^n)_{n\in\mathbb{N}}$ satisfying (10), $(n_k \in \mathbb{N})_{k\in\mathbb{N}}$ be a monotone increasing sequence, and choose*

$$\widetilde{w}_k \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$

for each $k \in \mathbb{N}$. If such a sequence exists, then for all $K \in \mathbb{N}$, $n_K \leq N < n_{K+1}$ we have

$$t_N^2 \mathrm{E}_0(u_N) + \sum_{n=1}^{N-1} \rho_n \mathrm{E}_0(u_n) + \frac{\|v_N - \widetilde{w}_K\|^2}{2} \leq C + \frac{\|\widetilde{w}_K\|^2}{2} + \frac{(N+1)^2 - n_K^2}{2} \mathrm{E}_0(\widetilde{w}_K) + \sum_{k=1}^{K} \frac{n_k^2 - n_{k-1}^2}{2} \mathrm{E}_0(\widetilde{w}_{k-1}) + \left\langle v_{n_k-1}, \widetilde{w}_{k-1} - \widetilde{w}_k \right\rangle \tag{82}$$

where $C = \frac{\|u_0 - \widetilde{w}_0\|^2 - \|\widetilde{w}_0\|^2}{2}$.

Proof This is just a telescoping of the right-hand side of (77) with the introduction of n_k and simplification $w_n = \widetilde{w}_k$,

$$\tfrac{1}{2}\|w_N\|^2 + \sum_{n=1}^{N} t_n \mathrm{E}_0(w_n) + \langle v_{n-1}, w_{n-1} - w_n \rangle = \tfrac{1}{2}\|\widetilde{w}_K\|^2 + \sum_{n=n_K}^{N} t_n \mathrm{E}_0(\widetilde{w}_K) + \sum_{k=1}^{K} \sum_{n=n_{k-1}}^{n_k-1} t_n \mathrm{E}_0(\widetilde{w}_{k-1}) + \left\langle v_{n_k-1}, \widetilde{w}_{k-1} - \widetilde{w}_k \right\rangle. \tag{83}$$

By Lemma 10, $t_n \leq n$ so we can further simplify

$$\sum_{n=a}^{b-1} t_n \leq \sum_{n=a}^{b-1} n = (b-a)\frac{b-1+a}{2} \leq \frac{b^2 - a^2}{2}$$

to get the required bound. □

A.2 Proof for Step 4

Lemma 12 (Lemma 4) *Suppose $\mathbb{U}^n$, u_n, v_n and n_k satisfy the conditions of Lemma 3 and $(\widetilde{w}_k)_{k\in\mathbb{N}}$ forms an $(a_\mathrm{U}, a_\mathrm{E})$-minimising sequence of E with*

$$\widetilde{w}_k \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}.$$

If either:

- *$a_\mathrm{U} > 1$ and $n_k^2 \lesssim a_\mathrm{E}^k a_\mathrm{U}^{2k}$,*
- *or $a_\mathrm{U} = 1$, $\sum_{k=1}^{\infty} n_k^2 a_\mathrm{E}^{-k} < \infty$, and $\sum_{k=1}^{\infty} \|\widetilde{w}_k - \widetilde{w}_{k+1}\| < \infty$,*

then

$$\mathrm{E}_0(u_N) \lesssim \frac{a_\mathrm{U}^{2K}}{N^2} \qquad \text{for all} \qquad n_K \leq N < n_{K+1}.$$

Proof Starting from Lemma 11 we have

$$\begin{aligned} t_N^2 \mathrm{E}_0(u_N) + \tfrac{1}{2}\|v_N - \widetilde{w}_K\|^2 &\leq C + \frac{\|\widetilde{w}_K\|^2}{2} + \frac{(N+1)^2 - n_K^2}{2} \mathrm{E}_0(\widetilde{w}_K) \\ &\quad + \sum_{k=1}^{K} \frac{n_k^2 - n_{k-1}^2}{2} \mathrm{E}_0(\widetilde{w}_{k-1}) + \left\langle v_{n_k-1}, \widetilde{w}_k - \widetilde{w}_{k+1} \right\rangle \qquad (84) \\ &\leq C + \frac{\|\widetilde{w}_K\|^2}{2} + \frac{n_{K+1}^2}{2} \mathrm{E}_0(\widetilde{w}_K) \\ &\quad + \sum_{k=1}^{K} \frac{n_k^2}{2} \mathrm{E}_0(\widetilde{w}_{k-1}) + \left\langle v_{n_k-1} - \widetilde{w}_{k-1} + \widetilde{w}_{k-1}, \widetilde{w}_{k-1} - \widetilde{w}_k \right\rangle. \qquad (85) \end{aligned}$$

The inductive step now depends on the value of a_U.

Case $a_\mathrm{U} > 1$: We simplify the inequality

$$\begin{aligned} t_N^2 \mathrm{E}_0(u_N) + \tfrac{1}{2}\|v_N - \widetilde{w}_K\|^2 &\lesssim a_\mathrm{U}^{2K} + n_{K+1}^2 a_\mathrm{E}^{-K} + \sum_{k=1}^{K} n_k^2 a_\mathrm{E}^{-k} + a_\mathrm{U}^k \left\| v_{n_k-1} - \widetilde{w}_{k-1} \right\| + a_\mathrm{U}^{2k} \qquad (86) \\ &\leq C_1 \left[a_\mathrm{U}^{2K+2} + \sum_{k=1}^{K} a_\mathrm{U}^{2k} + a_\mathrm{U}^k \left\| v_{n_k-1} - \widetilde{w}_{k-1} \right\| \right] \qquad (87) \end{aligned}$$

for some $C_1 > C$. Choose $C_2 \geq \|v_{n_1-1} - \widetilde{w}_{1-1}\| a_{\mathrm{U}}^{-1}$ such that

$$\frac{1}{2}C_2^2 \geq \frac{C_1}{a_{\mathrm{U}}^2 - 1}(C_2 + a_{\mathrm{U}}^2). \tag{88}$$

Assume $\|v_{n_k-1} - \widetilde{w}_{k-1}\| \leq C_2 a_{\mathrm{U}}^k$ for $1 \leq k \leq K$ (trivially true for $K = 1$), then for $N = n_{K+1} - 1$ we have

$$\tfrac{1}{2}\|v_{n_{K+1}-1} - \widetilde{w}_K\|^2 \leq C_1 \left[a_{\mathrm{U}}^{2K+2} + \sum_{k=1}^{K} a_{\mathrm{U}}^{2k} + a_{\mathrm{U}}^k \|v_{n_k-1} - \widetilde{w}_{k-1}\| \right] \tag{89}$$

$$\leq C_1 \left[a_{\mathrm{U}}^{2K+2} + (1 + C_2)\frac{a_{\mathrm{U}}^{2K+2}}{a_{\mathrm{U}}^2 - 1} \right] \tag{90}$$

$$\leq \frac{C_1 a_{\mathrm{U}}^{2K+2}}{a_{\mathrm{U}}^2 - 1}\left(a_{\mathrm{U}}^2 + C_2\right) \leq \tfrac{1}{2}(C_2 a_{\mathrm{U}}^{K+1})^2. \tag{91}$$

Case $a_{\mathrm{U}} = 1$: Denote $b_k = \|\widetilde{w}_k - \widetilde{w}_{k+1}\|$ and note that $\|\widetilde{w}_{k-1}\| \leq \|\widetilde{w}_0\| + \sum_0^\infty b_k \lesssim 1$. We therefore bound

$$t_N^2 \,\mathrm{E}_0(u_N) + \tfrac{1}{2}\|v_N - \widetilde{w}_K\|^2 \lesssim 1 + n_{K+1}^2 a_{\mathrm{E}}^{-K} + \sum_{k=1}^{K} n_k^2 a_{\mathrm{E}}^{-k} + (\|v_{n_k-1} - \widetilde{w}_{k-1}\| + 1) b_k \tag{92}$$

$$\leq C_1 \left[1 + \sum_{k=1}^{K} \|v_{n_k-1} - \widetilde{w}_{k-1}\| b_{k-1} \right] \tag{93}$$

for some $C_1 > 0$. Choose $C_2 \geq \frac{\|v_{n_1-1} - \widetilde{w}_{1-1}\|}{\sum_0^\infty b_k}$ such that

$$\frac{1}{2}C_2^2 \geq C_1 \left(1 + C_2 \sum_0^\infty b_k \right). \tag{94}$$

Assume $\|v_{n_k-1} - \widetilde{w}_{k-1}\| \leq C_2$ for $1 \leq k \leq K$ (trivially true for $K = 1$), then for $N = n_{K+1} - 1$ we have

$$\tfrac{1}{2}\|v_{n_{K+1}-1} - \widetilde{w}_K\|^2 \leq C_1 \left[1 + \sum_{k=1}^{K} \|v_{n_k-1} - \widetilde{w}_{k-1}\| b_{k-1} \right] \leq C_1 \left(1 + C_2 \sum_0^\infty b_k \right) \leq \frac{C_2^2}{2} \tag{95}$$

In both cases, the induction on $\|v_{n_{K+1}-1} - \widetilde{w}_K\|$ holds for all K, and we have $t_N^2 \,\mathrm{E}_0(u_N) \leq \frac{1}{2}C_2^2 a_{\mathrm{U}}^{2K}$ for all $N < n_K - 1$. □

A.3 Proof for Step 5

Lemma 13 **(Lemma 5)** *Suppose u_n and n_k are sequences satisfying*

$$\forall N \in [n_K, n_{K+1}),\ \mathrm{E}_0(u_N) \lesssim \frac{a_{\mathrm{U}}^{2K}}{N^2} \qquad \textit{where} \qquad n_K^2 \gtrsim a_{\mathrm{E}}^K a_{\mathrm{U}}^{2K},$$

then

$$\mathrm{E}_0(u_N) \lesssim \frac{1}{N^{2(1-\kappa)}} \qquad \textit{where} \qquad \kappa = \frac{\log a_{\mathrm{U}}^2}{\log a_{\mathrm{E}} + \log a_{\mathrm{U}}^2}.$$

Proof The proof is direct computation, note that

$$(a_{\mathrm{E}} a_{\mathrm{U}}^2)^\kappa = \exp\left(\kappa \log(a_{\mathrm{E}} a_{\mathrm{U}}^2)\right) = \exp(\log a_{\mathrm{U}}^2) = a_{\mathrm{U}}^2, \tag{96}$$

therefore

$$a_{\mathrm{U}}^{2K} = \left((a_{\mathrm{E}} a_{\mathrm{U}}^2)^K\right)^\kappa \lesssim n_K^{2\kappa} \leq N^{2\kappa}, \tag{97}$$

so $\mathrm{E}_0(u_N) \lesssim N^{-2(1-\kappa)}$ as required. □

A.4 Proofs for Step 6

Theorem 7 **(Theorem 3)** *Let $(\mathbb{U}^n \subset \mathbb{H})_{n\in\mathbb{N}}$ be a sequence of subsets satisfying (10), compute u_n and v_n by Algorithm 1. Suppose that there exists a monotone increasing sequence $n_k \in \mathbb{N}$ such that*

$$\widetilde{w}_k := u_{n_k-1} \in \mathbb{U}^{n_k} \cap \mathbb{U}^{n_k+1} \cap \ldots \cap \mathbb{U}^{n_{k+1}-1}$$

for all $k \in \mathbb{N}$.

If $(\widetilde{w}_k)_{k\in\mathbb{N}}$ is an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of E with $a_{\mathrm{U}} > 1$ and $n_k^2 \lesssim a_{\mathrm{E}}^k a_{\mathrm{U}}^{2k}$, then

$$\min_{n\leq N} \mathrm{E}_0(u_n) = \min_{n\leq N} \mathrm{E}(u_n) - \inf_{u\in\mathbb{H}} \mathrm{E}(u) \lesssim \frac{1}{N^{2(1-\kappa)}} \qquad \textit{where} \qquad \kappa = \frac{\log a_{\mathrm{U}}^2}{\log a_{\mathrm{E}} + \log a_{\mathrm{U}}^2}$$

uniformly for $N \in \mathbb{N}$.

Proof Let $C > 0$ satisfy $n_k^2 \leq C a_{\mathrm{E}}^k a_{\mathrm{U}}^{2k}$ for each $k \in \mathbb{N}$. Fix $N > C$ and choose k such that $C a_{\mathrm{E}}^{k-1} a_{\mathrm{U}}^{2k-2} \leq N < C a_{\mathrm{E}}^k a_{\mathrm{U}}^{2k}$. By construction, and using the equality from (96), we have

$$\min_{n \leq N} \mathrm{E}_0(u_N) \leq \mathrm{E}_0(\widetilde{w}_{k-1}) \lesssim a_{\mathrm{E}}^{-k} = (a_{\mathrm{E}} a_{\mathrm{U}}^2)^{-k(1-\kappa)} < C^{\kappa-1} N^{-2(1-\kappa)} \tag{98}$$

as required. □

Lemma 14 **(Lemma 6)** *Let $(\widetilde{w}_k)_{k\in\mathbb{N}}$ be a sequence in $\mathbb{H}$ with $\|\widetilde{w}_k\| \lesssim a_{\mathrm{U}}^k$. Suppose $\widetilde{w}_k \in \widetilde{\mathbb{U}}^k := \mathbb{U}^{n_k}$ and denote $\mathrm{E}_0(\widetilde{\mathbb{U}}^k) := \inf_{u\in\widetilde{\mathbb{U}}^k} \mathrm{E}_0(u)$. Any of the following conditions are sufficient to show that $\widetilde{w}_k$ is an $(a_{\mathrm{U}}, a_{\mathrm{E}})$-minimising sequence of* E*:*

1. *Small continuous gap refinement:* $\mathrm{E}_0(\widetilde{w}_k) \leq \beta a_{\mathrm{E}}^{-k}$ *for all* $k \in \mathbb{N}$*, some* $\beta > 0$*.*
2. *Small discrete gap refinement:* $\mathrm{E}_0(\widetilde{\mathbb{U}}^k) \leq \beta a_{\mathrm{E}}^{-k}$ *and* $\mathrm{E}_0(\widetilde{w}_k) - \mathrm{E}_0(\widetilde{\mathbb{U}}^{k-1}) \leq \beta a_{\mathrm{E}}^{-k}$ *for all* $k > 0$*, some* $\beta > 0$*.*

Otherwise, suppose there exists a Banach space $(\mathbb{U}, |||\cdot|||)$ which contains each $\widetilde{\mathbb{U}}^k$, $\sup_{k\in\mathbb{N}} |||\widetilde{w}_k||| < \infty$, and the sublevel sets of E *are $|||\cdot|||$-bounded. With the subdifferential $\partial \mathrm{E} \colon \mathbb{U} \rightrightarrows \mathbb{U}^*$, it is also sufficient if either:*

3. *Small continuous gradient refinement:* $\sup_{u\in\mathbb{U}} \inf_{v\in\partial \mathrm{E}(\widetilde{w}_k)} \frac{|\langle v,u\rangle|}{|||u|||} \leq \beta a_{\mathrm{E}}^{-k}$ *for all* $k \in \mathbb{N}$*, some* $\beta > 0$*.*
4. *Small discrete gradient refinement:* $\mathrm{E}_0(\widetilde{\mathbb{U}}^k) \leq \beta a_{\mathrm{E}}^{-k}$ *and* $\sup_{u,\widetilde{w}\in\widetilde{\mathbb{U}}^k} \inf_{v\in\mathbb{V}^k} \frac{|\langle v,u-\widetilde{w}\rangle|}{|||u-\widetilde{w}|||} \leq \beta a_{\mathrm{E}}^{-k}$ *for all* $k \in \mathbb{N}$*, some* $\beta > 0$*, where* $\mathbb{V}^k := \partial(\mathrm{E}|_{\widetilde{\mathbb{U}}^k})(\widetilde{w}_k)$*.*

Proof The conditions for a_{U} in Definition 1 are already met, it remains to be shown that $\mathrm{E}_0(\widetilde{w}_k) \leq C a_{\mathrm{E}}^{-k}$ for some fixed $C > 0$. For cases (3) and (4), fix $R > 0$ such that both $\{\widetilde{w}_k\}_{k\in\mathbb{N}}$ and the sublevel set $\{u \in \mathbb{U} \text{ s.t. } \mathrm{E}_0(u) \leq 1+\beta\}$ are contained in the ball of radius R. Any minimising sequences of E in $\mathbb{U}$ or $\widetilde{\mathbb{U}}^k$ are contained in this ball. We can therefore compute C in each case:

(1) $\mathrm{E}_0(\widetilde{w}_k) \leq \beta a_{\mathrm{E}}^{-k}$, so $C = \beta$ suffices.
(2) $\mathrm{E}_0(\widetilde{w}_k) \leq \mathrm{E}_0(\widetilde{\mathbb{U}}^k) + \beta a_{\mathrm{E}}^{-k} \leq (a_{\mathrm{E}}+1)\beta a_{\mathrm{E}}^{-k}$, so $C = (a_{\mathrm{E}}+1)\beta$ suffices.
(3) $\mathrm{E}_0(\widetilde{w}_k) - \mathrm{E}_0(u) \leq \inf_{v\in\partial \mathrm{E}(\widetilde{w}_k)} \langle v, \widetilde{w}_k - u\rangle \leq 2R\beta a_{\mathrm{E}}^{-k}$ for any $u \in \mathbb{U}$ with $|||u||| \leq R$. Maximising over u gives $C = 2R\beta$
(4) $\mathrm{E}_0(\widetilde{w}_k) - \mathrm{E}_0(u) \leq \inf_{v\in\partial \mathrm{E}(\widetilde{w}_k)} \langle v, \widetilde{w}_k - u\rangle \leq 2R\beta a_{\mathrm{E}}^{-k}$ for any $u \in \widetilde{\mathbb{U}}^k$ with $|||u||| \leq R$, so $\mathrm{E}_0(\widetilde{w}_k) \leq \mathrm{E}_0(\widetilde{\mathbb{U}}^k) + 2R\beta a_{\mathrm{E}}^{-k}$ and $C = (1+2R)\beta$.

This completes the requirements of Definition 1. □

B Proof of Theorem 4

First we recall the setting of Definition 2, fix: $p \geq 0$, $q \in [1,\infty]$, $h \in (0,1)$, $N \in \mathbb{N}$, connected and bounded domain $\Omega \subset \mathbb{R}^d$, and $u^* \in \operatorname{argmin}_{u\in\mathbb{U}} \mathrm{E}(u)$. We assume that $\mathbb{H} = L^2(\Omega)$, $\|\cdot\|_q \lesssim |||\cdot|||$, and there exist spaces $(\widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ with $\widetilde{\mathbb{U}}^k \subset \mathbb{U}$ containing a sequence $(\widetilde{w}_k \in \widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ such that $|||\widetilde{w}_k - u^*||| \lesssim h^{kp}$, c.f. (16). Furthermore, there exists constant $c_\alpha > 0$ and meshes $\mathbb{M}^k$ such that:

$$\exists \omega_0 \subset \Omega \quad \text{such that} \quad \forall \omega \in \mathbb{M}^k \quad \exists (\alpha_\omega, \vec{\beta}_\omega) \in \mathbb{R}^{d\times d} \times \mathbb{R}^d \quad \text{such that} \quad \vec{x} \in \omega_0 \iff \alpha_\omega \vec{x} + \vec{\beta}_\omega \in \omega, \qquad \text{and} \tag{99}$$

$$\forall (\widetilde{u}, \omega) \in \widetilde{\mathbb{U}}^k \times \mathbb{M}^k, \quad \exists u \in \widetilde{\mathbb{U}}^0 \quad \text{such that} \quad \det(\alpha_\omega) \geq c_\alpha h^{kd} \quad \text{and} \quad \forall \vec{x} \in \omega_0,\ u(\vec{x}) = \widetilde{u}(\alpha_\omega \vec{x} + \vec{\beta}_\omega). \tag{100}$$

In this section, these assumptions will be summarised simply by saying that $\mathbb{H}$ and $(\widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ satisfy Definition 2. We prove Theorem 4 as a consequence of Lemma 7, namely we compute exponents p', q' with $a_{\mathrm{U}} = h^{-q'}$ and $a_{\mathrm{E}} = h^{-p'}$. These values are computed as the result of the following three lemmas. The first, Lemma 15, is a quantification of the equivalence between L^q and L^2 norms on general sub-spaces. Lemma 16 applies this result to finite-element spaces to compute the value of q'. Finally, Lemma 17 then performs the computations for p' depending on the smoothness properties of E.

Lemma 15 **(Equivalence of norms for fixed k)** *Suppose $\mathbb{H} = L^2(\Omega)$ for some connected, bounded domain $\Omega \subset \mathbb{R}^d$ and $\|\cdot\|_q \leq C|||\cdot|||$ for some $q \in [1,\infty]$, $C > 0$. For any linear subspace $\widetilde{\mathbb{U}} \subset \mathbb{U}$ and $\widetilde{w} \in \widetilde{\mathbb{U}}$,*

$$\|\widetilde{w}\| \leq \sup_{u,\widetilde{u}\in\widetilde{\mathbb{U}}} \frac{\langle u, \widetilde{u}\rangle}{\|u\|\,|||\widetilde{u}|||} |||\widetilde{w}||| \quad \textit{where} \quad \sup_{u,\widetilde{u}\in\widetilde{\mathbb{U}}} \frac{\langle u, \widetilde{u}\rangle}{\|u\|\,|||\widetilde{u}|||} \leq C^{-1} \begin{cases} |\Omega|^{\frac{1}{2}-\frac{1}{q}} & \textit{if } q \geq 2, \textit{ otherwise} \\ |\Omega|^{1-\frac{1}{q}} \sup_{u\in\widetilde{\mathbb{U}}} \|u\|_\infty / \|u\| & \textit{if } q \in [1,2). \end{cases} \tag{101}$$

Proof The first statement of the result is by definition, for each $\widetilde{w} \in \widetilde{\mathbb{U}} \subset L^\infty(\Omega) \subset \mathbb{H}$ we have

$$\|\widetilde{w}\| = \frac{\langle \widetilde{w}, \widetilde{w}\rangle}{\|\widetilde{w}\|} \leq \sup_{u\in\widetilde{\mathbb{U}}} \frac{\langle u, \widetilde{w}\rangle}{\|u\|\,|||\widetilde{w}|||} |||\widetilde{w}||| \leq \sup_{u,\widetilde{u}\in\widetilde{\mathbb{U}}} \frac{\langle u, \widetilde{u}\rangle}{\|u\|\,|||\widetilde{u}|||} |||\widetilde{w}|||.$$

Recall $|||\cdot||| \geq C^{-1}\|\cdot\|_q$. To go further we use Hölder's inequality. If $\frac{1}{q} + \frac{1}{q^*} = 1$, then for any $u, \widetilde{u} \in \widetilde{\mathbb{U}}$

$$\frac{\langle u, \widetilde{u}\rangle}{\|u\|\,|||\widetilde{u}|||} \leq C^{-1} \frac{\langle u, \widetilde{u}\rangle}{\|u\|\,\|\widetilde{u}\|_q} \leq C^{-1} \frac{\|u\|_{q^*}}{\|u\|}. \tag{102}$$

If $q \geq 2$ we use Hölder's inequality a second time:

$$\int_\Omega |u(\vec{x})|^{q^*} \,\mathrm{d}\vec{x} \leq \left(\int_\Omega 1 \,\mathrm{d}\vec{x}\right)^{1-q^*/2} \left(\int_\Omega |u(\vec{x})|^2 \,\mathrm{d}\vec{x}\right)^{q^*/2} = \left(|\Omega|^{\frac{1}{2}-\frac{1}{q}} \|u\|\right)^{q^*}. \tag{103}$$

This confirms the inequality when $q \geq 2$. If $q < 2$, we can simply upper bound $\|\cdot\|_{q^*} \leq |\Omega|^{\frac{1}{q^*}} \|\cdot\|_\infty$ as required. □

Lemma 16 *Suppose* $\mathbb{H}$ *and* $(\widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ *satisfy Definition 2, then*

1. *If* $q \geq 2$, *then* $\|\widetilde{w}_k\| \lesssim 1$ *(i.e.* $q' = 0$*).*
2. *If* $q < 2$ *and* $\sup_{u\in\widetilde{\mathbb{U}}^0} \frac{\|u\|_{L^\infty(\omega_0)}}{\|u\|_{L^2(\omega_0)}} < \infty$, *then* $\|\widetilde{w}_k\| \lesssim h^{-\frac{kd}{2}}$ *(i.e.* $q' = -\frac{d}{2}$*).*

Proof Most of the conditions of Lemma 15 are already satisfied. Furthermore observe that $|||\widetilde{w}_k||| \lesssim |||u^*||| + h^{kp} \lesssim 1$. For the $q \geq 2$ case, this is already sufficient to conclude $\|\widetilde{w}_k\| \lesssim 1$ from Lemma 15, as required.

For the case $q < 2$, from Lemma 15 recall that we are required to bound

$$\sup_{\widetilde{u}\in\widetilde{\mathbb{U}}^k} \frac{\|\widetilde{u}\|_\infty}{|||\widetilde{u}|||} = \sup_{\widetilde{u}\in\widetilde{\mathbb{U}}^k} \sup_{\omega\in\mathbb{M}^k} \frac{\|\widetilde{u}\|_{L^\infty(\omega)}}{|||\widetilde{u}|||_{L^2(\Omega)}} \leq \sup_{\widetilde{u}\in\widetilde{\mathbb{U}}^k} \sup_{\omega\in\mathbb{M}^k} \frac{\|\widetilde{u}\|_{L^\infty(\omega)}}{\|\widetilde{u}\|_{L^2(\omega)}}. \tag{104}$$

However, due to the decomposition property (100), for each $\omega \in \mathbb{M}^k$ and $\widetilde{u} \in \widetilde{\mathbb{U}}^k$ there exists $u \in \widetilde{\mathbb{U}}^0$ such that

$$\|u\|_{L^\infty(\omega_0)} = \|\widetilde{u}\|_{L^\infty(\omega)}, \qquad \|u\|^2_{L^2(\omega_0)} = \int_{\omega_0} |u(\vec{x})|^2 \,\mathrm{d}\vec{x} = \int_{\omega_0} |\widetilde{u}(\alpha\vec{x}+\vec{\beta})|^2 \,\mathrm{d}\vec{x} = \det(\alpha)^{-1}\|\widetilde{u}\|^2_{L^2(\omega)}. \tag{105}$$

Combining these two equations with the assumed bound on $\frac{\|u\|_{L^\infty(\omega_0)}}{\|u\|_{L^2(\omega_0)}}$ confirms $\|\widetilde{w}_k\| \lesssim \sqrt{\det(\alpha)^{-1}} \leq c_\alpha^{-\frac{1}{2}} h^{-\frac{kd}{2}}$ as required. □

Lemma 17 *Suppose* $\mathbb{H}$ *and* $(\widetilde{\mathbb{U}}^k)_{k\in\mathbb{N}}$ *satisfy Definition 2 and* u^* *is the minimiser of* E *such that* $|||\widetilde{w}_k - u^*||| \lesssim h^{kp}$.

1. *If* E *is* $|||\cdot|||$*-Lipschitz at* u^*, *then* $\mathrm{E}(\widetilde{w}_k) - \mathrm{E}(u^*) \lesssim h^{kp}$ *(i.e.* $p' = p$*).*
2. *If* $\nabla\mathrm{E}$ *is* $|||\cdot|||$*-Lipschitz at* u^*, *then* $\mathrm{E}(\widetilde{w}_k) - \mathrm{E}(u^*) \lesssim h^{2kp}$ *(i.e.* $p' = 2p$*).*

Proof Both statements are direct by definition, observe

$$\mathrm{E}(\widetilde{w}_k) - \mathrm{E}(u^*) \leq \mathrm{Lip}(\mathrm{E})|||\widetilde{w}_k - u^*|||, \tag{106}$$

$$\mathrm{E}(\widetilde{w}_k) - \mathrm{E}(u^*) \leq \langle \nabla\mathrm{E}(w), \widetilde{w}_k - u^*\rangle = \langle \nabla\mathrm{E}(\widetilde{w}_k) - \nabla\mathrm{E}(u^*), \widetilde{w}_k - u^*\rangle \leq \mathrm{Lip}(\nabla\mathrm{E})|||\widetilde{w}_k - u^*|||^2. \tag{107}$$

The proof is concluded by using the approximation bounds of $\widetilde{w}_k$ in Definition 2. □

C Operator norms for numerical examples

Theorem 8 *Suppose* $\mathsf{A}\colon \mathbb{H} \to \mathbb{R}^m$ *has kernels* $\psi_j \in L^\infty([0,1]^d)$ *for* $j \in [m]$.

Case 1: *If* $\psi_j(\vec{x}) = \begin{cases} 1 & \vec{x} \in \mathbb{X}_j \\ 0 & \text{else} \end{cases}$ *for some collection* $\mathbb{X}_j \subset \Omega$ *such that* $\mathbb{X}_i \cap \mathbb{X}_j = \emptyset$ *for all* $i \neq j$, *then* $\|\mathsf{A}\|_{L^2\to\ell^2} = \max_{j\in[m]} \sqrt{|\mathbb{X}_j|}$.

Case 2: *If* $\psi_j(\vec{x}) = \cos(\vec{a}_j \bullet \vec{x})$ *for some frequencies* $\vec{a}_j \in \mathbb{R}^d$ *with* $|\vec{a}_j| \leq A$, *then*

$$\|\mathsf{A}\|_{L^2\to\ell^2} \leq \sqrt{m}, \qquad |\mathsf{A}^*\vec{r}|_{C^k} \leq m^{1-\frac{1}{q}} A^k \|\vec{r}\|_q, \quad \text{and} \quad |\mathsf{A}^*|_{\ell^2\to C^k} \leq \sqrt{m} A^k$$

for all $\vec{r} \in \mathbb{R}^m$ *and* $q \in [1,\infty]$.

Case 3: *Suppose* $\psi_j(\vec{x}) = (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{|\vec{x}-\vec{x}_j|^2}{2\sigma^2}\right)$ *for some regular mesh* $\vec{x}_j \in [0,1]^d$ *and separation* Δ. *i.e.*

$$\{\vec{x}_j \text{ s.t. } j \in [m]\} = \{\vec{x}_0 + (j_1\Delta, \ldots, j_d\Delta) \text{ s.t. } j_i \in [\widehat{m}]\}$$

for some $\vec{x}_0 \in \mathbb{R}^d$, $\widehat{m} \coloneqq \sqrt[d]{m}$. *For all* $\frac{1}{q} + \frac{1}{q^*} = 1$, $q \in (1,\infty]$, *we have*

$$\|\mathsf{A}\|_{L^2\to\ell^2} \leq \left((4\pi\sigma^2)^{-\frac{1}{2}} \sum_{j=-2\widehat{m},\ldots,2\widehat{m}} \exp(-\tfrac{\Delta^2}{4\sigma^2} j^2)\right)^d, \tag{108}$$

$$|\mathsf{A}^*\vec{r}|_{C^0} \leq (2\pi\sigma^2)^{-\frac{d}{2}} \left(\sum_{\vec{j}\in J} \exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2} \max(0, |\vec{j}| - \delta)^2\right)\right)^{\frac{1}{q^*}} \|\vec{r}\|_q, \tag{109}$$

$$|\mathsf{A}^*\vec{r}|_{C^1} \leq \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma} \frac{\Delta}{\sigma} \left(\sum_{\vec{j}\in J} (|\vec{j}| + \delta)^{q^*} \exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2} \max(0, |\vec{j}| - \delta)^2\right)\right)^{\frac{1}{q^*}} \|\vec{r}\|_q, \tag{110}$$

$$|\mathsf{A}^*\vec{r}|_{C^2} \leq \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma^2} \left(\sum_{\vec{j}\in J} \left(1 + \tfrac{\Delta^2}{\sigma^2}(|\vec{j}| + \delta)^2\right)^{q^*} \exp\left(-\tfrac{q^*\Delta^2}{2\sigma^2} \max(0, |\vec{j}| - \delta)^2\right)\right)^{\frac{1}{q^*}} \|\vec{r}\|_q, \tag{111}$$

where $\delta = \frac{\sqrt{d}}{2}$ *and* $J = \{\vec{j} \in \mathbb{Z}^d \text{ s.t. } \left\|\vec{j}\right\|_{\ell^\infty} \leq 2\widehat{m}\}$. *The case for* $q = 1$ *can be inferred from the standard limit of* $\|\cdot\|_{q^*} \to \|\cdot\|_\infty$ *for* $q^* \to \infty$.

Proof *(Case 1.)* From Lemma 8 we have

$$(\mathsf{A}\mathsf{A}^*)_{i,j} = \left\langle \mathbb{1}_{\mathbb{X}_i}, \mathbb{1}_{\mathbb{X}_j} \right\rangle = |\mathbb{X}_i \cap \mathbb{X}_j| = \begin{cases} |\mathbb{X}_i| & i = j \\ 0 & i \neq j \end{cases}. \tag{112}$$

Therefore, $\mathsf{A}\mathsf{A}^*$ is a diagonal matrix and $\|\mathsf{A}\mathsf{A}^*\|_{\ell^2\to\ell^2} = \max_{j\in[m]} |\mathbb{X}_j|$ completes the result. □

Proof (Case 2.) ψ_j are not necessarily orthogonal however $|\langle \psi_i, \psi_j \rangle| \leq 1$ therefore we can estimate

$$\|\mathsf{A}\mathsf{A}^*\|_{\ell^2 \to \ell^2} \leq \|\mathsf{A}\mathsf{A}^*\|_{\ell^\infty \to \ell^\infty} \leq m. \tag{113}$$

Now looking to apply Lemma 9, note $\left\|\nabla^k \psi_j\right\|_\infty \leq A^k$, therefore

$$|\mathsf{A}^* \vec{r}|_{C^k} \leq A^k m^{\frac{1}{q^*}} \|\vec{r}\|_q = A^k m^{1-\frac{1}{q}} \|\vec{r}\|_q \quad \text{and} \quad |\mathsf{A}^*|_{\ell^2 \to C^k} \leq A^k \min_{q \in [1,\infty]} m^{1-\frac{1}{q}} \sqrt{m}^{\max(0,2-q)} = \sqrt{m} A^k. \tag{114}$$

□

Proof (Case 3.) In the Gaussian case, we build our approximations around the idea that sums of Gaussians should converge very quickly. The first example can be used to approximate the operator norm. Computing the inner products gives

$$\langle \psi_i, \psi_j \rangle = (2\pi\sigma^2)^{-d} \int_{[0,1]^d} \exp\left(-\tfrac{|\vec{x}-\vec{x}_i|^2}{2\sigma^2} - \tfrac{|\vec{x}-\vec{x}_j|^2}{2\sigma^2} \right) \mathrm{d}\vec{x} \leq (2\pi\sigma^2)^{-d} (\pi\sigma^2)^{\frac{d}{2}} \exp\left(-\tfrac{|\vec{x}_i-\vec{x}_j|^2}{4\sigma^2} \right). \tag{115}$$

Estimating the operator norm,

$$\|\mathsf{A}\mathsf{A}^*\|_{\ell^2 \to \ell^2} \leq \|\mathsf{A}\mathsf{A}^*\|_{\ell^\infty \to \ell^\infty} = \max_{i \in [m]} \sum_{j=1}^m |\langle \psi_i, \psi_j \rangle| \tag{116}$$

$$= \max_{i \in [m]} (4\pi\sigma^2)^{-\frac{d}{2}} \sum_{j_1, \ldots, j_d \in [\widehat{m}]} \exp\left(-\frac{(j_1\Delta - i_1\Delta)^2 + \ldots + (j_d\Delta - i_d\Delta)^2}{4\sigma^2} \right) \tag{117}$$

$$\leq (4\pi\sigma^2)^{-\frac{d}{2}} \sum_{\vec{j} \in \mathbb{Z}^d \cap [-\widehat{m}, \widehat{m}]^d} \exp\left(-\frac{(j_1\Delta)^2 + \ldots + (j_d\Delta)^2}{4\sigma^2} \right) = \left[(4\pi\sigma^2)^{-\frac{1}{2}} \sum_{j=-\widehat{m}}^{\widehat{m}} \exp\left(-\frac{\Delta^2 j^2}{4\sigma^2} \right) \right]^d. \tag{118}$$

This is a nice approximation because it factorises simply over dimensions. Applying the results from Lemma 9, note

$$\begin{aligned}
|\psi_j(\vec{x})| &= \left|\psi_j(\vec{x})\right| &&= (2\pi\sigma^2)^{-\frac{d}{2}} \exp\left(-\frac{|\vec{x}-\vec{x}_j|^2}{2\sigma^2} \right), \\
|\nabla\psi_j(\vec{x})| &= \left| \frac{\vec{x}-\vec{x}_j}{\sigma^2} \psi_j(\vec{x}) \right| &&= \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma} \frac{|\vec{x}-\vec{x}_j|}{\sigma} \exp\left(-\frac{|\vec{x}-\vec{x}_j|^2}{2\sigma^2} \right), \\
|\nabla^2\psi_j(\vec{x})| &= \left| \frac{1}{\sigma^2} + \frac{(\vec{x}-\vec{x}_j)(\vec{x}-\vec{x}_j)^\top}{\sigma^4} \right| \psi_j(\vec{x}) &&= \frac{(2\pi\sigma^2)^{-\frac{d}{2}}}{\sigma^2} \left(1 + \frac{|\vec{x}-\vec{x}_j|^2}{\sigma^2} \right) \exp\left(-\frac{|\vec{x}-\vec{x}_j|^2}{2\sigma^2} \right).
\end{aligned}$$

We now wish to sum over $j = 1, \ldots, m$ and produce an upper bound on these, independent of t. To do so we will use the following lemma.

Lemma 18 *Suppose $q > 0$. If the polynomial $p(|\vec{x}|) = \sum p_k |\vec{x}|^k$ has non-negative coefficients and $\vec{x} \in [-m, m]^d$, then*

$$\sum_{\|\vec{j}\|_{\ell^\infty} \leq m} p(|\vec{j}-\vec{x}|) \exp\left(-\tfrac{q|\vec{j}-\vec{x}|^2}{2} \right) \leq \sum_{\|\vec{j}\|_{\ell^\infty} \leq 2m} p(|\vec{j}|+\delta) \exp\left(-\frac{q \max(0, |\vec{j}|-\delta)^2}{2} \right)$$

where $\delta \coloneqq \frac{\sqrt{d}}{2}$ and $\vec{j} \in \mathbb{Z}^d$.

Proof There exists $\widehat{\vec{x}} \in [-\frac{1}{2}, \frac{1}{2}]^d$ such that $\vec{x} + \widehat{\vec{x}} \in \mathbb{Z}^d$, therefore

$$\begin{aligned}
\sum_{\|\vec{j}\|_{\ell^\infty} \leq m} p(|\vec{j}-\vec{x}|) \exp\left(-\tfrac{q|\vec{j}-\vec{x}|^2}{2} \right) &= \sum_{\|\vec{j}\|_{\ell^\infty} \leq m} p(|\vec{j} - (\vec{x}+\widehat{\vec{x}}) + \widehat{\vec{x}}|) \exp\left(-\tfrac{q|\vec{j}-(\vec{x}+\widehat{\vec{x}})+\widehat{\vec{x}}|^2}{2} \right) \\
&\leq \sum_{\|\vec{j}\|_{\ell^\infty} \leq 2m} p(|\vec{j}+\widehat{\vec{x}}|) \exp\left(-\tfrac{q|\vec{j}+\widehat{\vec{x}}|^2}{2} \right) \\
&\leq \sum_{\substack{\vec{j} \in \mathbb{Z}^d \\ \|\vec{j}\|_{\ell^\infty} \leq 2m}} p(|\vec{j}|+\delta) \exp\left(-\tfrac{q \max(0, |\vec{j}|-\delta)^2}{2} \right)
\end{aligned}$$

as $|\widehat{\vec{x}}| \leq \delta$ and p has non-negative coefficients. □

Now, continuing the proof of Theorem 8, for $\widehat{m} = \sqrt[d]{m}$, $\delta = \frac{\sqrt{d}}{2}$ and $J = \{\vec{j} \in \mathbb{Z}^d \text{ s.t. } \left\|\vec{j}\right\|_{\ell^\infty} \leq 2\widehat{m}\}$, Lemma 18 bounds

$$\begin{aligned}
\sum_{j=1}^m |\psi_j(\vec{x})|^{q^*} &\leq (2\pi\sigma^2)^{-\frac{dq^*}{2}} \left[\sum_{\vec{j} \in J} \exp\left(-\frac{q^*\Delta^2}{2\sigma^2} \max(0, |\vec{j}|-\delta)^2 \right) \right] \\
\sum_{j=1}^m |\nabla\psi_j(\vec{x})|^{q^*} &\leq \frac{(2\pi\sigma^2)^{-\frac{dq^*}{2}}}{\sigma^{q^*}} \frac{\Delta^{q^*}}{\sigma^{q^*}} \left[\sum_{\vec{j} \in J} (|\vec{j}|+\delta)^{q^*} \exp\left(-\frac{q^*\Delta^2}{2\sigma^2} \max(0, |\vec{j}|-\delta)^2 \right) \right] \\
\sum_{j=1}^m |\nabla^2\psi_j(\vec{x})|^{q^*} &\leq \frac{(2\pi\sigma^2)^{-\frac{dq^*}{2}}}{\sigma^{2q^*}} \left[\sum_{\vec{j} \in J} \left(1 + \frac{\Delta^2}{\sigma^2} (|\vec{j}|+\delta)^2 \right)^{q^*} \exp\left(-\frac{q^*\Delta^2}{2\sigma^2} \max(0, |\vec{j}|-\delta)^2 \right) \right]
\end{aligned}$$

for all $\vec{x} \in \Omega$. In a worst case, this is $O(2^d m)$ time complexity however the summands all decay faster than exponentially and so should converge very quickly. □