Machine Learning (2021) 110:2941–2989
https://doi.org/10.1007/s10994-021-06045-z

Robust non-parametric regression via incoherent subspace projections

Bhaskar Mukhoty[1] · Subhajit Dutta[1] · Purushottam Kar[1]

Received: 24 December 2020 / Revised: 14 June 2021 / Accepted: 26 July 2021 /
Published online: 5 September 2021

Abstract

This paper establishes the algorithmic principle of alternating projections onto incoherent low-rank subspaces (APIS) as a unifying principle for designing robust regression algorithms that offer consistent model recovery even when a significant fraction of training points are corrupted by an adaptive adversary. APIS offers the first algorithm for robust non-parametric (kernel) regression with an explicit breakdown point that works for general PSD kernels under minimal assumptions. APIS also offers, as straightforward corollaries, robust algorithms for a much wider variety of well-studied settings, including robust linear regression, robust sparse recovery, and robust Fourier transforms. Algorithms offered by APIS enjoy formal guarantees that are frequently sharper than (especially in non-parametric settings) or competitive to existing results in these settings. They are also straightforward to implement and outperform existing algorithms in several experimental settings.

1 Introduction and problem statement

We have a regression problem with n training points with regression values (aka *responses* or *signal*) denoted by the vector $\mathbf{a}^* \in \mathbb{R}^n$ (see below for examples). An adversary introduces additive corruptions $\mathbf{b}^* \in \mathbb{R}^n$ so that the responses we actually observe are given by

$$\mathbf{y} = \mathbf{a}^* + \mathbf{b}^*. \tag{1}$$

Can, under some conditions on $\mathbf{a}^*$ and $\mathbf{b}^*$, we recover $\mathbf{a}^*$ (as well as parameters in its generative model) despite the corruptions? This paper develops the APIS framework to answer this question in the affirmative.

Editors: Annalisa Appice, Sergio Escalera, Jose A. Gamez, Heike Trautmann.

✉ Purushottam Kar
purushot@cse.iitk.ac.in

Bhaskar Mukhoty
bhaskarm@cse.iitk.ac.in

Subhajit Dutta
duttas@iitk.ac.in

[1] Indian Institute of Technology Kanpur, Kanpur, India

Key idea: If $\mathbf{a}^* \in \mathscr{A}$ and $\mathbf{b}^* \in \mathscr{B}$, where $\mathscr{A}$ and $\mathscr{B}$ are unions of low-rank subspaces that are *incoherent* with respect to each other (see Sect. 1.1 for an introduction to incoherence and Sect. 6.1 for details), then this paper shows that such recovery is indeed possible. Specifically, let $\mathbf{a}^* \in \mathscr{A}$ and $\mathbf{b}^* \in \mathscr{B}$, where $\mathscr{A} = \bigcup_{i=1}^{P} A_i$ and $\mathscr{B} = \bigcup_{j=1}^{Q} B_j$ are the unions of subspaces, with $\operatorname{rank}(A_i) \leq s$ for all $i \in [P]$ and $\operatorname{rank}(B_j) \leq k$ for all $j \in [Q]$ for some integers $s, k > 0$. Then, this paper shows that it is possible to recover $\mathbf{a}^*$ consistently using a simple strategy that involves alternating projections onto these unions using projection operators $\Pi_{\mathscr{A}}(\cdot)$ and $\Pi_{\mathscr{B}}(\cdot)$ (see Algorithm 1). As we shall see, such incoherent unions of subspaces implicitly arise in several learning settings, e.g., if $\mathbf{a}^*$ and $\mathbf{b}^*$ are known to have s- and k-sparse representations in two bases that are incoherent to each other. We denote the privileged subspaces within these unions to which $\mathbf{a}^*$ and $\mathbf{b}^*$ belong as $A^* \ni \mathbf{a}^*$ and $B^* \ni \mathbf{b}^*$.

Algorithm 1 The APIS Algorithmic Framework

```
Input: Corrupted responses $\mathbf{y}$, Projection operators $\Pi_{\mathscr{A}}(\cdot), \Pi_{\mathscr{B}}(\cdot)$ that project onto $\mathscr{A}, \mathscr{B}$
Output: An estimate $\hat{\mathbf{a}}$ of the clean responses
Initialize $\mathbf{a}^0 \leftarrow \mathbf{0}$ and $t \leftarrow 0$
for $T = 1, 2, \ldots, T-1$ do
    $\mathbf{b}^{t+1} \leftarrow \Pi_{\mathscr{B}}(\mathbf{y} - \mathbf{a}^t)$        //denote $B^{t+1} \in \mathscr{B}$ to be a subspace that contains $\mathbf{b}^{t+1}$
    $\mathbf{a}^{t+1} \leftarrow \Pi_{\mathscr{A}}(\mathbf{y} - \mathbf{b}^{t+1})$    //denote $A^{t+1} \in \mathscr{B}$ to be a subspace that contains $\mathbf{a}^{t+1}$
    $t \leftarrow t+1$
end for
return $\mathbf{a}^T$
```

What all is known to APIS: APIS requires the projection operators $\Pi_{\mathscr{A}}(\cdot)$, $\Pi_{\mathscr{B}}(\cdot)$ to be executable efficiently at runtime as its alternating strategy may invoke these operators multiple times. Thus, it needs the unions $\mathscr{A}$ and $\mathscr{B}$ to be (implicitly) known. The discussions in Sects. 2 and 5.1, 5.2 assure that these conditions are indeed satisfied in several interesting learning applications. However, APIS does not require the vectors $\mathbf{a}^*$ and $\mathbf{b}^*$ to be known, nor does it assume that the subspaces A^* and B^* to which they belong are known, nor does it require A^* and B^* to be unique either.

1.1 A key to the manuscript

The reader may be curious about several questions that need solutions for the above strategy to make sense. We summarize APIS's solutions to these questions below but provide more details in subsequent discussions (that are *underlined in italics* for easy identification).

1. *How are $\mathscr{A}$ and $\mathscr{B}$ known to the algorithm?* Sect. 2 shows how in the case of robust linear regression, the union $\mathscr{A}$ is implicitly known the moment training data is made available. Section 5.1 shows that the same is true in several other important learning applications. Section 5.2 on the other hand, shows how the union $\mathscr{B}$ is defined for several interesting corruption models.
2. *How are the projection operators for these unions constructed and efficiently executed?* Projection onto unions of spaces can be intractable in general. Nevertheless, Sect. 5 shows how for several interesting learning applications, the projection operators $\Pi_{\mathscr{A}}(\cdot)$ and $\Pi_{\mathscr{B}}(\cdot)$ can be executed efficiently. Moreover, Table 1 gives explicit time complexity for these projection operations in a variety of applications.

3. *What if $\mathscr{A}$ and $\mathscr{B}$ are not incoherent to each other?* As discussed in Sects. 3 and 6.4, APIS can exploit *local* incoherence properties to guarantee recovery even when strict notions of incoherence fail to hold. See Sect. 2 for an introduction to incoherence.
4. *Does the adversary know $\mathscr{A}$ (or, perhaps even* $\mathbf{a}^*$*) before deciding* $\mathbf{b}^*$? As discussed in Sect. 4, APIS allows a *fully adaptive adversary* that is permitted to decide the corruption vector $\mathbf{b}^*$ with complete knowledge of $\mathbf{a}^*, A^*$ as well as $\mathscr{A}$ and $\mathscr{B}$.
5. *Is the model in Eq.* (1) *general enough to capture interesting applications and can the unions $\mathscr{A}$ and $\mathscr{B}$ be data-dependent?* The discussion in Sect. 5 shows that the model does indeed capture several statistical estimation and signal processing problems such as low-rank kernel regression, sparse signal transforms, robust sparse recovery, and robust linear regression. In most of these settings, the union $\mathscr{A}$ is indeed data-dependent.
6. *What if* $\mathbf{a}^*$ and $\mathbf{b}^*$ *are only approximate members of these unions?* As discussed in Sect. 6.4, APIS can readily accommodate compressible signals, where the clean signal $\mathbf{a}^*$ does not belong to $\mathscr{A}$ but is *well-approximated* by vectors in $\mathscr{A}$, as well as handle unmodelled errors such as simultaneous sparse corruptions and dense Gaussian noise.
7. *How low-rank must the subspaces in the unions be, i.e., how large can s and k be?* The key result in this paper Theorem 1 guarantees recovery the moment a certain incoherence requirement is satisfied. The pursuit of satisfying this requirement results in bounds on s and k to emerge for various applications. Table 1 summarizes the signal-corruption pairings for which APIS guarantees perfect recovery and bounds how large s and k can be. Detailed derivations of these results are presented in the appendices and summarized in Sect. 6.

2 A gentle introduction to the intuition behind APIS

We recall our model from Sect. 1. We have $\mathbf{y} = \mathbf{a}^* + \mathbf{b}^*$ with $\mathbf{a}^* \in \mathscr{A}$ and $\mathbf{b}^* \in \mathscr{B}$, where $\mathscr{A} = \bigcup_{i=1}^{P} A_i$ and $\mathscr{B} = \bigcup_{j=1}^{Q} B_j$ are the unions of subspaces, with rank $(A_i) \leq s$ for all $i \in [P]$ and rank $(B_j) \leq k$ for all $j \in [Q]$ for some integers $s, k > 0$. To present the core ideas behind APIS , we consider a simplified scenario where $P = 1 = Q$, i.e., the unions consist of a single subspace each $\mathscr{A} = \{A\}, \mathscr{B} = \{B\}$. As Definition 1 shows, we say two subspaces $A, B \subseteq \mathbb{R}^n$ (of possibly different ranks) are *μ-incoherent* for some $\mu > 0$, if $\forall\, \mathbf{u} \in A$, $\left\|\Pi_B(\mathbf{u})\right\|_2^2 \leq \mu \cdot \|\mathbf{u}\|_2^2$ and $\forall\, \mathbf{v} \in B$, $\left\|\Pi_A(\mathbf{v})\right\|_2^2 \leq \mu \cdot \|\mathbf{v}\|_2^2$. As the discussion after Definition 1 shows, an alternate interpretion of this property is that for any two unit vectors $\mathbf{a} \in A, \mathbf{b} \in B$, we must always have $(\mathbf{a}^\top \mathbf{b})^2 \leq \mu$. This means that if μ is small, then no two vectors from these two subspaces can be very aligned to each other and thus, the vectors must be near-orthonormal. Definition 2 extends the concept of incoherence to unions of subspaces.

Figure 1 illustrates this concept using a toy example where A is a rank-2 subspace of $\mathbb{R}^3$ and B is a rank-1 subspace of $\mathbb{R}^3$, i.e., $s = 2, k = 1$. Notice that in Fig. 1a, the subspaces A, B are highly incoherent indicating a value of $\mu \to 0$. It is not possible for two vectors, one each from A, B, to be very aligned to each other. On the other hand, Fig. 1b illustrates an example of a pair of subspaces that are quite *coherent* and have a high value of $\mu \to 1$. Also, shown in Fig. 1b are examples of two vectors $\mathbf{a}^* \in A, \mathbf{b}^* \in B$ that are extremely aligned to each other since the subspaces A, B are not incoherent and allow vectors to get very aligned.

Why incoherence helps robust recovery? To appreciate the benefits of incoherence, consider an extreme example where A, B are *perfectly* incoherent with $\mu = 0$

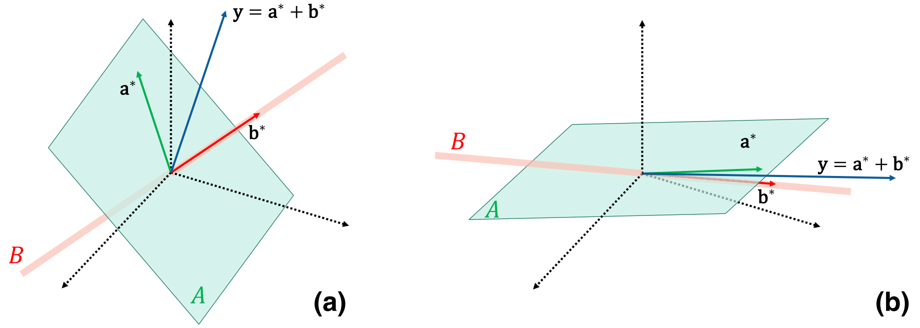


Fig. 1 Illustrating the distinction between a pair of incoherent subspaces (**a**) and a pair of coherent subspaces (**b**). For sake of simplicity, the unions $\mathscr{A}$ and $\mathscr{B}$ contain a single subspace in these examples

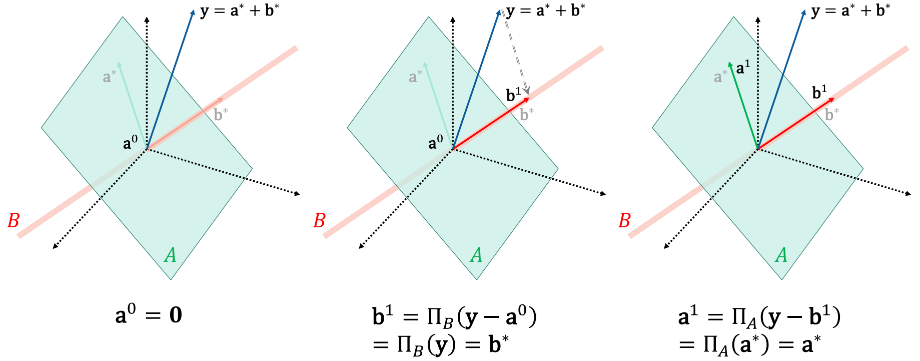


Fig. 2 An extreme example illustrating how perfect incoherence ($\mu = 0$) allows APIS to perform recovery in a single iteration of Algorithm 1. However APIS does not require perfect incoherence and guarantees recovery even if we only have $\mu < \frac{1}{9}$ (see Theorem 1) and even if only local incoherence is assured (see Sect. 6.4)

as illustrated in Fig. 2. For example, take $A = \left\{(x, y, z) \in \mathbb{R}^3, x + y + z = 0\right\}$ to be a rank-2 subspace of $\mathbb{R}^3$ and $B = \left\{(t, t, t) \in \mathbb{R}^3, t \in \mathbb{R}\right\}$ to be a rank-1 subspace of $\mathbb{R}^3$. Clearly, for any $(x, y, z) \in A, (t, t, t) \in B$, we have $\langle(x, y, z), (t, t, t)\rangle = t(x + y + z) = 0$. Now suppose the signal and corruption vectors are chosen as $\mathbf{a}^* \in A, \mathbf{b}^* \in B$ and we are presented with $\mathbf{y} = \mathbf{a}^* + \mathbf{b}^*$. Separating these two components is extremely simple in this case. To extract $\mathbf{b}^*$ from $\mathbf{y}$, we simply project $\mathbf{y}$ onto the subspace B to get $\Pi_B(\mathbf{y}) = \Pi_B(\mathbf{a}^* + \mathbf{b}^*) = \Pi_B(\mathbf{a}^*) + \Pi_B(\mathbf{b}^*) = \mathbf{0} + \mathbf{b}^* = \mathbf{b}^*$, where we have $\Pi_B(\mathbf{b}^*) = \mathbf{b}^*$ due to the idempotence of orthonormal projections and $\Pi_B(\mathbf{a}^*) = \mathbf{0}$ due to perfect incoherence between the two subspaces. Having done this, we can recover $\mathbf{a}^*$ by simply shaving off the contribution of $\mathbf{b}^*$ in $\mathbf{y}$ and projecting onto A to get $\Pi_A(\mathbf{y} - \mathbf{b}^*) = \Pi_A(\mathbf{a}^*) = \mathbf{a}^*$. It is easy to see that the above two steps are simply a single iteration of Algorithm 1. Thus, perfect incoherence allows straightforward recovery. Theorem 1 shows that APIS assures recovery even when the subspaces are reasonably incoherent but not perfectly incoherent. Section 6.4 extends this further to show how APIS offers recovery even in cases where only *local* incoherence is present in the task structure.

Table 1 Some signal and corruption models handled by APIS, and their corresponding breakdown points and per-iteration time complexity

Signal	Corruption	Breakdown point	Time per $\mathit{\Pi}_{\mathscr{A}}(\cdot)$	Time per $\mathit{\Pi}_{\mathscr{B}}(\cdot)$	References
Linear regression $\mathbf{x}^i \sim \mathscr{N}(\mathbf{0}, I_d)$	k-Sparse	$k < \frac{n}{154}$	$\mathscr{O}(nd)$	$\mathscr{O}(n \log n)$	Appendix B
Kernel regression w/gram matrix G $\mathbf{a}$ is s-sparse Sect. 5	k-Sparse	$3\Lambda_k^{\text{unif}}(G) < \lambda_s(G)$	$\mathscr{O}(ns) + \mathscr{O}\left(n^2 s\right)$ one time	As above	Appendix C
RBF kernel $\mathbf{x}^i \sim \text{unif}(S^{d-1})$	k-Sparse	$s < \mathscr{O}\left(\frac{\log n}{\log\log n}\right)$ $k < \mathscr{O}\left(\sqrt{n}\right)$	As above	As above	Appendix C
s-Sparse in either Fourier, Hadarmard or noiselet bases	k-Sparse	$sk < \frac{n}{9}$ e.g. $s, k \leq \frac{\sqrt{n}}{3}$	$\mathscr{O}(n \log n)$	As above	Foucart and Rauhut (2013)
s-Sparse in Fourier or wavelet (Haar, Daubechies D4/D8)	k Noiselet-sparse	$sk < \frac{n}{27}$	As above	As above	Candes and Wakin (2008) and Foucart and Rauhut (2013)
s Noiselet-sparse	k-Sparse in Fourier or wavelet (Haar, Daubechies D4/D8)	$sk < \frac{n}{27}$	As above	As above	Candes and Wakin (2008) and Foucart and Rauhut (2013)
s Haar-sparse and anti concentrated	k Fourier-sparse	$\frac{k^4}{s} + \frac{sk^2}{n} \leq \frac{1}{9}$	As above	As above	Section 6.4

For a vector $\mathbf{v}$, t-sparse means $\|\mathbf{v}\|_0 \leq t$ and t *blah*-sparse means $\mathbf{v}$ has a t-sparse representation in the basis *blah*. In the last two rows, the corruption signals, although sparse in some basis (e.g. noiselets), can be dense as vectors e.g. $\|\mathbf{b}^*\|_0 = n$, i.e. all n points suffer corruption. APIS offers recovery with such dense corruption whereas others e.g. Bafna et al. (2018), demand $\|\mathbf{b}^*\|_0 \ll n$. In the last row, APIS offers recovery with a pair of bases (Fourier, Haar wavelet), that is not incoherent but only satisfies *local* incoherence. The third row presents the special case for $d = 2$ (general case presented in Sect. 3)

Why lack of incoherence can make recovery impossible: To see why some form of incoherence is *essential* in general, consider a case similar to the one in Fig. 1b but taken to the extreme, i.e., where $\mu = 1$. To present a general case, let us take A, B to be higher rank spaces. For example, let $A = \left\{(a, b, c, d) \in \mathbb{R}^4, a + b = 0\right\}$ and $B = \left\{(p, q, 0, s) \in \mathbb{R}^4, p + q = 0\right\}$ be the two subspaces of $\mathbb{R}^4$ with ranks 3 and 2 respectively. Since the unit vector $\left(\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}}, 0, 0\right)$ lies in both A and B, the subspaces are coherent with $\mu = 1$. Suppose we are unlucky to have chosen the signal as $\mathbf{a}^* = (u, -u, 0, v) \in A$ for some $u, v \in \mathbb{R}$. Note that $\mathbf{a}^* \in A \cap B$. Then the adversary can readily choose $\mathbf{b}^* = (x, -x, 0, y) \in B$ for some secret values of $x, y \in \mathbb{R}$ that the adversary does not reveal to anybody. Recall that the adversary is allowed to choose $\mathbf{b}^*$ having seen the value of $\mathbf{a}^*$. Thus, we are presented with $\mathbf{y} = \mathbf{a}^* + \mathbf{b}^* = ((u + x), -(u + x), 0, (v + y))$. However, depending on x, y, this can be an arbitrary vector in the space B. Thus, recovering $\mathbf{a}^*$ becomes equivalent to recovering the secret values x, y which makes recovery impossible.

A real-life example: To make the above intuitions concrete, let us take the example of robust linear regression where we have $\mathbf{a}^* = X^\top \mathbf{w}^* \in \mathbb{R}^n$, where $X = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{d \times n}$ is the covariate matrix of the n data points and $\mathbf{w}^* \in \mathbb{R}^d$ is the linear model. In this case we always have $\mathbf{a}^* \in \mathrm{span}(\mathbf{x}^1, \ldots, \mathbf{x}^n)$, i.e. $P = 1$ and $\mathscr{A} = \{A\} = \mathrm{span}(\mathbf{x}^1, \ldots, \mathbf{x}^n)$. Suppose $\mathscr{B} = \{B\}$ with $B = A$ i.e., a completely coherent system with $\mu = 1$. In this case, the adversary can choose an *adversarial model* $\tilde{\mathbf{w}} \in \mathbb{R}^d$ and set $\mathbf{b}^* = X^\top(\tilde{\mathbf{w}} - \mathbf{w}^*) \in B$ so that we are presented with $\mathbf{y} = \mathbf{a}^* + \mathbf{b}^* = X^\top \tilde{\mathbf{w}}$. Since $\tilde{\mathbf{w}}$ is kept secret by the adversary, recovery yet again becomes impossible. On the other hand, as Table 1 and calculations in Sect. B in appendix show, if the adversary is restricted to impose only sparse corruptions, specifically $\mathscr{B} = \left\{\mathbf{b} \in \mathbb{R}^n, \|\mathbf{b}\|_0 \leq k\right\}$ for $k \leq \frac{n}{154}$, then $\mathscr{A}$ is sufficiently incoherent from $\mathscr{B}$ and APIS guarantees recovery.

3 Related works and our contributions in context

3.1 Summary of contributions

APIS presents a unified framework for designing robust (non-parametric) regression algorithms based on the principle of successive projections onto incoherent sub-spaces and applies it to various settings (see Sect. 5). APIS also offers explicit breakdown points and offers some of the fastest recoveries in experiments. Below, we give a survey of past works in these settings and offer comparisons with our contributions.

3.2 Robust non-parametric regression

Classical results in this are mostly relying on *robust* estimators such as Huber, L_1 and median (Cizek and Sadikoglu 2020; Fan et al. 1994), some of which (e.g., those based on Tukey's depth) are computationally intractable (Du et al. 2018). Please refer to recent reviews in Cizek and Sadikoglu (2020) and Du et al. (2018) for details. More recent work includes the LBM method (Du et al. 2018) that uses binning and median-based techniques.

Comparison: We compare experimentally to all these methods in Sect. 7. Classical techniques mostly do not offer explicit breakdown points; instead, they analyze the *influence function* of their estimators (Cizek and Sadikoglu 2020). Classical works and LBM also consider only Huber contamination models where the adversary is essentially stochastic.

In contrast, APIS offers explicit breakdown points against a fully adaptive adversary (see Sect. 4). LBM does not scale well with dimension d. Unless it receives $n = (\Omega(1))^d$ training points, it has to settle for coarse bins that increase the bias or face a situation where most bins are unpopulated, affecting the recovery. In contrast, APIS requires kernel ridge regression problems to be solved, for which efficient routines exist even for large d.

3.3 Robust linear regression

Past works adopt various strategies such as robust gradient methods e.g., SEVER (Diakonikolas et al. 2019), RGD (Prasad et al. 2018), hard thresholding techniques TORRENT (Bhatia et al. 2015), and reweighing techniques STIR (Mukhoty et al. 2019), apart from classical techniques based on *robust* loss functions such as Tukey's Bisquare and constrained L1-minimization based morphological component analysis (MCA) (McCoy and Tropp 2014).

Comparison: We compare experimentally to all these methods in Sect. 7. On the theoretical side, APIS offers more attractive guarantees. SEVER requires $n > d^5$ samples, whereas APIS requires $n > \Omega(d \log(d))$ samples. RGD offers theoretical guarantees only for Huber and heavy-tailed contamination models where the adversary is essentially stochastic, whereas APIS can tolerate a fully adaptive adversary (see Sect. 4). APIS offers much sharper guarantees (see Sect. 6.4) than TORRENT and STIR in the *hybrid* corruption case where apart from sparse corruptions, all points face Gaussian noise. However, TORRENT and STIR offer better breakdown points than APIS .

3.4 Robust Fourier and other signal transforms

Several works offer recovery of Fourier-sparse functions under sparse outliers, with the discrete cube or torus being candidate domains, and propose algorithms based on linear programming (Chen and De 2020; Guruswami and Zuckerman 2016). These offer good theoretical guarantees but are expensive (poly(n) runtime) to implement (the authors themselves offer no experimental work). On the other hand, APIS only requires "fast" transforms such as FFT to be carried out several times (and consequently, APIS offers an $\mathcal{O}(n \log n)$ runtime in these settings). Under sparse corruptions, APIS guarantees robust versions of several other transforms such as robust Hadamard transforms (see Table 1). APIS is additionally able to handle dense corruptions in special cases as well (see Sect. 6). The work of Bafna et al. (2018) uses an algorithm proposed by Baraniuk et al. (2010), in the context of performing robust Fourier transforms in the presence of sparse corruptions. However, their RIP-based analysis is restrictive and only applies to transforms such as Fourier, for which every entry of the design matrix is $\mathcal{O}\left(1/\sqrt{n}\right)$ (see Bafna et al. 2018, Theorem 2.2). This is not true of transforms, e.g., Haar wavelet, where design matrix entries can be $\Omega(1)$. APIS continues to give recovery guarantees even in such cases, and it can handle certain cases where corruptions are dense, i.e., $\|\mathbf{b}^*\|_0 = \Omega(n)$ which Bafna et al. (2018) do not consider.

3.5 Use of (local) incoherence in literature

The general principle of alternating projections and the notion of incoherence has been used in prior work. For example, Hegde and Baraniuk (2012) apply this principle to the problem of signal recovery on incoherent manifolds. However, our application of the alternating projection principle to robust non-parametric regression is novel and not addressed by prior work. Notions of incoherence and incoherent bases are also well-established in compressive sensing (Candes and Wakin 2008) and matrix completion (Chen 2015). However, to the best of our knowledge, APIS offers the first application of these notions to robust non-parametric recovery. It is well-known (Krahmer and Ward 2014; Zhou et al. 2016) that (global) incoherence may be unavailable in practical situations (e.g., Fourier and wavelet bases are not incoherent). Nevertheless, several results in compressive sensing (Krahmer and Ward 2014), matrix completion (Chen et al. 2014) and robust PCA (Zhang et al. 2015) assert that *local* notions of incoherence can still guarantee recovery. *Section 6.4 shows that* APIS *can as well exploit local incoherence properties to guarantee recovery in settings where strict notions of incoherence fail to hold.*

3.6 Learning incoherent spaces

An interesting line of work has pursued the goal of *learning* incoherent dictionaries for the task of classification (Schnass and Vandergheynst 2010; Barchiesi and Plumbley 2013, 2015). Specifically, a set of discriminative subspaces (sub-dictionaries) are learnt, one per class so as to offer discriminative advantage in supervised classification tasks. However, in the problem setting for APIS, as described after Eq. (1), the subspaces $\mathscr{A}, \mathscr{B}$ are well defined once once training data has been obtained and the corruption model has been fixed and thus, do not need to be learnt. For this reason, these works do not directly relate to the work in the current paper.

4 APIS: alternating projections onto incoherent subspaces

Adversary model: APIS *allows a fully adaptive adversary that is permitted to decide the corruption vector* $\mathbf{b}^*$ *with complete knowledge of* $\mathbf{a}^*, A^*$ *as well as* $\mathscr{A}$ and $\mathscr{B}$.

We note that this is the most potent adversary model considered in the literature. Specifically, given a pair of incoherent unions $\mathscr{A}$ and $\mathscr{B}$, first a subspace A^* and $\mathbf{a}^* \in A^*$ are chosen arbitrarily. The adversary is now told $A^*, \mathbf{a}^*$ and is then free to choose a subspace B^* in the union $\mathscr{B}$ and $\mathbf{b}^* \in B^*$ using its knowledge in any way.

APIS is described in Algorithm 1 and involves alternately projecting onto unions of subspaces $\mathscr{A}, \mathscr{B}$. For specific applications, the projection steps take on various forms, e.g., solving a (kernel) least-squares problem, a Fourier transform, or hard-thresholding. These are discussed in Sect. 5.

Notation: For $\mathbf{v} \in \mathbb{R}^d$ and set $F \subseteq [d]$, let $\mathbf{v}_F \in \mathbb{R}^d$ denote a vector with coordinates in the set F identical to those in $\mathbf{v}$ and others set to zero. For any matrix $X \in \mathbb{R}^{d\times n}$ and any sets $S \subseteq [n], F \subseteq [d]$, we let $X_S^F = [\tilde{x}_{ij}] \in \mathbb{R}^{d\times n}$ be a matrix such that $\tilde{x}_{ij} = x_{ij}$ if $i \in F, j \in S$ and $\tilde{x}_{ij} = 0$ otherwise. We similarly let $X^F = [z_{ij}]$ denote the matrix with entries in the rows in F identical to those in X and other entries zeroed out i.e. $z_{ij} = x_{ij}$ if $i \in F$ and $z_{ij} = 0$ otherwise. X_S is similarly defined as a matrix with entries in the columns in S identical to those in X and other entries zeroed out.

Projections: For any subspace $S \subseteq \mathbb{R}^n$, Π_S denotes orthonormal projection onto S and $\Pi_S^\perp$ denotes the orthonormal projection onto the ortho-complement of S so that for any $\mathbf{v} \in \mathbb{R}^n$ and any subspace S, we always have $\mathbf{v} = \Pi_S(\mathbf{v}) + \Pi_S^\perp(\mathbf{v})$. We abuse notation to extend the projection operator $\Pi_\cdot(\cdot)$ to unions of low-rank subspaces. Let $\mathscr{A} = \bigcup_{i=1}^{P} A_i \subseteq \mathbb{R}^n$ be a union of P subspaces, then for any $\mathbf{v} \in \mathbb{R}^n$ we define $\Pi_{\mathscr{A}}(\mathbf{v}) = \arg\min_{\mathbf{z}=\Pi_{A_i}(\mathbf{v}), i\in[P]} \|\mathbf{v} - \mathbf{z}\|_2^2$. Projection onto a union of subspaces is expensive in general (requiring time linear in P, the number of subspaces in the union) but will be efficient in all cases we consider (see Table 1).

Hard thresholding: The hard-thresholding operator will be instrumental in allowing efficient projections in APIS. For any $n, k < n$, let $\mathscr{S}_k^n = \left\{\mathbf{z} \in \mathbb{R}^n : \|\mathbf{z}\|_0 \leq k\right\}$ be the set of all k-sparse vectors. For any $\mathbf{z} \in \mathbb{R}^n, k < n$, let $\mathrm{HT}_k(\mathbf{z}) := \Pi_{\mathscr{S}_k^n}(\mathbf{z})$ denote the projection of $\mathbf{z}$ onto $\mathscr{S}_k^n$. Note that this operation is possible in $\mathscr{O}(n \log n)$ time by sorting all the coordinates by magnitude, retaining the top k coordinates (in magnitude) and setting rest to 0.

5 Applications and projection details

The signal and corruption model in Eq. (1) does indeed capture several statistical estimation and signal processing problems. In most of these settings, the union $\mathscr{A}$ is indeed data-dependent. The discussion below shows that in each case, the union of subspaces $\mathscr{A}$ is well-defined once training data is available. On the other hand, the union of subspaces $\mathscr{B}$ is well-defined once the corruption model has been identified. The projection operators $\Pi_{\mathscr{A}}(\cdot)$ and $\Pi_{\mathscr{B}}(\cdot)$ can be executed in polynomial time (see Table 1 for time complexity details).

5.1 Examples of signal models supported by APIS

Linear regression: Here we have $\mathbf{a}^* = X^\top \mathbf{w}^*$, where $X = [\mathbf{x}^1, \ldots, \mathbf{x}^n] \in \mathbb{R}^{d\times n}$ is the covariate matrix of the n data points and $\mathbf{w}^* \in \mathbb{R}^d$ is the linear model. It is easy to see that Eq. (1) recovers robust linear regression as a special case with $P = 1$ and $\mathscr{A} = A$, where $A = \mathrm{span}(\mathbf{x}^1, \ldots, \mathbf{x}^n)$. Using the SVD $X = U\Sigma V^\top$, we can project onto $\mathscr{A}$ simply by solving a least squares problem i.e. we have $\Pi_{\mathscr{A}}(\mathbf{z}) = VV^\top \mathbf{z} = (X^\top X^\dagger)\mathbf{z}$.

Low-rank kernel regression: Consider a Mercer kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such as the RBF kernel and let $G \in \mathbb{R}^{n\times n}$ be the Gram matrix with $G_{ij} = K(\mathbf{x}^i, \mathbf{x}^j)$. Low-rank kernel regression corresponds to the case when the uncorrupted signal satisfies $\mathbf{a}^* = G\boldsymbol{\alpha}^*$ where $\boldsymbol{\alpha}^* \in \mathbb{R}^n$ belongs to the span of the some s eigenvectors of G. Specifically, consider the eigendecomposition $G = V\Sigma V^\top$, $V = [\mathbf{v}^1, \ldots, \mathbf{v}^r] \in \mathbb{R}^{n\times r}$ is the matrix of eigenvectors (r is the rank of the Gram matrix) and $\Sigma = \mathrm{diag}\,(s_1, \ldots, s_r) \in \mathbb{R}^{r\times r}$ is the diagonal matrix of strictly positive eigenvalues (assume $s_1 \geq s_2 \geq \ldots \geq s_r > 0$). APIS offers the strongest guarantees in the case when $\boldsymbol{\alpha}^* \in \mathrm{span}(\mathbf{v}^1, \ldots, \mathbf{v}^s)$, i.e., when $\boldsymbol{\alpha}^*$ lies in the span of the the top eigenvectors. Note that in this case, $\mathbf{a}^*$ too is spanned by the top s eigenvectors of G since $\mathbf{a}^* = G\boldsymbol{\alpha}^*$. We stress that the guarantees continue to hold (see Sect. C in appendix) but deteriorate if $\boldsymbol{\alpha}^*, \mathbf{a}^*$ are spanned by s eigenvectors that include the lower ones as well. This is because Gram matrices corresponding to popular kernels such as the RBF kernel are often very ill-conditioned. Then, we can see that Eq. (1) recovers the robust low-rank kernel regression problem as a special case with $P = 1$ and $\mathscr{A} = A$, where $A = \mathrm{span}(\mathbf{v}^1, \ldots, \mathbf{v}^s)$. Projection onto $\mathscr{A} = A$ is given by $\Pi_{\mathscr{A}}(\mathbf{z}) = \Pi_A(\mathbf{z}) = V_s V_s^\top \mathbf{z}$, where $V_s = [\mathbf{v}^1, \ldots, \mathbf{v}^s] \in \mathbb{R}^{n\times s}$. Section 6.5 shows how this restriction of the signal to the span of the top-s eigenvectors does not affect the universality of popular kernels such as the RBF kernel.

Sparse signal transforms: Consider signal transforms such as Fourier, Hadamard, wavelet, etc. Sparse signal transforms correspond to the case when $\mathbf{a}^* = M^\top \boldsymbol{\alpha}^*$, where $M \in \mathbb{R}^{n\times n}$ is the design matrix of the transform (for sake of simplicity, assume M to be orthonormal as is often the case) and $\boldsymbol{\alpha}^* \in \mathbb{R}^n$ is an s-sparse vector, i.e., $\|\boldsymbol{\alpha}^*\|_0 \le s$. It is easy to see that Eq. (1) recovers the robust sparse signal transform problem with $P = \binom{n}{s}$ and $\mathscr{A} = \bigcup_{F\subset[n],|F|=s} A_F$ with $A_F = \mathrm{span}(\{\mathbf{m}^i\}_{i\in F})$, where $\mathbf{m}^i$ is the i^{th} column of the design matrix M. Given that M is orthonormal, projection onto a given subspace A_F is given by $\Pi_{A_F}(\mathbf{z}) = M_F M_F^\top \mathbf{z}$. The orthonormality of M can be further exploited to carry out projection onto the union $\mathscr{A}$ in $\mathscr{O}(n\log n)$ time by using "fast" versions of these transforms followed by a hard-thresholding operation. Specifically, we have $\Pi_{\mathscr{A}}(\mathbf{z}) = M\mathbf{v}$, where $\mathbf{v} = \mathrm{HT}_s(M^\top \mathbf{z})$.

Sparse recovery: In the sparse recovery signal model, the uncorrupted signal satisfies $\mathbf{a}^* = X^\top \mathbf{w}^*$ where $\mathbf{w}^* \in \mathbb{R}^d$ is an s^*-sparse linear model, i.e., $\|\mathbf{w}^*\|_0 \le s^*$ for some $s^* < d$. Equation (1) recovers the robust sparse recovery problem as a special case with $P = \binom{d}{s^*}$ and $\mathscr{A} = \bigcup^{*}_{F\subset[d],|F|=s} A_F$, where the subspace A_F is given by $A_F = \mathrm{span}(\mathbf{x}_F^1, \ldots, \mathbf{x}_F^n)$. Projection onto a given subspace A_F can be easily seen to be $\Pi_{A_F}(\mathbf{z}) = ((X^F)^\top (X^F)^\dagger)\mathbf{z}$. Projection onto the union $\mathscr{A}$ can then be shown to be simply the classical sparse recovery problem that can be solved efficiently if X satisfies properties such as RIP or RSC (see Agarwal et al. (2012)). Specifically, we have $\Pi_{\mathscr{A}}(\mathbf{z}) = X^\top \hat{\mathbf{w}}$, where $\hat{\mathbf{w}} = \arg\min^{*}_{\|\mathbf{w}\|_0 \le s} \|X^\top \mathbf{w} - \mathbf{z}\|_2^2$. The projection step $\Pi_{\mathscr{A}}(\cdot)$ can be carried out in $\mathscr{O}(nd)$ time here as well by employing projected gradient and iterative hard-thresholding methods (see Agarwal et al. (2012)).

5.2 Examples of corruption models supported by APIS

Sparse fully adaptive adversarial corruptions This is most widely studied case in literature and assumes a sparse corruption vector i.e. $\|\mathbf{b}^*\|_0 \le k$ for some $k < n$. The model in Eq. (1) recovers this case with $Q = \binom{n}{k}$ and $\mathscr{B} = \bigcup_{T\subset[n],|T|=k} B_T$ with B_T as the subspace of all vectors with support within the set T. Note that the convergence guarantees for APIS do not impose any restrictions on the magnitude of corruptions. Instead, the number of iterations required for recovery merely scale logarithmically with the L_2 norm of the corruption vector i.e. the runtime scales as $\log(\|\mathbf{b}^*\|_2)$ (see Theorem 1). It can be easily seen that the hard-thresholding operator $\mathrm{HT}_k(\cdot)$ (Sect. 4) offers projection onto the union $\mathscr{B}$.

Dense fully adaptive adversarial corruptions Unlike several previous works, APIS also allows corruption vectors that are dense $\|\mathbf{b}^*\|_0 = \Omega(n)$, i.e., most points suffer corruption. This is because APIS only requires the unions $\mathscr{A}, \mathscr{B}$ to be incoherent and does not care if $\mathscr{B}$ contains dense vectors. We will see such examples in Sect. 6, with noiselet corruptions, and in Sect. 6.4 where we will exploit *local incoherence* results to guarantee recovery when the signal is Fourier-sparse, and the corruptions are Wavelet-sparse. In Sect. 7, we will establish that APIS offers recovery in such dense corruption settings, experimentally as well.

As noted earlier, in both the corruption models, the adversary has full knowledge of $\mathbf{a}^*, A^*$ before choosing $\mathbf{b}^*, B^*$ in any manner, i.e., the adversary is *fully adaptive*.

5.3 Do the subspaces really need to be low-rank? What if this is too strict and $\mathbf{a}^* \notin \mathscr{A}$?

For exact recovery guarantees (which APIS does offer), some low-rank restriction seems to be necessary, especially when working with universal models such as the Gaussian kernel whose Gram matrix is often full-rank (and ill-conditioned), or the Fourier transform whose design matrix is also full-rank (but well-conditioned). Given such full-rank designs, unless additional restrictions are put (e.g., low-rank), recovery remains an ill-posed problem. However, in Sect. 6.4, we will see that APIS *offers non-trivial recovery even if the clean signal* $\mathbf{a}^*$ *does not belong to* $\mathscr{A}$ *but is well-approximated by vectors in* $\mathscr{A}$. Specifically, these are cases when $\mathbf{a}^* \notin \mathscr{A}$ but rather $\mathbf{a}^* + \mathbf{e}^* \in \mathscr{A}$ and $\|\mathbf{e}^*\|_2$ is small. It is common in signal processing tasks to consider signals (images, etc.) that are well-approximated by a sparse wavelet/Fourier representation but not exactly sparse themselves.

6 Recovery, breakdown points, misspecified models and universality

All detailed proofs and derivations are provided in the appendices.

6.1 Incoherence

A key requirement for robust recovery in model presented in Eq. (1) is for the unions $\mathscr{A}$ and $\mathscr{B}$ to be *incoherent* with respect to each other. We present below a notion of subspace incoherence suitable to our technique. We note that the notion presented below is similar to notions of subspace incoherence prevalent in literature but suited for our setting.

Definition 1 (*Subspace incoherence*) For any $\mu > 0$, we say two subspaces $A, B \subseteq \mathbb{R}^n$ (of possibly different ranks) are *μ-incoherent* if for all $\mathbf{u} \in A$, $\left\|\Pi_B(\mathbf{u})\right\|_2^2 \leq \mu \cdot \|\mathbf{u}\|_2^2$ and for all $\mathbf{v} \in B$, $\left\|\Pi_A(\mathbf{v})\right\|_2^2 \leq \mu \cdot \|\mathbf{v}\|_2^2$.

Note that the above definition uses the same incoherence constant μ for projections both ways. This is justified since $\max_{\mathbf{u}\in A} \frac{\|\Pi_B(\mathbf{u})\|_2^2}{\|\mathbf{u}\|_2^2} = \max_{\mathbf{v}\in B} \frac{\|\Pi_A(\mathbf{v})\|_2^2}{\|\mathbf{v}\|_2^2}$. To see why, let U and V be orthonormal matrices whose columns span A and B resp. and notice that $\mu = \left\|V^\top U\right\|_{\text{op}}^2 = \left\|U^\top V\right\|_{\text{op}}^2$ where $\|\cdot\|_{\text{op}}$ is the operator norm. Since orthonormal projections are non-expansive, we always have $\mu \leq 1$. However, our results demand stronger contractions which we will establish for our application settings discussed in Sect. 5. However, first we extend the notion of incoherence to unions of subspaces.

Definition 2 (*Subspace union (SU) incoherence*) For any $\mu > 0$, we say that a pair of unions of subspaces $\mathscr{A} = \bigcup_{i=1}^P A_i$ and $\mathscr{B} = \bigcup_{j=1}^Q B_j$ is μ-SU incoherent if for all $i \in [P]$ and all $j \in [Q]$, the subspaces A_i and B_j are μ-incoherent.

Theorem 1 states the main claim of this paper. Restrictions on s, k and breakdown points emerge when trying to satisfy the incoherence criterion demanded by Theorem 1.

Theorem 1 *Suppose we obtain data as described in Eq.* (1) *where the two unions* $\mathscr{A}$ and $\mathscr{B}$ *are* μ*-incoherent with* $\mu < \frac{1}{9}$*. Then, for any* $\epsilon > 0$ *within* $T = \mathscr{O}\left(\log \frac{\|\mathbf{a}^*\|_2+\|\mathbf{b}^*\|_2}{\epsilon}\right)$ *iterations,* APIS *offers* $\left\|\mathbf{a}^T - \mathbf{a}^*\right\|_2 \le \epsilon$*. Moreover, in the known signal support case when* $P = 1$ *(see below), the requirement is further relaxed to* $\mu < \frac{1}{3}$*.*

Proof (*Sketch*) We present the main steps in deriving the result for the known signal support case when $P = 1$. We recall the notation from Algorithm 1 where $A^{t+1} \ni \mathbf{a}^{t+1}, B^{t+1} \ni \mathbf{b}^{t+1}$, and let $\mathbf{p}^t = \Pi_A(\mathbf{b}^* - \mathbf{b}^t)$ and $\mathbf{p}^{t+1} = \Pi_A(\mathbf{b}^* - \mathbf{b}^{t+1})$. Thus, we have $\mathbf{a}^{t+1} = \Pi_A(\mathbf{y} - \mathbf{b}^{t+1}) = \mathbf{a}^* + \Pi_A(\mathbf{b}^* - \mathbf{b}^{t+1})$ (since $\mathbf{a}^* \in A$ and orthonormal projections are idempotent) which gives us $\left\|\mathbf{a}^{t+1} - \mathbf{a}^*\right\|_2 = \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^{t+1})\right\|_2 = \left\|\mathbf{p}^{t+1}\right\|_2$. Let $\mathfrak{Q} := B^{t+1} \cap B^*$ denote the meet of the two subspaces, as well as denote the symmetric difference subspaces $\mathfrak{P} := B^{t+1} \cap (B^*)^\perp$ and $\mathfrak{R} = B^* \cap (B^{t+1})^\perp$ (recall that $A \ni \mathbf{a}^*$ and $B^* \ni \mathbf{b}^*$).

Below we show that $\left\|\mathbf{p}^{t+1}\right\|_2 \le 3\mu \cdot \|\mathbf{p}^t\|_2$ that establishes a linear rate of convergence if $\mu < \frac{1}{3}$ as it grants $\left\|\mathbf{a}^{t+1} - \mathbf{a}^*\right\|_2 = \left\|\mathbf{p}^{t+1}\right\|_2 \le 3\mu \cdot \|\mathbf{p}^t\|_2 = 3\mu \cdot \|\mathbf{a}^t - \mathbf{a}^*\|_2$. To show that $\left\|\mathbf{p}^{t+1}\right\|_2 \le 3\mu \cdot \|\mathbf{p}^t\|_2$, we note that

$$\mathbf{b}^{t+1} = \Pi_{B^{t+1}}(\mathbf{a}^* + \mathbf{b}^* - \mathbf{a}^t) = \Pi_{B^{t+1}}(\mathbf{b}^* - \Pi_A(\mathbf{b}^* - \mathbf{b}^t)) = \Pi_{B^{t+1}}(\mathbf{b}^* - \mathbf{p}^t),$$

and thus $\mathbf{b}^* - \mathbf{b}^{t+1} = \mathbf{b}^* - \Pi_{B^{t+1}}(\mathbf{b}^* - \mathbf{p}^t) = \Pi_{\mathfrak{R}}(\mathbf{b}^*) + \Pi_{B^{t+1}}(\mathbf{p}^t)$. This gives us, by an application of the triangle inequality, $\left\|\mathbf{p}^{t+1}\right\|_2 = \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^{t+1})\right\|_2 \le \left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 + \left\|\Pi_A(\Pi_{B^{t+1}}(\mathbf{p}^t))\right\|_2$. Applying incoherence now tells us that, since $\mathbf{p}^t \in \mathscr{A}$ by projection, we have

$$\left\|\Pi_A(\Pi_{B^{t+1}}(\mathbf{p}^t))\right\|_2 \le \sqrt{\mu} \cdot \left\|\Pi_{B^{t+1}}(\mathbf{p}^t)\right\|_2 \le \mu \cdot \left\|\mathbf{p}^t\right\|_2$$

Using other arguments given in the full proof, it can be shown that $\left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 \le 2\mu \cdot \|\mathbf{p}^t\|_2$ which gives us $\left\|\mathbf{p}^{t+1}\right\|_2 \le 3\mu \cdot \|\mathbf{p}^t\|_2$ concluding the proof sketch.

Section A in the appendix gives the complete proof this result. APIS offers a stronger guarantee, requiring $\mu < \frac{1}{3}$, in the *known signal support* case (Chen and De 2020). These are cases when the union $\mathscr{A}$ consists of a single subspace, i.e., $P = 1$. Note that this is indeed the case (see Sect. 5) for linear regression and low-rank kernel regression. We now derive breakdown points, as well as restrictions on s, k for various applications that arise when we attempt to satisfy the incoherence requirements of Theorem 1. Table 1 summarizes the signal-corruption pairings for which APIS guarantees perfect recovery and their corresponding breakdown points essentially bounding how large s, k can be. Detailed derivations of these results are presented in the appendix and summarized below.

6.2 Cases with sparse corruptions

In this case $\mathscr{B} = \mathscr{S}_k^n$, the set of all k-sparse vectors. Calculating the incoherence constants then reduces to application-specific derivations which we sketch below.

Linear regression If the covariate matrix is X, then we get $\mu \le \max_{\substack{S \subset [n] \\ |S| = k}} \frac{\|X_S\|_{\text{op}}^2}{\lambda_{\min}(XX^\top)}$

, where $\|\cdot\|_{\text{op}}$ is the operator norm (see Appendix B for proofs). It turns out that this is satisfied in several natural settings. For example, if the covariates are Gaussian i.e. $\mathbf{x}^i \sim \mathcal{N}(\mathbf{0}, I_d)$ then $\mu < \frac{1}{3}$ (as required by Theorem 1) with high probability whenever $k < \frac{n}{154}$. We stress that our results do not require data points to be sampled from a standard Gaussian per se. The requirement $\mu < \frac{1}{3}$ is satisfied by other data distributions as well (see Appendix B).

Kernel regression For a Gram matrix G (calculated on covariates $\mathbf{x}^i, i \in [n]$ using a PSD kernel), we get $\mu \leq \frac{\Lambda_k^{\text{unif}}(G)}{\lambda_s(G)}$ where λ_s is the s^{th} largest eigenvalue of G and Λ_k^{unif} is the largest eigenvalue of any principal $k \times k$ submatrix of G. For the special case of RBF kernel, further calculations show that $\mu < \frac{1}{9}$ is satisfied, for instance, when covariates are sampled uniformly over the unit sphere and we have $s < \tilde{\boldsymbol{\Omega}}(\log n), k \leq \mathcal{O}\left(\sqrt{n}\right)$. Yet again, these settings (RBF kernel, unit sphere etc) are not essential, but merely sufficient conditions where APIS is guaranteed to succeed.

Signal transforms For a variety of signal transforms including Fourier, Hadamard, noiselet, we are assured $\mu < \frac{1}{9}$, as desired by Theorem 1, whenever $sk < \frac{n}{9}$. This can be realized in several ways, e.g., $s = \mathcal{O}(1), k = \mathcal{O}(n)$ or $s, k = \mathcal{O}\left(\sqrt{n}\right)$, etc. See Table 1 for a summary.

6.3 Cases with dense corruptions

Notably, APIS offers exact recovery even in certain cases where the corruption vector is completely dense, i.e., $\|\mathbf{b}^*\|_0 = n$. Note that the adversary is still allowed to be completely adaptive. One such case is when the signal is s-sparse in the Fourier or wavelet (Haar or Daubechies D4/D8) bases, and the corruptions are k-sparse in noiselet basis (Coifman et al. 2001). Since wavelets are known to represent natural signals well, this is a practically useful setting. Note that a vector $\mathbf{b}^*$ with a k-sparse noiselet representation even for $k = 1$ can be completely dense, i.e., $\|\mathbf{b}^*\|_0 = n$. APIS also supports dense Gaussian noise the responses as is discussed below.

6.4 Handing model misspecifications

In certain practical situations, the model outlined in Eq. (1) may not be satisfied. For instance, we could have $\mathbf{a}^* \notin \mathscr{A}$ if we have an image that is not entirely (but only approximately) sparse in the wavelet basis. Similarly, the unions $\mathscr{A}$ and $\mathscr{B}$ could fail to be incoherent (as is the case in the Fourier-Wavelet pair). In this section, we show how APIS can still offer non-trivial recovery in these settings.

Unmodelled error In this case we modify Eq. (1) to include an unmodelled error term.

$$\mathbf{y} = \tilde{\mathbf{a}} + \mathbf{b}^* + \mathbf{e}^*, \tag{2}$$

where $\tilde{\mathbf{a}} \in \mathscr{A}, \mathbf{b}^* \in \mathscr{B}$ and $\mathbf{a}^* = \tilde{\mathbf{a}} + \mathbf{e}^* \notin \mathscr{A}$. We make no assumptions on $\mathbf{e}^*$ belonging to any union of subspaces etc and allow it to be completely arbitrary. Section E in the appendix shows that if $\mu < \frac{1}{9}$ is satisfied, then for any $\epsilon > 0$, within $T \leq \mathcal{O}\left(\log((\left\|\mathbf{a}_2^*\right\| + \left\|\mathbf{b}_2^*\right\|)/\epsilon)\right)$ iterations, APIS guarantees a recovery error of

$$\left\|\mathbf{a}^T - \tilde{\mathbf{a}}\right\|_2 \leq \epsilon + \mathcal{O}\left(\max_{A \in \mathscr{A}} \|\Pi_A(\mathbf{e}^*)\|_2 + \max_{B \in \mathscr{B}} \|\Pi_B(\mathbf{e}^*)\|_2\right).$$

We now look at two applications of this result.

Simultaneous sparse corruptions and dense Gaussian noise. Consider linear regression where, apart from k adversarially corrupted points, all n points get Gaussian noise i.e. $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}^* + \mathbf{e}^*$, where $\mathbf{e}^* \sim \mathcal{N}(\mathbf{0}, \sigma^2 \cdot I_n)$. The above result shows that within $T = \mathcal{O}(\log n)$ iterations, APIS guarantees $\|\mathbf{w}^T - \mathbf{w}^*\|_2^2 \leq \mathcal{O}\left(\sigma^2\left(\frac{(d+k)\ln n}{n}\right)\right)$. As $n \to \infty$, the model error behaves as $\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \mathcal{O}(k \log n/n)$. This guarantees consistent recovery if $k \log n/n \to 0$ as $n \to \infty$. This is a sharper result than previous works (Bhatia et al. 2015; Mukhoty et al. 2019) that do not offer consistent estimation even if $k \log n/n \to 0$.

Compressible signals. Given an image $\mathbf{a}^*$ that is not itself wavelet-sparse, but still (s, ϵ) -approximately wavelet sparse i.e. there exists an image $\tilde{\mathbf{a}}$ that is s wavelet-sparse, and $\|\mathbf{a}^* - \tilde{\mathbf{a}}\|_2 \leq \epsilon \cdot \|\mathbf{a}^*\|_2$. In particular, $\tilde{\mathbf{a}}$ can be taken to be the best s wavelet-sparse approximation of $\mathbf{a}^*$. The above shows that even if $\mathbf{a}^*$ is subjected to adversarial corruptions, APIS offers a recovery of $\tilde{\mathbf{a}}$ to within $\mathcal{O}(\epsilon \cdot \|\mathbf{a}^*\|_2)$ error within $\mathcal{O}(\log(1/\epsilon))$ iterations.

Handling lack of incoherence: Pairs of bases that are not incoherent are well-known (Krahmer and Ward 2014; Zhou et al. 2016), the most famous example being the Fourier-Wavelet pair which can only assure $\mu \approx 1$ no matter how small s, k are. Thus, Theorem 1, if applied directly, would fail to offer a non-trivial recovery result if the signal is wavelet-sparse and corruptions are Fourier-sparse. However, in Sect. F in the appendix, we show that using local incoherence properties of these two bases [which are also well-studied e.g. Krahmer and Ward (2014)], APIS can be shown to continue to offer exact recovery if the signal is not just sparse in the wavelet domain, but also *anti-concentrated* as well i.e. it spreads its mass over its wavelet support elements (please see Sect. F in the appendix for details). For this setting, we show that the incoherence constant satisfies $\mu \leq \frac{k^4}{s} + \frac{sk^2}{n}$. Now $\mu < \frac{1}{9}$ can be ensured if, for example, $sk^2 \leq n/18$ (i.e. s, k are small compared to n which controls the second term) and $k^4 < s/18$ (i.e. $s \gg k$ which controls the first term). We note that some form of signal restriction, for example, signal anti-concentration, seems to be necessary since a *spike* signal having support over a single wavelet-basis element, can be irrevocably corrupted by an adaptive Fourier-sparse signal, given that the bases are not incoherent. Also, notice that this is yet another instance of APIS guaranteeing recovery when the corruptions are dense since a Fourier-sparse vector $\mathbf{b}^*$ can still have $\|\mathbf{b}^*\|_0 = \Omega(n)$.

6.5 Does APIS retain universality?

Kernel (ridge) regression with the RBF kernel is known to be a *universal* estimator (Micchelli et al. 2006). However, APIS requires the signal $\mathbf{a}^*$ to have a low-rank representation in terms of the top-s eigenvectors of the Gram matrix. As Table 1 indicates, for the RBF kernel, Theorem 1 allows s to be as large as $\mathcal{O}(\log n/\log\log n)$ for n points. Does this model retain universality despite this restriction? What sort of functions can $\mathbf{a}^*$ still approximate? We sketch an argument below that indicates an answer in the affirmative, along with a qualitative outline of functions that can be still described by this low-rank model.

Several results on random matrix approximation guarantee that if data covariates are chosen from nice distributions then, as the number of covariates $n \to \infty$, not only do the eigenvalues of the Gram matrix closely approximate those of the integral operator induced the PSD kernel (Minh et al. 2006; Rosasco et al. 2010), but the empirical operator also approaches the integral operator in the Hilbert-Schmidt norm. This assures us that eigenvectors of the Gram matrix closely approximate the eigenfunctions of the integral operator. For instance, Rosasco et al. (2010) offer an explicit two-way relation between the eigenvectors and the eigenfunctions. Now, in the uni-dimensional case ($d = 1$), for any $i \leq n$, the i^{th} largest eigenfunction of the integral operator for the RBF kernel is represented in terms of the i^{th}-order Hermite polynomial (Rasmussen and Williams 2006). The i^{th} Hermite polynomial is of degree i and Hermite polynomials form a universal basis (as they constitute an orthogonal polynomial sequence). In particular, the first s Hermite polynomials span all degree-s polynomial functions. Thus, even with the restriction on s, APIS does allow signals $\mathbf{a}^*$ that are (upto vanishing approximation errors) spanned by Hermite polynomials of order upto s. Now, APIS allows $s \leq \mathscr{O}(\log n / \log\log n)$ and as $n \to \infty$, $s \to \infty$ as well (albeit slowly). Thus, $\mathbf{a}^*$ can represent functions that are (upto approximation errors) arbitrarily high degree polynomials.

A similar argument holds for multi-dimensional spaces and *product kernels* e.g. RBF since, for such kernels, the eigenfunctions and eigenvalues in the multi-dimensional case are products of their uni-dimensional counterparts (Fasshauer 2011). Although it would be interesting to make the above arguments rigorous, they nevertheless indicate that APIS offers robust recovery for a model that is still universal in the limit. In Sect. 7, we will see that APIS offers excellent reconstruction for sinusoids, polynomials as well as their combinations over multi-dimensional spaces, even under adversarial corruptions.

7 Experiments

Extensive experiments were carried out comparing APIS to state-of-the-art competitor algorithms on three key robust regression tasks

1. Robust non-parametric regression to learn (multi-dimensional) sinusoidal and polynomial functions (see Figs. 3, 8, and Table 2) and robust Fourier transform (see Fig. 8)
2. Robust linear regression (see Fig. 4)
3. Image denoising on the benchmark Set12 images (see Fig. 5) with sparse adversarial salt-and-pepper corruption, as well as dense-checkerboard pattern corruptions (see Figs. 6, 7, and Table 3).

7.1 System configuration

Experiments for which runtimes for various algorithms were recorded were carried out on a 64-bit machine with Intel® Core™ i7-6500U CPU @ 2.50 GHz, 4 cores, 16 GB RAM and Ubuntu 16.04 OS, except for the Deep CNN model from (Zhang et al. 2017) which was trained on NVidia K80 GPUs (made available by the Kaggle platform for which the authors are grateful). All methods were implemented in Python, except those from (Gu et al. 2014; Cizek and Sadikoglu 2020), for which codes were made available by the authors themselves, in R and MATLAB, respectively. All figures e.g. Figs. 3, 4, 6 and 7 show actual predictions by various algorithms, and all results are reported over a single run of the algorithms.

7.2 Baselines and competitor algorithms

Below are described the state-of-the-art competitor algorithms chosen alongside APIS in various experiments.

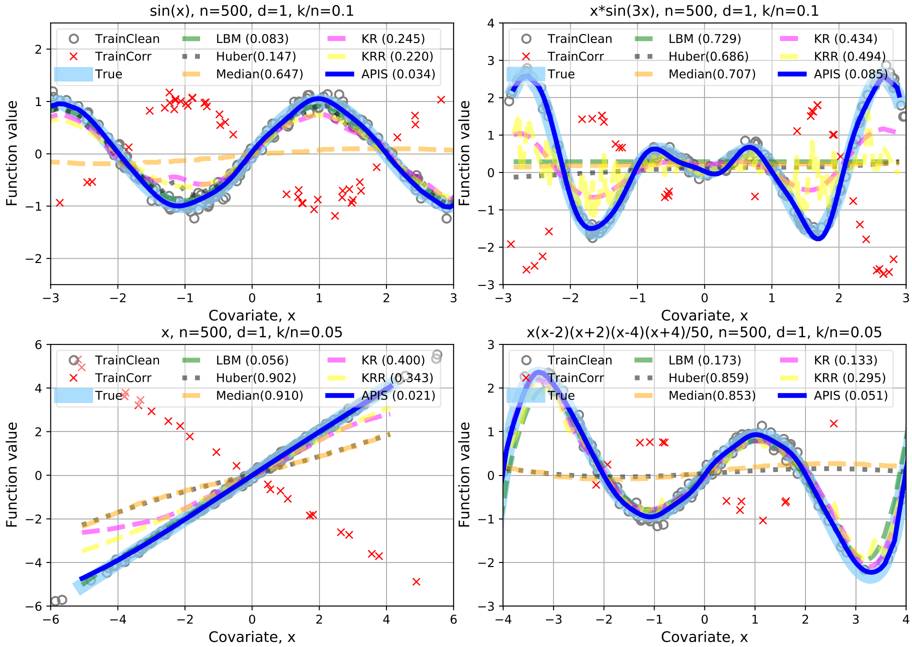
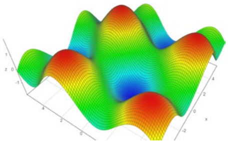


Fig. 3 Non-parametric kernel regression with RBF kernel. The four panels demonstrate the performance of various algorithms on four functions, a sinusoid $f(x) = \sin(x)$, a hybrid function $f(x) = x \cdot \sin(3x)$, a linear function $f(x) = x$ and a quintic polynomial $f(x) = x(x-2)(x+2)(x-4)(x+4)/50$. The headings of the panels show the function being learnt, the number of training points, and the fraction of corrupted training points. The RMSE values offered by various algorithms is mentioned in parentheses against method names in the legends. The true function curve is plotted using a thick light blue curve. 500 training points were sampled from $\mathcal{N}(0, 2^2)$. A fraction $\left(\frac{k}{n}\right)$ these points were sampled to be subjected to adversarial corruption with probability proportional to the magnitude of their function value i.e. $|f(x)|$. Responses were modified for the corrupted points by setting $y = -f(x)$ i.e. flipping the sign of the response but retaining the magnitude. Gaussian noise sampled from $\mathcal{N}(0, 0.1^2)$ was added to all points (even *clean* ones). In the figures, corrupted points are depicted using a red cross and clean points using an empty gray circle. Hyperparameter tuning was done for all methods as described in the main text. 1000 test points were sampled from $\mathcal{N}(0, 1.5^2)$ for estimating the RMSE for various algorithms. No corruption or Gaussian noise was added to test responses. In all cases, APIS offers the best test RMSE that is 2 to 5× smaller than the next best method

Table 2 Multidimensional non-parametric regression with RBF kernel

d	n	LBM	Huber	Median	KR	KRR	APIS	
2	500	0.399	0.99	0.965	0.5	0.445	**0.029**	
5	2000	KR	1.552	1.526	1.089	0.710	**0.474**	
7	3000	KR	1.934	1.876	1.564	1.352	**0.803**	

The problem settings (data, corruption model, Gaussian noise) are identical to Fig. 3 except that the covariates are multi-dimensional now. The first two columns report the dimensionality of the covariates and the number of training points. The rest of the columns report test RMSE values for various algorithms. Bold values indicate the best performance in each row i.e. for each dimensionality. The true function being learnt was a sum of uni-dimensional sinusoids i.e. $f(\mathbf{x}) = \sum_{i=1}^{d} \sin(\mathbf{x}_i)$. The figure on the right depicts this function for the $d = 2$ case. Figure courtesy https://academo.org

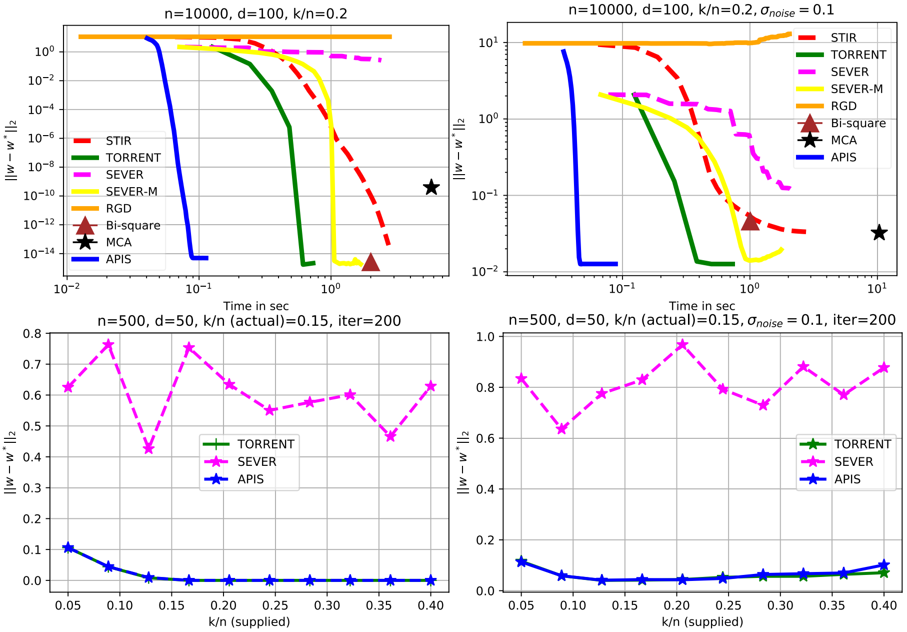


Fig. 4 Linear regression. The figures demonstrate convergence of the algorithms with respect to time (top) and sensitivity with respect to hyper-parameter k (bottom). The headings of the panels show the number of training points, the fraction of corrupted training points, and the variance of the dense Gaussian noise (if present). For the top row, 10000 training points were generated as $\mathbf{x}^i \sim \mathcal{N}(\mathbf{0}, I_d)$ for $d = 100$. A true model and an *adversarial* model were both chosen as $\mathbf{w}^*, \mathbf{w}^{\mathrm{a}} \sim \mathcal{N}(\mathbf{0}, I_d)$. For clean points, $y = \mathbf{x}^\top \mathbf{w}^*$, for 20% corrupted points $y = \mathbf{x}^\top \mathbf{w}^{\mathrm{a}}$. The left column has no Gaussian noise, while the plots on the right column have Gaussian noise $\mathcal{N}(0, 0.1^2)$, added to all points (even *clean* ones). The timed convergence plots show that in absence of Gaussian noise, most methods offer recovery of $\mathbf{w}^*$ to machine precision level although APIS offers fastest recovery (RGD, SEVER offer slow recovery). The bottom row shows APIS offer stable recovery in a wide range of settings

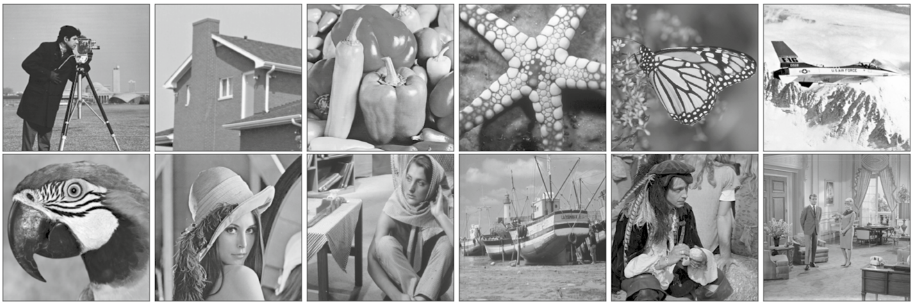


Fig. 5 The Set-12 Dataset (Fan et al. 2019; Zhang et al. 2017) is a collection of 12 popular gray scale images. All images were scaled to 512 × 512 pixels for our experiments

Robust non-parametric kernel regression: Based on recommendations of the recent survey by Cizek and Sadikoglu (2020), the widely studied Huber and median estimators were chosen as baselines. The Nadaraya-Watson (kernel regression) and kernel ridge regression

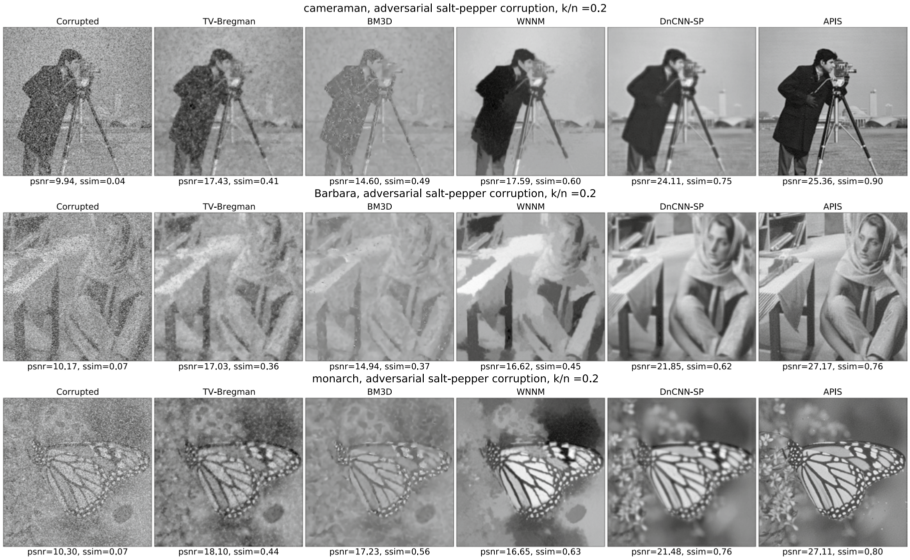


Fig. 6 Image denoising with sparse adversarial corruptions. For each image, sparse adversarial salt-and-pepper corruptions were added to 20% of the pixels. Corruptions were added to these pixels by setting lighter pixels i.e. those with grayscale value > 127 to completely black i.e., the grayscale value of 0, and setting darker pixels i.e. those with grayscale value < 127 to completely white i.e., the grayscale value of 255. The first column shows the image obtained as a result of these corruptions. The subsequent columns show the image recovered by various algorithms. Below each image is shown the PSNR and SSIM values for that image. APIS not only offers visually better recovery, but also the best PSNR and SSIM values, often significantly better than the next best method. In particular, APISis able to preserve the minute checkerboard pattern on the tablecloth and the fine stripes on the scarf in the Barbara image (second row), and the fine details of the flowers and the extremely thin antennae of the butterfly in the Monarch image (third row), whereas all the other methods lose these fine details.

estimators were also chosen as baselines. In addition, the recently proposed LBM method (Du et al. 2018) for non-parametric regression was chosen as a competitor.

Robust linear regression: Recent state-of-the-art algorithms were chosen as competitors including SEVER (Diakonikolas et al. 2019), RGD (Prasad et al. 2018), STIR (Mukhoty et al. 2019), TORRENT (Bhatia et al. 2015) and the constrained L_1 minimization based morphological component analysis (MCA) approach of McCoy and Tropp (2014). The classical robust M-estimator based on Tukey's bi-square loss was also chosen as a baseline. We note that the SEVER algorithm as described in Diakonikolas et al. (2019) eliminates corrupted points sluggishly and is slow. A modification was done to offer the algorithm a handicap by revealing the true number of corrupted points to speed up the algorithm. Results are reported for both SEVER as well as this modified variant SEVER-M.

Image denoising: For image experiments, comparisons are reported against a wide variety of standard state-of-the-art image denoising methods, including local methods, as well as methods based on total-variation and deep learning. We briefly summarize these here. In image denoising, non-local self-similarity (NSS) based methods exploit the fact that image patches often repeat themselves. In the BM3D method (Dabov et al. 2007), similar patches found using block-matching (BM) are stacked in 3D and taken to a transform domain for noise attenuation. A weighted average of the cleaned patches is then used to estimate each pixel of the output image. Another NSS-based method is WNNM (Gu et al. 2014) that

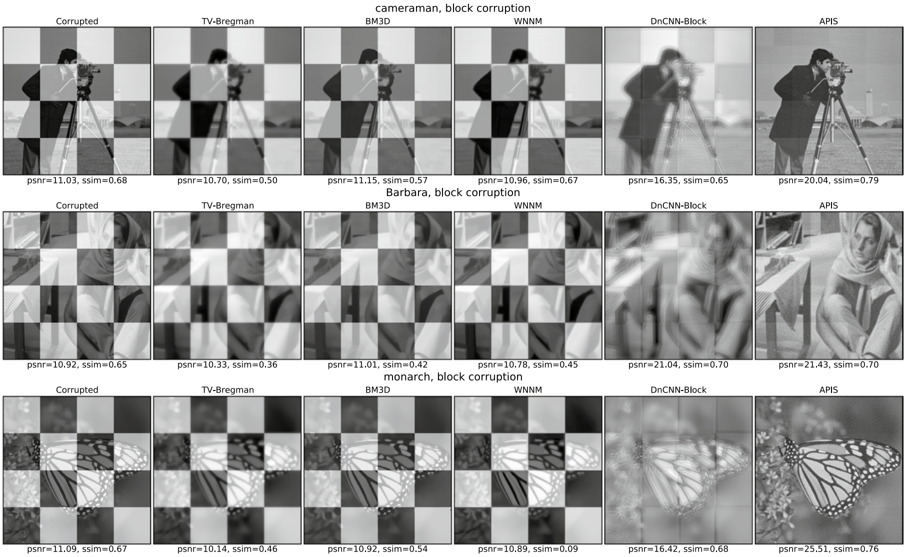


Fig. 7 Image denoising with dense corruptions. For each image, dense corruptions were introduced by superimposing a dense checkerboard pattern on the entire image, thus corrupting all pixels. The first column shows the image obtained as a result of these corruptions. The subsequent columns show the image recovered by various algorithms. Below each image is shown the PSNR and SSIM values for that image. Similar to the observation made in the caption for Fig. 6, APIS offers superior recovery and preserves fine details in both images. In contrast, TV-Bregman, BM3D, and WNNM fail to offer any reasonable recovery in either of the examples, while DnCNN-Block retains much of the block corruption

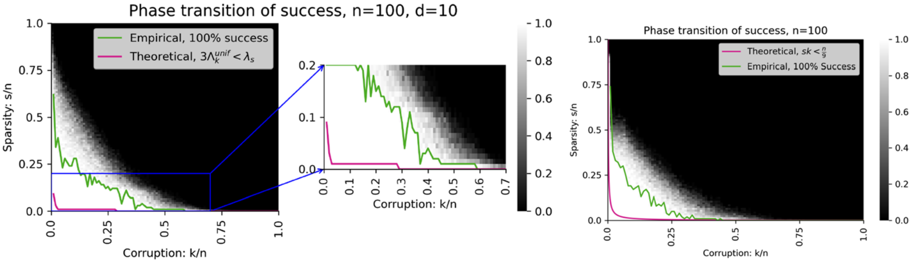


Fig. 8 Phase transition. This figure demonstrates the phase transition behavior of APIS on robust non-parametric regression with RBF kernel (left) and robust Fourier transform (right) tasks, by considering various sparsity and corruption levels. For each combination of $(s, k) \in [n] \times [n]$, a synthetic dataset $(\mathbf{a}^*, \mathbf{b}^*)$ was created randomly and APIS was executed to obtain its estimate $\hat{\mathbf{a}}$ of the signal. An approximation error of $\|\mathbf{a}^* - \hat{\mathbf{a}}\|_2 < 10^{-4} \cdot \|\mathbf{a}^*\|_2$ was considered a "success". The experiment was repeated 50 times for each (s, k) pair giving us a success likelihood $p_{(s,k)} \in [0, 1]$ for each pair (s, k). The figures plot these success likelihood values with a white pixel indicating $p_{(s,k)} = 1$ i.e. all 50 experiments succeeded for that value of (s, k), a black pixel indicating $p_{(s,k)} = 0$ i.e. none of the 50 experiments succeeded for that value of (s, k) and shades of gray indicating intermediate likelihood values in the range (0, 1). For robust non-parametric regression with RBF kernel, a Gram matrix was first generated by sampling 100 data points, from the surface of the unit sphere in 10 dimensions, i.e. S^9. Next, the signal was generated as a random vector in the span of the top s eigen-vectors of the Gram matrix. For robust Fourier transform, first a random s-sparse vector $\boldsymbol{\alpha}^*$ was generated and then the signal was created as $\mathbf{a}^* = F\boldsymbol{\alpha}^*$ where F is the matrix corresponding to the Fourier transform. In both cases, adversarial corruption was introduced to $\mathbf{a}^*$ as done in Fig.3. In both figures, the recovery limit guaranteed by Theorem 1 is outlined in magenta whereas the empirical 100% success region is outlined in green. Additionally, the figure on the left for robust non-parametric regression with RBF kernel is zoomed in for easy viewing

Table 3 PSNR, SSIM metric values and recovery times for various methods, averaged over all 12 images of the Set12 dataset (see Fig. 5)

	Sparse salt-and-pepper corruptions			Dense checkerboard corruptions		
	PSNR	SSIM	Time (sec)	PSNR	SSIM	Time (sec)
APIS	**28.92**	**0.856**	1.85	**22.89**	**0.787**	82.15
DnCNN-SP/Block	23.04	0.714	2.14	20.12	0.69	2.03
TV-Bregman	17.24	0.407	**0.09**	10.58	0.465	**0.09**
WNNM	16.24	0.523	642.51	10.40	0.486	527.36
BM3D	15.54	0.465	206.58	10.90	0.459	232.69
Noisy image	10.21	0.050	–	10.99	0.677	–

Bold values indicate the best performance in each column i.e. for each metric viz. PSNR, SSIM and running time in seconds. For sparse adversarial salt-and-pepper corruptions, APIS improves PSNR by 25% and SSIM by 20% over the next best method that happens to be DnCNN-SP. For dense checkerboard corruptions, APISimproves PSNR by 13% and SSIM by 14% with respect to the next best method that happens to be DnCNN-Block. It is notable though that the training time of DnCNN models i.e. DnCNN-SP and DnCNN-Block is quite large (around 6 hours each on NVidia K80 GPUs), and is not shown in the table

utilizes the low-rank structure of similar patches. Weighted nuclear norm minimization is used as a convex relaxation of low rank approximation. Given a set of image patches $\mathbf{Y}$, a low rank estimate $\hat{\mathbf{X}}$ is found such that:

$$\hat{\mathbf{X}} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{X}\|_F + \|\mathbf{X}\|_{\mathbf{w},*}$$

where $\|\mathbf{X}\|_{\mathbf{w},*} = \sum_i w_i \sigma_i(\mathbf{X})$ is the weighted nuclear norm of $\mathbf{X}$. Comparisons are also reported against the total variation regularization based method TV-Bregman [solved using the split-Bregman method (Getreuer 2012)] that uses local smoothness property of natural images. The superiority of the WNNM method, among non-deep methods, was reported by Bouwmans et al. (2018). Comparisons are also reported with a deep CNN-based technique (DnCNN) that performs Gaussian denoising based on residual learning. This network has 17 CNN layers with the ReLU activation function and intermediate layers additionally using batch normalization. However, to adapt the network to adversarial corruption settings, the architecture was retrained separately on salt-paper noise (DnCNN-SP) and block noise (DnCNN-Block) using the 400 image dataset used by the authors (Zhang et al. 2017). Recent studies (Fan et al. 2019) have demonstrated the effectiveness of DnCNN over filtering methods on image denoising tasks.

7.3 Performance metrics

Robust non-parametric regression: For the results in Fig. 3 and Table 2, the root mean squared error (RMSE) on test points was used as a performance metric. It is notable that these experiments were conducted on synthetic data and in order to enable an unbiased evaluation of the recovery of the function being learnt (e.g. sinusoid, polynomial etc), test points were not subjected to adversarial corruptions or Gaussian noise.

Robust linear regression: For the results in Fig. 4, again synthetic settings were used, and performance was measured in terms of model recovery error i.e. the L_2 distance of the model recovered by an algorithm and the true model. Note that directly measuring model recovery error was not performed in non-parametric settings since in those settings, the function being learnt by various algorithms is, in general, not expressible in terms of a compact parameter.

Image denoising: Performance was measured in experiments reported in Figs. 6, 7 and Table 3 using standard performance measures, namely peak signal to noise ratio (PSNR) and structural similarity index measure (SSIM). The mean squared error (MSE) of a noisy $n \times m$ image N w.r.t. the original image I is defined as $\text{MSE} = \frac{1}{mn}\|N - I\|_2^2$. The PSNR of the noisy image N is then defined to be

$$\text{PSNR} = 10 \log_{10} \frac{\max(I)^2}{\text{MSE}}$$

where $\max(I)$ defined as the maximum value of a pixel in the image I. SSIM on the other hand, tries to combine luminance, contrast and structural similarity between two images, but at a patch level rather than the global level. Given two image patches X, Y, their SSIM is computed as

$$\text{SSIM} = \frac{(2\mu_X\mu_Y + c_1)(2\sigma_{XY} + c_2)}{(\mu_X^2 + \mu_Y^2 + c_1)(\sigma_X^2 + \sigma_Y^2 + c_2)}$$

where $\mu_X, \sigma_X, \sigma_{XY}$ denote the mean, variance and co-variance respectively, and c_1, c_2 denote constants. A complete description of this metric can be found in (Wang et al. 2004).

7.4 Hyperparameter tuning

Hyperparameter tuning becomes non-trivial with corrupted data since responses in a given *validation* set are also expected to be corrupted. Specifically, let the observed and predicted responses in a validation set be denoted by $\mathbf{y}$ and $\hat{\mathbf{y}}$. Since $\mathbf{y}$ itself contains corruptions, the residual $\|\mathbf{r}\|_2, \mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, is not a reliable indicator of performance. Instead, APIS proposes using $\left\|\mathbf{r} - \Pi_{\mathscr{B}}(\mathbf{r})\right\|_2$ as an indicator. Note that if $\hat{\mathbf{y}} \approx \mathbf{a}^*$, i.e., excellent performance, then $\mathbf{r} \approx \mathbf{b}^*$ and indeed we have $\left\|\mathbf{r} - \Pi_{\mathscr{B}}(\mathbf{r})\right\|_2 \to 0$. For sparse corruptions, the above involves simply rejecting the top few validation points with the highest prediction error. Under this strategy, algorithms need to have an (estimate of) k to construct the projection operator $\Pi_{\mathscr{B}}(\cdot)$. Methods that do not have k as a hyperparameter were offered the true value of k to allow them to perform hyperparameter tuning. APIS was not offered this handicap as it did have k as a tunable hyperparameter.

Hyperparameter ranges: All hyperparameters that were tuned for various algorithms are reported using Python syntax to specify ranges. For non-parametric regression, all algorithms were offered the RBF kernel whose bandwidth was tuned in the range `logspace(-2,2,5)` separately for all algorithms (except for the Huber and median methods, where code from the authors of (Cizek and Sadikoglu 2020) was used which itself performed all hyperparameter tuning). For LBM, the number of bins M is shown to have an optimal value $M \approx \sqrt{n}$ by Du et al. (2018) and was thus tuned in the range `sqrt(n)*linspace(0.5,1.5,10)`. For APIS, both s, k were tuned in the range `n*linspace(0.05,0.2,10)`. For linear regression, all algorithms except STIR required k to be tuned which was done in the range `n*linspace(0.01,0.2,20)`. For STIR, an η parameter (that needs to be strictly greater than 1) was tuned in the range `np.linspace(1.01,3.01,21)`. For the RGD algorithm, two parameters

η, δ were respectively tuned in the ranges `np.linspace(0.05,0.2,20)` and `np.linspace(0.1,0.9,9)`.

7.5 Results

Figures 3 and 4 summarize the results of an empirical comparison of APIS with competitor methods on non-parametric regression and linear regression, respectively. APIS offers far superior convergence speeds on linear regression tasks (and vanishing model recovery error comparable to that offered by several other algorithms) and the best RMSE values on non-parametric regression tasks. Table 2 reports results on non-parametric regression on multidimensional data. For $d > 2$, LBM results were identical to that of its base estimator KR, as number of bins $M \approx (\sqrt{n})^d$ (at $\sqrt{n}$ bins per dimension), far exceeded number of train points resulting in bins having a single data point.

For image denoising experiments, gray-scale images of popular Set12 dataset (Fan et al. 2019; Zhang et al. 2017) (shown in Fig. 5) were used. Images were subjected to sparse adversarial salt-and-pepper corruptions, as well as dense checkerboard corruptions. Figures 6 and 7 present visualizations of the performance of APIS and state-of-the-art denoising methods. APIS demonstrates superior recovery both in terms of subjective visual perception, as well as in terms of PSNR and SSIM. The PSNR, SSIM metrics, and prediction times offered by all methods averaged over the 12 images in the Set12 dataset, are reported in Table 3. For sparse adversarial salt-and-pepper corruptions, APISimproves PSNR by 25% and SSIM by 20% over the next best model DnCNN-SP. For dense checkerboard corruptions, APISimproves PSNR by 20% and SSIM by 14% to the next best DnCNN-Block.

Appendix

A A generic recovery guarantee for APIS: a proof of Theorem 1

In this section, we will prove Theorem 1. We will present the proof in two parts, presenting the main proof ideas in Lemma 1 with the special case of $P = 1$ (the so-called *known signal support* case (Chen and De 2020)) where the union $\mathscr{A}$ consists of a single subspace A. Recall that we denote using P (resp. Q), the number of subspaces in the union $\mathscr{A} = \bigcup_{i=1}^{P} A_i$ in which the uncorrupted signal $\mathbf{a}^*$ resides (resp the union $\mathscr{B} = \bigcup_{j=1}^{Q} B_j$ in which the corruption $\mathbf{b}^*$ resides). We also recall that the known signal support case does capture linear regression and low-rank kernel ridge regression. Note however, that even in the known signal support case, we may still have $Q > 1$ i.e. $\mathscr{B}$ may still be a general non-trivial union of subspaces e.g. the set of k-sparse vectors which has $Q = \begin{pmatrix} n \\ k \end{pmatrix}$. We will then extend the proof to the general case in Lemma 2 where both $P, Q \geq 1$. We reproduce the APIS algorithm below for ease of reading.

Algorithm 2 The APIS Algorithmic Framework (reproduced from Algorithm 1)

```
Input: Corrupted responses $\mathbf{y}$, Projection operators $\Pi_{\mathscr{A}}(\cdot), \Pi_{\mathscr{B}}(\cdot)$ that project onto $\mathscr{A}, \mathscr{B}$
Output: An estimate $\hat{\mathbf{a}}$ of the clean responses
Initialize $\mathbf{a}^0 \leftarrow \mathbf{0}$ and $t \leftarrow 0$
for $T = 1, 2, \ldots, T-1$ do
    $\mathbf{b}^{t+1} \leftarrow \Pi_{\mathscr{B}}(\mathbf{y} - \mathbf{a}^t)$          //let $B^{t+1} \in \mathscr{B}$ be a subspace that contains $\mathbf{b}^{t+1}$
    $\mathbf{a}^{t+1} \leftarrow \Pi_{\mathscr{A}}(\mathbf{y} - \mathbf{b}^{t+1})$      //let $A^{t+1} \in \mathscr{B}$ be a subspace that contains $\mathbf{a}^{t+1}$
    $t \leftarrow t + 1$
end for
return $\mathbf{a}^T$
```

A.1 Convergence analysis for $P = 1$ i.e. $\mathscr{A} = A$ but still $Q \geq 1$

We now present the proof in the case of known signal support.

Lemma 1 *Suppose we obtain data as described in Eq.* (1) *where the two unions $\mathscr{A}, \mathscr{B}$ are μ-incoherent with $\mu < \frac{1}{3}$ and in addition, the union $\mathscr{A}$ contains a single subspace (the "known signal support" model). Then, for any $\epsilon > 0$ within $T = \mathscr{O}\left(\log \frac{\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2}{\epsilon}\right)$ iterations,* APIS *offers* $\left\|\mathbf{a}^T - \mathbf{a}^*\right\|_2 \leq \epsilon$.

Proof To simplify notation, we denote $\mathbf{a}^t =: \mathbf{a}, \mathbf{b}^t =: \mathbf{b}, \mathbf{a}^{t+1} =: \mathbf{a}^+, \mathbf{b}^{t+1} =: \mathbf{b}^+, B^{t+1} =: B^+$ (please refer to Algorithm 2 for notation). Let $\mathfrak{Q} := B^+ \cap B^*$ denote the meet of the two subspaces, as well as denote the symmetric difference subspaces $\mathfrak{P} := B^+ \cap (B^*)^\perp$ and $\mathfrak{R} = B^* \cap (B^+)^\perp$ (recall that $A \ni \mathbf{a}^*, B^* \ni \mathbf{b}^*$).

Denote $\mathbf{p} = \Pi_A(\mathbf{b}^* - \mathbf{b})$ and $\mathbf{p}^+ = \Pi_A(\mathbf{b}^* - \mathbf{b}^+)$. In this case we have $\mathbf{a}^+ = \Pi_A(\mathbf{y} - \mathbf{b}^+) = \mathbf{a}^* + \Pi_A(\mathbf{b}^* - \mathbf{b}^+)$ (since $\mathbf{a}^* \in A$ and orthonormal projections are idempotent) and thus, $\|\mathbf{a}^+ - \mathbf{a}^*\|_2 = \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2 = \|\mathbf{p}^+\|_2$. We will show below that $\|\mathbf{p}^+\|_2 \leq 3\mu \cdot \|\mathbf{p}\|_2$ which will establish, if $\mu < \frac{1}{3}$, a linear rate of convergence since we will have $\|\mathbf{a}^+ - \mathbf{a}^*\|_2 = \|\mathbf{p}^+\|_2 \leq 3\mu \cdot \|\mathbf{p}\|_2 = 3\mu \cdot \|\mathbf{a} - \mathbf{a}^*\|_2$.

We have

$$\mathbf{b}^+ = \Pi_{B^+}(\mathbf{a}^* + \mathbf{b}^* - \mathbf{a}) = \Pi_{B^+}(\mathbf{b}^* - \Pi_A(\mathbf{b}^* - \mathbf{b})) = \Pi_{B^+}(\mathbf{b}^* - \mathbf{p}),$$

and thus $\mathbf{b}^* - \mathbf{b}^+ = \mathbf{b}^* - \Pi_{B^+}(\mathbf{b}^* - \mathbf{p}) = \Pi_{\mathfrak{R}}(\mathbf{b}^*) + \Pi_{B^+}(\mathbf{p})$. This gives us, by an application of the triangle inequality,

$$\left\|\mathbf{p}^+\right\|_2 = \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2 \leq \left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 + \left\|\Pi_A(\Pi_{B^+}(\mathbf{p}))\right\|_2$$

Now, the projection step assures us that projecting onto B^+ was the best option out of all the subspaces in $\mathscr{B}$ and thus, if we denote $\mathbf{z} = \mathbf{b}^* - \mathbf{p}$, then we have, for any subspace $B \in \mathscr{B}$,

$$\left\|\Pi_{B^+}(\mathbf{z}) - \mathbf{z}\right\|_2^2 \leq \left\|\Pi_B(\mathbf{z}) - \mathbf{z}\right\|_2^2.$$

Now, $\Pi_B(\mathbf{z}) - \mathbf{z} = \Pi_B^\perp(\mathbf{z})$. Using this, in particular, we have, setting $B = B^*$

$$\left\|\Pi_{B^+}^\perp(\mathbf{z})\right\|_2^2 \leq \left\|\Pi_{B^*}^\perp(\mathbf{z})\right\|_2^2$$

Canceling components in the subspace $(B^+)^\perp \cap (B^*)^\perp$, as well as those in the subspace $\mathfrak{Q}$ gives us

$$\left\|\Pi_{\mathfrak{R}}(\mathbf{z})\right\|_2^2 \le \left\|\Pi_{\mathfrak{P}}(\mathbf{z})\right\|_2^2 = \left\|\Pi_{\mathfrak{P}}(\mathbf{p})\right\|_2^2$$

since $\Pi_{B^*}^{\perp}(\mathbf{b}^*) = \mathbf{0}$. Now, $\Pi_{\mathfrak{R}}(\mathbf{z}) = \Pi_{\mathfrak{R}}(\mathbf{b}^*) - \Pi_{\mathfrak{R}}(\mathbf{p})$ since projections are linear operators. Applying the triangle inequality gives us $\left\|\Pi_{\mathfrak{R}}(\mathbf{z})\right\|_2 \ge \left\|\Pi_{\mathfrak{R}}(\mathbf{b}^*)\right\|_2 - \left\|\Pi_{\mathfrak{R}}(\mathbf{p})\right\|_2$. This gives us

$$\begin{aligned}\left\|\Pi_{\mathfrak{R}}(\mathbf{b}^*)\right\|_2 &\le \left\|\Pi_{\mathfrak{P}}(\mathbf{p})\right\|_2 + \left\|\Pi_{\mathfrak{R}}(\mathbf{p})\right\|_2\\ &\le \left\|\Pi_{B^+}(\mathbf{p})\right\|_2 + \left\|\Pi_B^*(\mathbf{p})\right\|_2,\end{aligned}$$

where the second step follows since orthonormal projections are always non-expansive. Applying incoherence results now tells us that, since $\mathbf{p} \in \mathscr{A}$, we have

$$\left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 \le \sqrt{\mu}\cdot\left\|\Pi_{\mathfrak{R}}(\mathbf{b}^*)\right\|_2 = \sqrt{\mu}(\left\|\Pi_{B^+}(\mathbf{p})\right\|_2 + \left\|\Pi_B^*(\mathbf{p})\right\|_2) \le 2\mu\cdot\|\mathbf{p}\|_2$$

This gives us, upon applying contraction due to incoherence,

$$\begin{aligned}\left\|\mathbf{p}^+\right\|_2 &\le \left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 + \left\|\Pi_A(\Pi_{B^+}(\mathbf{p}))\right\|_2\\ &\le 3\mu\cdot\|\mathbf{p}\|_2\end{aligned}$$

Thus, in the known signal support case, APIS offers a linear rate of convergence whenever $\mu < \frac{1}{3}$. Now, APIS initializes $\mathbf{a}^0 = \mathbf{0}$ which means that initially, we have

$$\mathbf{p}^1 = \Pi_A(\mathbf{b}^* - \mathbf{b}^1) = \Pi_A(\mathbf{b}^* - \Pi_{B^+}(\mathbf{a}^* + \mathbf{b}^*))$$

and thus, $\left\|\mathbf{p}^1\right\|_2 \le \|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2$ since projections are always non-expansive. The linear rate of convergence implies that within $T = \mathscr{O}\left(\log\frac{\|\mathbf{a}^*\|_2+\|\mathbf{b}^*\|_2}{\epsilon}\right)$ iterations, we will have $\left\|\mathbf{p}^T\right\|_2 \le \epsilon$. Using our earlier observation $\left\|\mathbf{a}^T - \mathbf{a}^*\right\|_2 = \left\|\mathbf{p}^T\right\|_2$ then finishes the proof.

A.2 Convergence analysis for general case i.e both $P, Q \ge 1$

We now present the proof in the general case.

Lemma 2 *Suppose we obtain data as described in Eq.* (1) *where the two unions* $\mathscr{A}, \mathscr{B}$ *are* μ *-incoherent with* $\mu < \frac{1}{9}$ *(we allow both* $P, Q > 1$ *in this case). Then, for any* $\epsilon > 0$ *within* $T = \mathscr{O}\left(\log\frac{\|\mathbf{a}^*\|_2+\|\mathbf{b}^*\|_2}{\epsilon}\right)$ *iterations,* APIS *offers* $\left\|\mathbf{a}^T - \mathbf{a}^*\right\|_2 \le \epsilon$.

Proof As before, to simplify notation, we denote $\mathbf{a}^t{=}:\mathbf{a}, \mathbf{b}^t{=}:\mathbf{b}, \mathbf{a}^{t+1}{=}:\mathbf{a}^+, \mathbf{b}^{t+1}{=}:\mathbf{b}^+, A^{t+1}{=}:A^+, B^{t+1}{=}:B^+$. Let $\mathfrak{Q}:=B^+ \cap B^*$ denote the meet of the two subspaces, as well as denote the symmetric difference subspaces $\mathfrak{P}:=B^+ \cap (B^*)^\perp$ and $\mathfrak{R} = B^* \cap (B^+)^\perp$ (recall that $B^* \ni \mathbf{b}^*$). Also let $\mathfrak{M}:=A^+ \cap A^*$ denote the meet of the two subspaces, as well as denote the symmetric difference subspaces $\mathfrak{L}:=A^+ \cap (A^*)^\perp$ and $\mathfrak{N} = A^* \cap (A^+)^\perp$ (recall that $A^* \ni \mathbf{a}^*$). We also introduce the additional notation $p := \max_{A \in \mathscr{A}} \left\|\Pi_A(\mathbf{b}^* - \mathbf{b})\right\|_2, p^+ := \max_{A \in \mathscr{A}} \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2$ as well as $q := \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{a}^* - \mathbf{a})\right\|_2, q^+ := \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{a}^* - \mathbf{a}^+)\right\|_2$.

Note that the update step gives us $\mathbf{a}^+ = \Pi_{A^+}(\mathbf{a}^* + \mathbf{b}^* - \mathbf{b}^+)$ which gives us

$$\mathbf{a}^+ - \mathbf{a}^* = \Pi_{\mathfrak{R}}(\mathbf{a}^*) + \Pi_{A^+}(\mathbf{b}^+ - \mathbf{b}^*),$$

i.e.

$$\left\|\mathbf{a}^+ - \mathbf{a}^*\right\|_2 \leq \left\|\Pi_{\mathfrak{R}}(\mathbf{a}^*)\right\|_2 + \left\|\Pi_{A^+}(\mathbf{b}^+ - \mathbf{b}^*)\right\| \leq \left\|\Pi_{\mathfrak{R}}(\mathbf{a}^*)\right\|_2 + p^+,$$

by applying the triangle inequality. A similar analysis of the projection step, as we did to analyze the special case for $P = 1$, then gives us

$$\left\|\Pi_{\mathfrak{R}}(\mathbf{a}^*)\right\|_2 \leq \left\|\Pi_{A^+}(\mathbf{b}^+ - \mathbf{b}^*)\right\|_2 + \left\|\Pi_A^*(\mathbf{b}^+ - \mathbf{b}^*)\right\|_2 \leq 2p^+,$$

giving us

$$\left\|\mathbf{a}^+ - \mathbf{a}^*\right\|_2 \leq 3p^+.$$

We now show that we have $p^+ \leq 9\mu \cdot p$ i.e. the quantity p decreases at a linear rate whenever $\mu < \frac{1}{9}$. Since the update step gives us $\mathbf{b}^+ = \Pi_{A^+}(\mathbf{b}^* + \mathbf{a}^* - \mathbf{a})$, an analysis similar to the one done for the special case for $P = 1$ gives us, for any $A \in \mathscr{A}$,

$$\begin{aligned}\left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2 &\leq \left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 + \left\|\Pi_A(\Pi_{B^+}(\mathbf{a} - \mathbf{a}^*))\right\|_2 \\ &\leq \sqrt{\mu}(\left\|\Pi_{\mathfrak{R}}(\mathbf{b}^*)\right\|_2 + q).\end{aligned}$$

Going as before also gives us

$$\left\|\Pi_{\mathfrak{R}}(\mathbf{b}^*)\right\|_2 \leq \left\|\Pi_{B^+}(\mathbf{a} - \mathbf{a}^*)\right\|_2 + \left\|\Pi_B^*(\mathbf{a} - \mathbf{a}^*)\right\|_2 \leq 2q,$$

and thus, putting the results together gives us

$$p \leq \max_{A \in \mathscr{A}} \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2 \leq 3\sqrt{\mu} \cdot q,$$

or considering this result for a different iterate, we get $p^+ \leq 3\sqrt{\mu} \cdot q^+$. Since the updates w.r.t $\mathbf{a}$ and $\mathbf{b}$ are absolutely symmetric, a similar analysis to the above also gives us $q^+ \leq 3\sqrt{\mu} \cdot p$ and consequently, $p^+ \leq 9\mu \cdot p$. Thus, APIS offers a linear rate of convergence in the general case whenever $\mu < \frac{1}{9}$. A similar analysis as before confirms $p^1 = \max_{A \in \mathscr{A}} \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^1)\right\|_2 = \mathscr{O}\left(\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2\right)$ and that within $T = \mathscr{O}\left(\log \frac{\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2}{\epsilon}\right)$ iterations, we would have $p^T \leq \frac{\epsilon}{3}$. Since we already saw above that $\left\|\mathbf{a}^T - \mathbf{a}^*\right\|_2 \leq 3p^T$, this confirms the upper bound on the number of iterations required.

B Robust linear regression using APIS

We recall that in this case, we have known signal support i.e. $P = 1$ with $\mathscr{A} = A$ being the row span of the covariate matrix $X \in \mathbb{R}^{d \times n}$ and $\mathscr{B}$ being the union of subspaces of k-sparse vectors.

Lemma 3 *If the corruption vectors are (adaptive adversarial) k-sparse vectors and the covariates* $\mathbf{x}^i \in \mathbb{R}^d, i \in [n]$ *are sampled i.i.d. from a standard Gaussian i.e.* $\mathbf{x}^i \sim \mathcal{N}(\mathbf{0}, I_d)$ *and* $n = \Omega(d)$*, then with probability at least* $1 - \frac{1}{d^2}$*,* APIS *offers exact recovery at a linear rate if* $k < \frac{n}{154}$*. Moreover, the projection operation* $\Pi_{\mathcal{A}}(\cdot)$ *can be performed in* $\mathcal{O}(nd)$ *time in this case.*

Proof Let V denote the d right singular vectors of the covariate matrix $X \in \mathbb{R}^{d\times n}$. Then the projection operator Π_A is given as $\Pi_A(\mathbf{z}) = VV^\top \mathbf{z}$ where $VV^\top = X^\top(XX^\top)^\dagger X$. Note that this can also be accomplished by simply solving a least squares problem which can be done in $\mathcal{O}(nd)$ time using various (conjugate, stochastic) gradient descent techniques. This settles the time complexity of the projection operation $\Pi_{\mathcal{A}}(\cdot)$. However, $\Pi_A(\mathbf{z}) = VV^\top \mathbf{z}$ also gives us the following expression for the SU-incoherence constant.

$$\mu = \max_{\substack{\mathbf{u},\mathbf{v} \in S^{n-1} \\ \|\mathbf{v}\|_0 \le k}} \left(\mathbf{v}^\top VV^\top \mathbf{u}\right)^2 \le \max_{\substack{\mathbf{v} \in S^{n-1} \\ \|\mathbf{v}\|_0 \le k}} \left\|VV^\top \mathbf{v}\right\|_2^2 = \max_{\substack{S \subset [n] \\ |S| = k}} \frac{\|X_S\|_2^2}{\lambda_{\min}(XX^\top)}.$$

The above constants are readily available from prior works e.g. TORRENT (Bhatia et al. 2015) and are reproduced here (see Bhatia et al. 2015, Lemma 14 and Theorem 15). For any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

1. $\lambda_{\min}(XX^\top) \ge n - 3\sqrt{513dn + 178n\log\frac{2}{\delta}}$
2. $\max_{\substack{S \subset [n] \\ |S| = k}} \|X_S\|_2^2 \le k\left(1 + 3e\sqrt{6\log\frac{en}{k}}\right) + 3\sqrt{513dk + 178k\log\frac{1}{\delta}}$

To simplify the above bounds, we set $\delta = \frac{1}{d^2}$ and notice that for large enough d, we have $\log(2d^2) < \frac{d}{100}$ so that we get $\lambda_{\min}(XX^\top) \ge n - 3\sqrt{515dn}$ and $\max_{\substack{S \subset [n] \\ |S| = k}} \|X_S\|_2^2 \le k\left(1 + 3e\sqrt{6\log\frac{en}{k}}\right) + 3\sqrt{515dk}$, each with confidence at least $1 - \frac{1}{d^2}$. We also assume that n is large enough so that $\sqrt{515dn} < \frac{n}{300}$ ($n > 300^2 \cdot 515 \cdot d$ i.e. $n = \Omega(d)$ suffices to ensures this) so that we get $\lambda_{\min}(XX^\top) \ge \frac{99n}{100}$ and $\max_{\substack{S \subset [n] \\ |S| = k}} \|X_S\|_2^2 \le k\left(\frac{101}{100} + 3e\sqrt{6\log\frac{en}{k}}\right)$.

This gives us

$$\mu \le \left(\frac{100}{99}\right)\frac{k}{n}\left(\frac{101}{100} + 3e\sqrt{6\log\frac{en}{k}}\right).$$

Elementary calculations show that we have $\mu < \frac{1}{3}$ whenever $k \le \frac{n}{154}$. Since Theorem 1 assures a linear rate of convergence for APIS in the known support case whenever $\mu < \frac{1}{3}$, this finishes the proof.

However, we note that similar breakdown points can be obtained even if the data covariates come from other nice distributions, for example, sub-Gaussian distributions that include all distributions with bounded support, arbitrary (non-standard) Gaussian distributions, mixtures of Gaussian distributions, and many more. The following result sketches that APIS offers a linear rate of convergence even in this general setting. However, the breakdown point is less explicit due to the generality of the result.

Lemma 4 *If the corruption vectors are (adaptive adversarial) k-sparse vectors and the covariates $\mathbf{x}^i \in \mathbb{R}^d, i \in [n]$ are sampled i.i.d. from a sub-Gaussian distribution with sub-Gaussian norm R and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$, and $n = \Omega(d)$, then with probability at least $1 - \frac{1}{d^2}$,* APIS *offers exact recovery at a linear rate if $k < \frac{n}{\mathcal{O}(1)}$. The constants hidden in the $\mathcal{O}(\cdot), \Omega(\cdot)$ notations used in this statement are either universal or depend only on the sub-Gaussian norm R of the distribution.*

Proof We note that the projection operation $\Pi_{\mathscr{A}}(\cdot)$ can still be performed in $\mathcal{O}(nd)$ time in this case (by solving a least squares problem). As before, we have

$$\mu = \max_{\substack{\mathbf{u}, \mathbf{v} \in S^{n-1} \\ \|\mathbf{v}\|_0 \le k}} \left(\mathbf{v}^\top VV^\top \mathbf{u}\right)^2 \le \max_{\substack{\mathbf{v} \in S^{n-1} \\ \|\mathbf{v}\|_0 \le k}} \left\|VV^\top \mathbf{v}\right\|_2^2 = \max_{\substack{S \subset [n] \\ |S| = k}} \frac{\|X_S\|_2^2}{\lambda_{\min}(XX^\top)},$$

where $VV^\top = X^\top (XX^\top)^\dagger X$. For the case of sub-Gaussian distributions, the following relevant results are available (see Bhatia et al. 2015, Lemma 16 and Theorem 17). For any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have the following where c, C are universal constants that depend only on the sub-Gaussian norm R of the distribution.

1. $\lambda_{\min}(XX^\top) \ge n \cdot \lambda_{\min}(\Sigma) - C \cdot \sqrt{dn} - \sqrt{\frac{n}{c} \log \frac{2}{\delta}}$
2. $\max_{\substack{S \subset [n] \\ |S| = k}} \|X_S\|_2^2 \le k\left(\lambda_{\max}(\Sigma) + \sqrt{\frac{n}{ck} \log \frac{en}{k}}\right) + C \cdot \sqrt{kd} + \sqrt{\frac{n}{c} \log \frac{2}{\delta}}$

As before, to simplify the above bounds, we set $\delta = \frac{1}{d^2}$ and notice that for large enough d and $n = \Omega(d)$, we have $\lambda_{\min}(XX^\top) \ge \frac{99n}{100} \cdot \lambda_{\min}(\Sigma)$ and $\max_{\substack{S \subset [n] \\ |S| = k}} \|X_S\|_2^2 \le k\left(\lambda_{\max}(\Sigma) \cdot \frac{101}{100} + \sqrt{\frac{n}{ck} \log \frac{en}{k}}\right)$ which gives us

$$\mu \le \left(\frac{100}{99}\right)\frac{k}{n}\left(\frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} \cdot \frac{101}{100} + \frac{1}{\lambda_{\min}(\Sigma)}\sqrt{\frac{n}{ck} \log \frac{en}{k}}\right).$$

Assuming w.l.o.g. $\lambda_{\max}(\Sigma) \ge 1$ and denoting $\boldsymbol{\kappa} := \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)}$ as the condition number of the covariance matrix Σ gives us

$$\mu \le \mathcal{O}\left(\boldsymbol{\kappa} \cdot \frac{k}{n}\left(1 + \sqrt{\frac{n}{k} \log \frac{en}{k}}\right)\right),$$

which can be shown to assure $\mu < \frac{1}{3}$ when $k \le \mathcal{O}\left(\frac{n}{\boldsymbol{\kappa}}\right)$. Now notice that the above breakdown point depends on the condition number of the covariance matrix. This dependence is superfluous and can be removed, as we show below.

Notice that if we let $\tilde{X} = \Sigma^{-\frac{1}{2}} X$ where X is the covariate matrix used by the algorithm and Σ is the covariance matrix of the distribution generating the covariates, then we have

$$VV^\top = X^\top (XX^\top)^\dagger X = \tilde{X}^\top (\tilde{X}\tilde{X}^\top)^\dagger \tilde{X},$$

where $\tilde{X}$ is now a matrix of covariates assumed to be sampled from a (still) sub-Gaussian distribution but with identity covariance. This allows us to use the following improved upper bound on the incoherence constant

$$\mu = \max_{\substack{S \subset [n] \\ |S| = k}} \frac{\|\tilde{X}_S\|_2^2}{\lambda_{\min}(\tilde{X}\tilde{X}^\top)},$$

as well as

1. $\lambda_{\min}(\tilde{X}\tilde{X}^\top) \geq n - C \cdot \sqrt{dn} - \sqrt{\frac{n}{c}\log\frac{2}{\delta}}$
2. $\max_{\substack{S \subset [n] \\ |S| = k}} \|\tilde{X}_S\|_2^2 \leq k\left(1 + \sqrt{\frac{n}{ck}\log\frac{en}{k}}\right) + C \cdot \sqrt{kd} + \sqrt{\frac{n}{c}\log\frac{2}{\delta}}$

The above in turn give us

$$\mu \leq \mathcal{O}\left(\frac{k}{n}\left(1 + \sqrt{\frac{n}{k}\log\frac{en}{k}}\right)\right),$$

which can be shown to assure $\mu < \frac{1}{3}$ when $k < \frac{n}{\mathcal{O}(1)}$ where the constants hidden in the $\mathcal{O}(\cdot)$ notation are either universal or depend only on the sub-Gaussian norm R of the distribution. Note that the algorithm does not need to know Σ at all (either exactly or even approximately) for the above trick to work. The algorithm can continue to perform the $\Pi_{\mathscr{A}}(\cdot)$ projections using $VV^\top = X^\top(XX^\top)^\dagger X$ but the analysis uses the (equivalent) $VV^\top = \tilde{X}^\top(\tilde{X}\tilde{X}^\top)^\dagger\tilde{X}$ instead.

C Robust low-rank kernel regression using APIS

We recall that in this case, the uncorrupted signal satisfies $\mathbf{a}^* = G\boldsymbol{\alpha}^*$ where $G \in \mathbb{R}^{n\times n}$ be the Gram matrix with $G_{ij} = K(\mathbf{x}^i, \mathbf{x}^j)$ corresponding to a Mercer kernel $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such as the RBF kernel. Moreover, $\boldsymbol{\alpha}^*$ belongs to the span of the some s eigenvectors of G i.e. $\boldsymbol{\alpha}^* = V\boldsymbol{\gamma}^*$ where $\|\boldsymbol{\gamma}^*\|_0 \leq s$ and $V = [\mathbf{v}^1, \ldots, \mathbf{v}^r] \in \mathbb{R}^{n\times r}$ is the matrix of eigenvectors of G and r is the rank of G. As we will see, APIS offers the strongest guarantees in the case when $\boldsymbol{\alpha}^* \in \mathrm{span}(\mathbf{v}^1, \ldots, \mathbf{v}^s)$, i.e., when $\boldsymbol{\alpha}^*$ lies in the span of the the top eigenvectors.

Thus, in this case, we have known signal support i.e. $P = 1$ with $\mathscr{A} = A$ being the span of the top s eigenvectors of G and $\mathscr{B}$ being the union of subspaces of k-sparse vectors. Here we derive breakdown points for the case of kernel ridge regression. Lemma 5 presents this result for general Mercer kernels, whereas Lemma 6 will yield a specific breakdown point for the special case of the RBF kernel.

Lemma 5 *If the corruption vectors are (adaptive adversarial) k-sparse vectors and the uncorrupted signal lies in the span of the top s eigenvectors of a Gram matrix G corresponding to a Mercer kernel, then* APIS *offers exact recovery at a linear rate if* $3 \cdot \Lambda_k^{\mathrm{unif}}(G) < \lambda_s(G)$ *where* $\lambda_s(G)$ *is the* s^{th}*-largest eigenvalue of G and for any* $k > 0$, $\Lambda_k^{\mathrm{unif}}(G)$ *denotes the largest eigenvalue of any principal* $k \times k$ *sub-matrix of G. Moreover,* *the projection operation $\Pi_{\mathscr{A}}(\cdot)$ can be performed in $\mathscr{O}(ns)$ time in this case apart from a one-time cost of $\mathscr{O}(n^2 s)$.*

Proof Let $\mathbf{v}^1, \ldots, \mathbf{v}^s \in \mathbb{R}^n$ be the top-s eigenvectors of G i.e. $\tilde{V} = [\mathbf{v}^1, \ldots, \mathbf{v}^s] \in \mathbb{R}^{n \times s}$. Also, let the diagonal matrix containing the corresponding top-s eigenvalues $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_s > 0$ be denoted by $\tilde{\Sigma} = \operatorname{diag}(\lambda_1, \ldots, \lambda_s) \in \mathbb{R}^{s \times s}$. The time complexity of the projection step is settled by noting that the projection operator Π_A is given as $\Pi_A(\mathbf{z}) = \tilde{V}\tilde{V}^\top \mathbf{z}$. Calculating $\tilde{V}$ takes a one-time cost of $\mathscr{O}(n^2 s)$ whereas applying the projection operator requires two multiplications with an $n \times s$ matrix which takes $\mathscr{O}(ns)$ time.

Consider the matrix $\tilde{X} = \tilde{\Sigma}^{-\frac{1}{2}} \tilde{V}^\top G$. It is easy to see that

$$\tilde{X}^\top (\tilde{X}\tilde{X}^\top)^{-1} \tilde{X} = \tilde{V}\tilde{V}^\top.$$

Notice the parallels between the above and a similar expression derived in the linear regression case in the proof of Lemma 3. An identical analysis then gives us

$$\mu = \max_{\substack{\mathbf{u}, \mathbf{v} \in S^{n-1} \\ \|\mathbf{v}\|_0 \leq k}} \left(\mathbf{v}^\top \tilde{V}\tilde{V}^\top \mathbf{u}\right)^2 \leq \max_{\substack{\mathbf{v} \in S^{n-1} \\ \|\mathbf{v}\|_0 \leq k}} \left\|\tilde{V}\tilde{V}^\top \mathbf{v}\right\|_2^2 = \max_{\substack{S \subset [n] \\ |S| = k}} \frac{\|\tilde{X}_S\|_2^2}{\lambda_{\min}(\tilde{X}\tilde{X}^\top)}.$$

Now, clearly we have $\lambda_{\min}(\tilde{X}\tilde{X}^\top) \geq \lambda_s$ which lower bounds the denominator in the last expression. To upper bound the numerator, notice that $\tilde{X}_S = [\tilde{\mathbf{x}}^i]_{i \in S} \in \mathbb{R}^{s \times k}$ where $\tilde{\mathbf{x}}^i = \tilde{\Sigma}^{-\frac{1}{2}} \tilde{V}^\top G_i$ where G_i is the i^{th} column of the matrix G. Now consider $\hat{\mathbf{x}}^i = \Sigma^{-\frac{1}{2}} V^\top G_i$ where $\Sigma \in \mathbb{R}^{r \times r}$ is the diagonal matrix of all the eigenvalues of G, not just the top-s ones (assuming G is of rank r) and $V \in \mathbb{R}^{n \times r}$ is the matrix of all the eigenvectors of G and let $\hat{X} = [\hat{\mathbf{x}}^i]_{i=1}^n \in \mathbb{R}^{s \times n}$ and, in particular, $\hat{X}_S = [\hat{\mathbf{x}}^i]_{i \in S} \in \mathbb{R}^{s \times k}$.

Since $\tilde{X}_S$ is a projection of $\hat{X}_S$ onto the top-s eigenvectors of G, we conclude that $\|\tilde{X}_S\|_2^2 \leq \left\|\hat{X}_S\right\|_2^2$. However, notice that $\hat{X}^\top \hat{X} = G$ and thus, $\left\|\hat{X}_S\right\|_2^2$ is upper bounded by the largest eigenvalue of the principal sub-matrix G_S^S. Thus, we have

$$\mu \leq \max_{\substack{S \subset [n] \\ |S| = k}} \frac{\|\tilde{X}_S\|_2^2}{\lambda_{\min}(\tilde{X}\tilde{X}^\top)} \leq \frac{\Lambda_k^{\text{unif}}(G)}{\lambda_s(G)}.$$

This concludes the proof upon noting that we get $\mu < \frac{1}{3}$ as desired by Theorem 1 for APIS to offer exact recovery at a linear rate whenever $3 \cdot \Lambda_k^{\text{unif}}(G) < \lambda_s(G)$.

We note that the above proof does not use anywhere the fact that the signal has support only among the top-s eigenvectors of the Gram matrix. However, if we start considering other sets of s eigenvectors as possible support, we will run into adverse incoherence constants. Specifically, if the set of eigenvectors contains the smallest eigenvector of G as well, then we would have

$$\mu \leq \max_{\substack{S \subset [n] \\ |S| = k}} \frac{\|\tilde{X}_S\|_2^2}{\lambda_{\min}(G)}.$$

Notice that the denominator now has $\lambda_{\min}(G)$ instead of $\lambda_s(G)$. Since the eigenvalues of Gram matrices w.r.t popular kernels such as RBF decay rapidly (see proof of Lemma 6 below), this would mean that μ could take a very large value and it may be impossible to satisfy $\mu < \frac{1}{9}$ no matter how small the value of k. That is why we restrict the support to the top-s eigenvectors. However, Appendix E.1 shows that APIS offers recovery even if signals are not totally represented by the top-s eigenvectors but merely well-approximated by them.

C.1 Breakdown point derivations for the RBF kernel

Our goal in this discussion will be to establish the following breakdown point result for robust kernel ridge regression settings.

Lemma 6 *If the corruption vectors are (adaptive adversarial) k-sparse vectors and the uncorrupted signal lies in the span of the top s eigenvectors of a Gram matrix G corresponding to the RBF kernel $\kappa(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|_2^2}{h^2}\right)$ with $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ for $d > 1$ and h being the bandwidth parameter of the kernel, with the data covariates $\mathbf{x}^1, \ldots, \mathbf{x}^n \in \mathbb{R}^d$ sampled from the uniform distribution over the unit sphere S^{d-1}, then* APIS *offers exact recovery at a linear rate in the following settings. We note that these conditions are neither exhaustive nor necessary but merely some sufficient conditions in which recovery is guaranteed by* APIS.

1. *Case 1: $d = 2$, $s \geq e$: if $k \leq \sqrt{n}, s^s \leq n^{\frac{1}{5}}$ (i.e. $s \leq \mathcal{O}(\log n / \log\log n)$), and $h \in \left[\sqrt{\frac{40}{\log n}}, \frac{1.13}{(20.4)^{\frac{2.5}{\log n}}}\right]$, then with probability at least $1 - 4\exp(-n^{\frac{2}{5}})$, we have $3 \cdot \Lambda_k^{\text{unif}}(G) < \lambda_s(G)$ i.e. $\mu < \frac{1}{3}$ as guaranteed by Lemma 5.*
2. *Case 2: $d > 2$, $s \geq e$: if $k \leq \sqrt{n}$, $\frac{(s+\frac{d}{2}-1)^{s+\frac{d}{2}-1}}{\exp(s-1)\left(\frac{d}{2}\right)^{\frac{d+1}{2}}} \leq n^{\frac{1}{5}}$, and $h \in \left[\sqrt{\frac{40}{\log n}}, \left(\frac{n^{\frac{1}{20}}}{18.8}\right)^{\frac{1}{2s}}\right]$, then with probability at least $1 - 4\exp(-n^{\frac{2}{5}})$, we have $3 \cdot \Lambda_k^{\text{unif}}(G) < \lambda_s(G)$ i.e. $\mu < \frac{1}{3}$ as guaranteed by Lemma 5.*

Note that in both cases, the range which the bandwidth is allowed to take while ensuring recovery expands with n. For example, in the $d = 2$ case, in the limit $n \to \infty$, the range expands to $[0, 1.13]$ since $(20.4)^{\frac{2.5}{\log n}} \to 1$ as $n \to \infty$ since the exponent $\frac{2.5}{\log n} \to 0$.

Sample complexity Before giving derivations for the above results, we put in a word about the sample complexity.

1. *Case 1: $d = 2, s \geq e$*: $n = \Omega(1)$ samples and $s \leq \mathcal{O}(\log n / \log\log n)$ clearly suffice in this case.
2. *Case 2: $d > 2, s \geq e$*: we first simplify the expression $\frac{(s+\frac{d}{2}-1)^{s+\frac{d}{2}-1}}{\exp(s-1)\left(\frac{d}{2}\right)^{\frac{d+1}{2}}}$ using simple inequalities such as $(x + y)^p \leq 2^p(x^p + y^p)$ for any $x, y \in \mathbb{R}_+, p \in \mathbb{N}$ to obtain the following inequality (using the shorthand $D := \frac{d}{2} - 1$ to avoid clutter)

$$2^D\left(D^s + \frac{s^s}{D^D} + \left(\frac{s}{D}\right)^D\right) < n^{\frac{1}{5}}$$

Simple calculations show that $n = (\Omega(1))^d$ as well as $s < \mathcal{O}\big(\log_d(n)\big)$ suffice to satisfy the above requirement.

Note that in both cases, we can tolerate upto $k \leq \sqrt{n}$ corruptions.

C.2 Some pre-calculations

Let the data points $\{\mathbf{x}^i\}$ be sampled from the uniform distribution over S^{d-1} with $d > 1$, and the RBF kernel, $\kappa(\mathbf{x}^i, \mathbf{x}^j) = \exp\left(-\frac{\|\mathbf{x}^i - \mathbf{x}^j\|^2}{h^2}\right)$. Let π_r be the r^{th} largest, $r \in \mathbb{N} \cup \{0\}$ distinct eigenvalue of the integral transform operator corresponding to the kernel function κ, then (Minh et al. 2006, Theorem 2) states

$$\pi_r = \exp\left(-\frac{2}{h^2}\right)h^{d-2}I_{r+\frac{d}{2}-1}\left(\frac{2}{h^2}\right)\Gamma\left(\frac{d}{2}\right), \tag{3}$$

where, I denotes the modified Bessel function of the first kind. Here each π_r occurs with multiplicity $\frac{(2r+d-2)(r+d-3)!}{r!(d-2)!}$. The eigenvalues also satisfy

$$\left(\frac{2e}{h^2}\right)^r \frac{A_1}{(2r+d-2)^{r+\frac{d-1}{2}}} < \pi_r < \left(\frac{2e}{h^2}\right)^r \frac{A_2}{(2r+d-2)^{r+\frac{d-1}{2}}}, \tag{4}$$

where A_1, A_2 being independent of r are given as follows:

$$\begin{aligned} A_1 &= \frac{2^{\frac{d}{2}-1}}{\sqrt{\pi}} \exp\left(-\frac{2}{h^2} - \frac{1}{12} + \frac{d}{2} - 1\right)\Gamma\left(\frac{d}{2}\right) \\ A_2 &= \frac{2^{\frac{d}{2}-1}}{\sqrt{\pi}} \exp\left(-\frac{2}{h^2} + \frac{1}{h^4} + \frac{d}{2} - 1\right)\Gamma\left(\frac{d}{2}\right) \end{aligned},$$

with Γ denoting the Gamma function.

Let $\lambda_r^{(n)}$ be the r^{th}-largest eigenvalue of the $n \times n$ gram matrix $G_{ij} = k(\mathbf{x}^i, \mathbf{x}^j)$ over n data points. We have from Rosasco et al. (2010, Theorems 5 and 7) that for a normalized mercer kernel $\kappa(\mathbf{x}^i, \mathbf{x}^i) \leq 1$ (which the RBF kernel does satisfy), with probability $1 - 2\exp(-\tau)$,

$$\left|\lambda_r^{(n)} - n\pi_r\right| \leq 2\sqrt{2n\tau},$$

Hence, for a given principal sub-matrices of size k,

$$
\begin{aligned}
&\Pr\left(\lambda_0^{(k)} > k\pi_0 + 2\sqrt{2k\tau_1}\right) \leq 2e^{-\tau_1}\\
&\quad \implies \Pr\left(\bigcup_{\text{sub-matrices of size } k} \{\lambda_0^{(k)} > k\pi_0 + 2\sqrt{2k\tau_1}\}\right) \leq \binom{n}{k} 2e^{-\tau_1} \leq \left(\frac{ne}{k}\right)^k 2e^{-\tau_1}\\
&\quad \iff \Pr\left(\Lambda_k^{unif} > k\pi_0 + 2\sqrt{2k\tau_1}\right) \leq \left(\frac{ne}{k}\right)^k 2e^{-\tau_1} =: \frac{\delta}{2}\\
&\quad \iff \Pr\left(\Lambda_k^{unif} > k\pi_0 + 2\sqrt{2k\left(k\ln\left(\frac{ne}{k}\right) + \ln\frac{4}{\delta}\right)}\right) \leq \frac{\delta}{2} \quad \text{putting, } \tau_1 = k\ln\left(\frac{ne}{k}\right) + \ln\frac{4}{\delta}.
\end{aligned}
\tag{5}
$$

Also we have,

$$
\Pr\left(\lambda_s^{(n)} < n\pi_s - 2\sqrt{2n\ln\frac{4}{\delta}}\right) \leq \frac{\delta}{2} \tag{6}
$$

Combining Eqs. (5) and (6):

$$
\begin{aligned}
&\Pr\left(\left[3\Lambda_k^{unif} > 3k\pi_0 + 6\sqrt{2k(k\ln\left(\frac{ne}{k}\right) + \ln\frac{4}{\delta})}\right] \cup \left[\lambda_s^{(n)} < n\pi_s - 2\sqrt{2n\ln\frac{4}{\delta}}\right]\right) \leq \delta\\
&\quad \iff \Pr\left(\left[3\Lambda_k^{unif} \leq 3k\pi_0 + 6\sqrt{2k(k\ln\left(\frac{ne}{k}\right) + \ln\frac{4}{\delta})}\right] \cap \left[n\pi_s - 2\sqrt{2n\ln\frac{4}{\delta}} \leq \lambda_s^{(n)}\right]\right) \geq 1-\delta\\
&\quad \iff \Pr\left(3\Lambda_k^{unif} \leq \lambda_s^{(n)}\right) \geq 1-\delta
\end{aligned}
$$

whenever,

$$
\begin{aligned}
&3k\pi_0 + 6\sqrt{2k\left(k\ln\frac{ne}{k} + \ln\frac{4}{\delta}\right)} \leq n\pi_s - 2\sqrt{2n\ln\frac{4}{\delta}}\\
&\quad \iff \frac{3k}{n}\left(\pi_0 + 2\sqrt{2}\sqrt{\ln\frac{ne}{k} + \frac{1}{k}\ln\frac{4}{\delta}}\right) + 2\sqrt{\frac{2}{n}\ln\frac{4}{\delta}} \leq \pi_s\\
&\quad \Longleftarrow \frac{3k}{n}\left(\pi_0 + 2\sqrt{2}\sqrt{\ln\frac{ne}{k}}\right) + 2\sqrt{2}\sqrt{\frac{1}{n}\ln\frac{4}{\delta}}\left(3\sqrt{\frac{k}{n}} + 1\right) \leq \pi_s \quad \text{using, } \sqrt{a+b} \leq \sqrt{a} + \sqrt{b}
\end{aligned}
\tag{7}
$$

We break the remaining proof into the two cases $d = 2$ and $d > 2$ in the following two subsections.

C.3 Case 1: $d = 2, s \geq e$

From Eq. (3) and using $I_0(x) = \frac{1}{\pi}\int_0^{\pi} \exp(x\cos(\theta))d\theta \leq \exp(x)$ we have,

$$
\pi_0 = \exp\left(-\frac{2}{h^2}\right) I_0\left(\frac{2}{h^2}\right) \leq \exp\left(-\frac{2}{h^2}\right)\exp\left(\frac{2}{h^2}\right) \leq 1
$$

From Eq. (4) we have, for $s \geq e$ and $s^s \leq n^{\epsilon_2}$

$$\begin{aligned}
\pi_s &\geq \left(\frac{2e}{h^2}\right)^s \frac{\exp\left(-\frac{2}{h^2} - \frac{1}{12}\right)}{\sqrt{\pi}(2s)^{s+\frac{1}{2}}} = \frac{\exp\left(-\frac{2}{h^2}\right)}{h^{2s}} \frac{\exp(s - \frac{1}{12})}{\sqrt{2\pi} s^{s+\frac{1}{2}}} \\
&\geq \frac{\exp\left(-\frac{2}{h^2}\right)}{h^{2s}} \frac{\exp\left(\frac{11}{12}\right)}{\sqrt{2\pi}} \frac{\exp(s-1)}{s^{s+\frac{1}{2}}} \\
&\geq \frac{\exp\left(-\frac{2}{h^2}\right)}{h^{2s}} \frac{\exp\left(\frac{11}{12}\right)}{\sqrt{2\pi}} \frac{1}{s^{s-\frac{1}{2}}} \quad \text{using, } \exp(s-1) \geq s \\
&\geq \frac{\exp\left(-\frac{2}{h^2}\right)}{h^{2s}} \frac{1}{s^s} \quad \text{using, } s \geq e \\
&\geq \frac{\exp\left(-\frac{2}{h^2}\right)}{h^{2s}} \frac{1}{n^{\epsilon_2}}
\end{aligned}$$

From Eq. (7) we require:

$$\frac{1}{n^{\epsilon_2}} \geq h^{2s} \exp\left(\frac{2}{h^2}\right)\left(\frac{3k}{n} + \frac{6\sqrt{2}k}{n}\sqrt{\ln\frac{ne}{k}} + 2\sqrt{2}\sqrt{\frac{1}{n}\ln\frac{4}{\delta}}(3\sqrt{\frac{k}{n}} + 1)\right)$$

Let $k \leq n^{\epsilon_3}$. To satisfy the above requirement we break it into following cases:

$$\begin{aligned}
\frac{1}{(9+8\sqrt{2})n^{\epsilon_2}} &\geq h^{2s} \exp\left(\frac{2}{h^2}\right)\frac{3k}{n} \\
\Longleftarrow n^{1-\epsilon_2-\epsilon_3} &\geq (9+8\sqrt{2})h^{2s}\exp\left(\frac{2}{h^2}\right)
\end{aligned} \tag{8}$$

Since $\frac{k}{n}\sqrt{\ln\frac{ne}{k}} \leq \left(\frac{k}{n}\right)^{\frac{3}{5}} \leq n^{\frac{3(\epsilon_3-1)}{5}}$, for $0 \leq \frac{k}{n} \leq 0.5$

$$\begin{aligned}
\frac{6\sqrt{2}}{(9+8\sqrt{2})}\frac{1}{n^{\epsilon_2}} &\geq 6\sqrt{2}\frac{k}{n}\sqrt{\ln\frac{ne}{k}}h^{2s}\exp\left(\frac{2}{h^2}\right) \\
\Longleftarrow n^{\frac{3}{5}(1-\epsilon_3)-\epsilon_2} &\geq (9+8\sqrt{2})h^{2s}\exp\left(\frac{2}{h^2}\right)
\end{aligned} \tag{9}$$

Assume, $\frac{1}{n^{\epsilon_4}}\sqrt{\ln\frac{4}{\delta}} = 1$ so that, $\delta = 4\exp(-n^{2\epsilon_4})$with, $0 < \epsilon_4 < \frac{1}{2}$

$$\frac{6+2\sqrt{2}}{(9+8\sqrt{2})}\frac{1}{n^{\epsilon_2}} \geq 2\sqrt{2}h^{2s}\exp\left(\frac{2}{h^2}\right)\sqrt{\frac{1}{n}\ln\frac{4}{\delta}}\left(3\left(\frac{k}{n}\right)^{0.5}+1\right)$$

$$\Longleftarrow \frac{6+2\sqrt{2}}{(9+8\sqrt{2})}\frac{1}{n^{\epsilon_2}} \geq h^{2s}\exp\left(\frac{2}{h^2}\right)\sqrt{\frac{1}{n}\ln\frac{4}{\delta}}(6+2\sqrt{2}) \quad \text{since, } \frac{1}{2} \geq \frac{k}{n} \tag{10}$$

$$\Longleftrightarrow n^{\frac{1}{2}-\epsilon_4-\epsilon_2} \geq (9+8\sqrt{2})\,h^{2s}\exp\left(\frac{2}{h^2}\right) \text{using, } \frac{1}{n^{\epsilon_4}}\sqrt{\ln\frac{4}{\delta}} = 1$$

We now summarize, last three conditions in order to satisfy Eq. (7)

- *Breakdown point:* we set $\epsilon_3 = \frac{1}{2}$, so as to obtain $\frac{k}{n} \leq n^{\epsilon_3-1} = n^{-\frac{1}{2}}$
- *Confidence bound:* we set $\frac{1}{2} - \epsilon_4 - \epsilon_2 = \frac{3}{5}(1-\epsilon_3) - \epsilon_2$, so that we get $\epsilon_4 = \frac{1}{5}$. This gives us, $\delta = 4\exp(-n^{2\epsilon_4}) = 4\exp(-n^{\frac{2}{5}})$
- *Generality:* we need $\frac{1}{2} - \epsilon_4 - \epsilon_2 \geq 0 \implies \frac{3}{10} - \epsilon_2 \geq 0$ Set $\epsilon_2 = \frac{2}{10}$, so that $s^s \leq n^{\epsilon_2} = n^{\frac{1}{5}}$
- *Bandwidth:* Using $s \leq s\ln s \leq \epsilon_2\ln(n) = \frac{\ln(n)}{5}$. We require, $n^{\frac{1}{10}} \geq 20.4h^{\frac{\ln(n)}{2.5}}\exp\left(\frac{2}{h^2}\right)$, which is satisfied if:

$$n^{\frac{1}{20}} \geq \exp\left(\frac{2}{h^2}\right) \quad \text{and} \quad n^{\frac{1}{20}} \geq 20.4h^{\frac{\ln(n)}{2.5}}$$

$$\sqrt{\frac{40}{\ln(n)}} \leq h \leq \frac{1.13}{(20.4)^{\frac{2.5}{\ln(n)}}}$$

Note that the permissible range for h improves with n.

C.4 Case: $d > 2, s \geq e$

For $d > 2$ we have from eq. 4,

$$\pi_0 < \frac{(2e)^{\frac{d}{2}-1}\exp\left(-\frac{2}{h^2}+\frac{1}{h^4}\right)\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}(d-2)^{\frac{d-1}{2}}} \leq 2\exp\left(\frac{1}{h^4}\right),$$

where we have used that for $d > 2$, we always have $\frac{(2e)^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}(d-2)^{\frac{d-1}{2}}} \leq 2$. A short proof of this is given below. From[1], we deduce that we always have $\Gamma(x) \leq \sqrt{2\pi}x^{x-\frac{1}{2}}\exp\left(\frac{1}{12x}-x\right)$ so that,

$$\begin{aligned}\frac{(2e)^{\frac{d}{2}-1}\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}(d-2)^{\frac{d-1}{2}}} &\leq \frac{2^{\frac{d}{2}-1}\sqrt{2\pi}\exp\left(\frac{d}{2}-1+\frac{1}{6d}-\frac{d}{2}\right)\left(\frac{d}{2(d-2)}\right)^{\frac{d-1}{2}}}{\sqrt{\pi}} \\ &= \exp\left(\frac{1}{6d}-1\right)\left(\frac{d}{d-2}\right)^{\frac{d-1}{2}} \\ &\leq 3\exp\left(\frac{1}{18}-1\right) \text{ since, both are strictly decreasing on } d > 2 \\ &= 1.11 < 1.2\end{aligned}$$

[1] https://dlmf.nist.gov/5.6.

Coming back to the original argument, assume, $\frac{\exp(s-1)\left(\frac{d}{2}\right)^{\frac{d+1}{2}}}{(s+\frac{d}{2}-1)^{s+\frac{d}{2}-1}} \geq \frac{1}{n^{\epsilon_2}}$ and $s \geq e$ so that,

$$\pi_s > \left(\frac{2e}{h^2}\right)^s \frac{(2e)^{\frac{d}{2}-1}\exp\left(-\frac{2}{h^2}-\frac{1}{12}\right)\Gamma\left(\frac{d}{2}\right)}{\sqrt{\pi}(2s+d-2)^{s+\frac{d-1}{2}}} \geq \left(\frac{e}{h^2}\right)^s \frac{(e)^{\frac{d}{2}-1}\exp\left(-\frac{2}{h^2}-\frac{1}{12}\right)\sqrt{2\pi\frac{d}{2}}\left(\frac{d}{2e}\right)^{\frac{d}{2}}}{\sqrt{2\pi}(s+\frac{d}{2}-1)^{s+\frac{d-1}{2}}}$$

$$= \frac{1}{h^{2s}} \frac{\exp(s-\frac{2}{h^2}-\frac{13}{12})\left(\frac{d}{2}\right)^{\frac{d+1}{2}}}{(s+\frac{d}{2}-1)^{s+\frac{d-1}{2}}}$$

$$\geq \frac{\exp\left(-\frac{2}{h^2}\right)}{h^{2s}} \frac{\exp(s-1)\left(\frac{d}{2}\right)^{\frac{d+1}{2}}}{(s+\frac{d}{2}-1)^{s+\frac{d}{2}-1}} \quad \text{using, } \exp\left(-\frac{1}{12}\right)\sqrt{s+\frac{d}{2}-1} \geq \exp\left(-\frac{1}{12}\right)\sqrt{e} \geq 1$$

$$\geq \frac{\exp\left(-\frac{2}{h^2}\right)}{h^{2s}} \frac{1}{n^{\epsilon_2}}$$

From Eq. (7), we require,

$$\frac{1}{n^{\epsilon_2}} \geq h^{2s}\exp\left(\frac{2}{h^2}\right)\left(\frac{3k}{n}\left(1.2\exp(\frac{1}{h^4}) + 2\sqrt{2}\sqrt{\ln\frac{ne}{k}}\right) + 2\sqrt{2}\sqrt{\frac{1}{n}\ln\frac{4}{\delta}}(3\left(\frac{k}{n}\right)^{\frac{1}{2}}+1)\right)$$

Let $k \leq n^{\epsilon_3}$. In order to satisfy the above requirement we break it into following three cases:

$$\frac{1.4}{18.8}\frac{1}{n^{\epsilon_2}} \geq 3.6h^{2s}\exp\left(\frac{2}{h^2}+\frac{1}{h^4}\right)\frac{k}{n}$$
$$\Longleftarrow n^{1-\epsilon_2-\epsilon_3} \geq 18.8h^{2s}\exp\left(\left(1+\frac{1}{h^2}\right)^2\right) \tag{11}$$

Since $\frac{k}{n}\sqrt{\ln\left(\frac{ne}{k}\right)} \leq \left(\frac{k}{n}\right)^{\frac{3}{5}} \leq n^{\frac{3(\epsilon_3-1)}{5}}$,

$$\frac{8.5}{18.8}\frac{1}{n^{\epsilon_2}} \geq 6\sqrt{2}\frac{k}{n}\sqrt{\ln\left(\frac{ne}{k}\right)}h^{2s}\exp\left(\frac{2}{h^2}\right)$$
$$\Longleftarrow n^{\frac{3}{5}(1-\epsilon_3)-\epsilon_2} \geq 18.8h^{2s}\exp\left(\frac{2}{h^2}\right) \tag{12}$$

Assume, $\frac{1}{n^{\epsilon_4}}\sqrt{\ln\frac{4}{\delta}} = 1$ so that, $\delta = 4\exp(-n^{2\epsilon_4})$with, $0 < \epsilon_4 < \frac{1}{2}$

$$\frac{8.9}{18.8}\frac{1}{n^{\epsilon_2}} \geq 2\sqrt{2}h^{2s}\exp\left(\frac{2}{h^2}\right)\sqrt{\frac{1}{n}\ln\frac{4}{\delta}}\left(3\left(\frac{k}{n}\right)^{0.5}+1\right)$$
$$\Longleftarrow \frac{8.9}{18.8}\frac{1}{n^{\epsilon_2}} \geq h^{2s}\exp\left(\frac{2}{h^2}\right)\sqrt{\frac{1}{n}\ln\frac{4}{\delta}}(6+2\sqrt{2}) \quad \text{since, } \frac{1}{2} \geq \frac{k}{n} \tag{13}$$
$$\Longleftrightarrow n^{\frac{1}{2}-\epsilon_4-\epsilon_2} \geq 18.8\,h^{2s}\exp\left(\frac{2}{h^2}\right) \text{using, } \frac{1}{n^{\epsilon_4}}\sqrt{\ln\frac{4}{\delta}} = 1$$

Below we instantiate the variables ϵ_2, ϵ_3 and ϵ_4 which satisfies the above three conditions simultaneously.

- *Breakdown point:* set $\epsilon_3 = \frac{1}{2}, \frac{k}{n} \leq n^{\epsilon_3 - 1} = n^{-\frac{1}{2}}$
- *Confidence bound:* set $\frac{1}{2} - \epsilon_4 - \epsilon_2 = \frac{3}{5}(1 - \epsilon_3) - \epsilon_2$, so that $\epsilon_4 = \frac{1}{5}$. This gives, $\delta = 4\exp(-n^{2\epsilon_4}) = 4\exp(-n^{\frac{2}{5}})$
- *Universality:* we need $\frac{1}{2} - \epsilon_4 - \epsilon_2 \geq 0 \implies \frac{3}{10} - \epsilon_2 \geq 0$. Set $\epsilon_2 = \frac{2}{10}$, so that $\frac{(s+\frac{d}{2}-1)^{s+\frac{d}{2}-1}}{\exp(s-1)\left(\frac{d}{2}\right)^{\frac{d+1}{2}}} \leq n^{\epsilon_2} = n^{\frac{1}{5}}$
- *Bandwidth:* instantiating Eq. (13) ,$n^{\frac{1}{10}} \geq 18.8h^{2s}\exp\left(\frac{2}{h^2}\right)$. we require:

$$n^{\frac{1}{20}} \geq \exp\left(\frac{2}{h^2}\right) \qquad \text{and,} \quad n^{\frac{1}{20}} \geq 18.8h^{2s}$$

$$\iff h \geq \sqrt{\frac{40}{\ln(n)}} \qquad \text{and,} \quad h \leq \left(\frac{n^{\frac{1}{20}}}{18.8}\right)^{\frac{1}{2s}}$$

Assume Eq. (13) holds. In order to satisfy Eq. (11) we further require,

$$\begin{aligned} & n^{\frac{3}{10}} \geq 18.8h^{2s}\exp\left(\left(1+\frac{1}{h^2}\right)^2\right) \\ \Longleftarrow\ & n^{\frac{3}{10}} \geq 18.8h^{2s}\exp\left(\frac{2}{h^2}\right)\exp\left(1+\frac{1}{h^4}\right) \\ \Longleftarrow\ & n^{\frac{3}{10}} \geq n^{\frac{1}{10}}\exp\left(1+\frac{1}{h^4}\right) \\ \Longleftarrow\ & n^{\frac{1}{5}} \geq \exp\left(\frac{1}{h^4}\right) \quad \Longleftarrow h \geq \left(\frac{5}{\ln(n)-5}\right)^{\frac{1}{4}} \end{aligned}$$

Hence to satisfy all conditions on the bandwidth we require,

$$\max\left\{\sqrt{\frac{40}{\ln(n)}}, \left(\frac{5}{\ln(n)-5}\right)^{\frac{1}{4}}\right\} \leq h \leq \left(\frac{n^{\frac{1}{20}}}{18.8}\right)^{\frac{1}{2s}}$$

Note that here as well, the acceptable range of bandwidth improves with s that in turn improves with n.

D Robust signal transforms using APIS

Table 4 is a subset of Table 1, it only presents the rows concerning signals that have a sparse representation in a basis such as Fourier, wavelet, etc., with the corruption being either a sparse vector or having a sparse representation in the noiselet basis.

A proof of the breakdown points for examples in the first row i.e. when the signal has an s-sparse representation in Fourier, Hadamard, or noiselet bases, is given below. The proof is quite generic and holds for all transformations. The $n \times n$ design matrix has all its entries of magnitude $\mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$ which is true of the design matrices of the Fourier, Hadamard, and noiselet transforms.

Table 4 This table is a subset of Table 1 and presents only the rows concerning signals that have a sparse representation in a basis such as Fourier, wavelet etc with the corruption being either a sparse vector or having a sparse representation in the noiselet basis

Signal	Corruption	Breakdown point	Time per $\mathit{\Pi}_{\mathscr{A}}(\cdot)$	Time per $\mathit{\Pi}_{\mathscr{B}}(\cdot)$	References
s-Sparse in either Fourier, Hadarmard or noiselet bases	k-sparse	$sk < \frac{n}{9}$ e.g. $s, k \leq \frac{\sqrt{n}}{3}$	$\mathscr{O}(n \log n)$	$\mathscr{O}(n \log n)$	Foucart and Rauhut (2013)
s-Sparse in Fourier or wavelet (Haar, Daubechies D4/D8)	k noiselet-sparse	$sk < \frac{n}{27}$	As above	As above	Candes and Wakin (2008) and Foucart and Rauhut (2013)
s Noiselet-sparse	k-Sparse in Fourier or wavelet (Haar, Daubechies D4/D8)	$sk < \frac{n}{27}$	As above	As above	Candes and Wakin (2008) and Foucart and Rauhut (2013)

We note that noiselet corruptions can be dense vectors i.e. have $\|\mathbf{b}^*\|_0 = n$ despite having a sparse representation in the noiselet basis

Lemma 7 *Consider the $n \times n$ (orthonormal) design matrix U corresponding to a transformation such as Fourier etc. Let $\mathscr{A}$ be the union of subspaces of all signals that have an s-sparse representation in this basis i.e. $\mathscr{A} = \left\{\mathbf{a}^* : \mathbf{a}^* = U\boldsymbol{\alpha}^*, \|\boldsymbol{\alpha}^*\|_0 \leq s\right\}$. Also let $\mathscr{B}$ be the union of subspaces corresponding to k-sparse vectors i.e. $\mathscr{B} = \left\{\mathbf{b}^* : \|\mathbf{b}^*\|_0 \leq k\right\}$. Then the pair $(\mathscr{A}, \mathscr{B})$ is μ-SU incoherent (see Sect. 6.1) for $\mu \leq \frac{sk}{n}$ if every entry of the design matrix U satisfies $|U_{ij}| \leq \frac{1}{\sqrt{n}}$.*

Proof Note that to bound the SU incoherence constant, we only need to bound

$$\max_{\substack{\boldsymbol{\alpha} \in S^{n-1},\, \|\boldsymbol{\alpha}\|_0 \leq s \\ \mathbf{b} \in S^{n-1},\, \|\mathbf{v}\|_0 \leq k}} \left((U\boldsymbol{\alpha})^\top \mathbf{b}\right)^2$$

Since $\boldsymbol{\alpha}, \mathbf{b}$ are sparse vectors, we have $\boldsymbol{\alpha}^\top U^\top \mathbf{b} \leq \left\|U_S^K\right\|_2$ where $S = \operatorname{supp}(\boldsymbol{\alpha})$ and $K = \operatorname{supp}(\mathbf{b})$ are the supports of $\boldsymbol{\alpha}, \mathbf{b}$. Now, for any matrix $A \in \mathbb{R}^{s \times k}$, we have $\|A\|_2 \leq \sqrt{sk} \cdot \|A\|_\infty$. Since U_S^K is effectively an $s \times k$ matrix since its other rows and columns are zeroed out, this gives us $\left\|U_S^K\right\|_2 \leq \sqrt{sk} \cdot \nu$ where $\nu := \left\|U_S^K\right\|_\infty$. However, by assumption, $\|U\|_\infty \leq \frac{1}{\sqrt{n}}$ which gives us $(U\boldsymbol{\alpha})^\top \mathbf{b} \leq \sqrt{\frac{sk}{n}}$ and thus, $\mu \leq \max_{\substack{\boldsymbol{\alpha} \in S^{n-1},\, \|\boldsymbol{\alpha}\|_0 \leq s \\ \mathbf{b} \in S^{n-1},\, \|\mathbf{v}\|_0 \leq k}} \left((U\boldsymbol{\alpha})^\top \mathbf{b}\right)^2 \leq \frac{sk}{n}$ which finishes the proof.

An equivalent incoherence bound can also be derived from the results of Foucart and Rauhut (2013, Ch. 12) who effectively show that $\nu \leq \frac{1}{\sqrt{n}}$, but we presented the above proof in our notation for the sake of convenience.

Corollary 1 APIS *offers a linear rate of recovery when the signal is s-sparse in either the Fourier, Hadarmard or noiselet bases and the corruption is a k-sparse vector, whenever $sk < \frac{n}{9}$.*

Proof Lemma 7 shows that the SU-incoherence constant in these cases is bounded by $\mu \leq \frac{sk}{n}$. Theorem 1 shows that APIS has a linear rate of recovery when $\mu < \frac{1}{9}$. Combining the two finishes the proof.

A proof of the breakdown points in the second and the third rows of Table 4 i.e. when the signal has an s-sparse representation in the Fourier or wavelet (Haar, Daubechies D4/D8) bases and the corruption has a k-sparse representation in the noiselet basis or the vice-versa, is given below. We note that corruptions having a sparse representation in the noiselet, wavelet, or Fourier bases can nevertheless be dense as vectors i.e. have $\|\mathbf{b}^*\|_0 = n$.

Lemma 8 APIS *offers a linear rate of recovery when the signal is s-sparse in Fourier or wavelet (Haar, Daubechies D4/D8) bases and the corruption has a k-sparse representation in the noiselet basis, or vice versa, whenever $sk < \frac{n}{27}$.*

Proof The proof is a generalization of the one used for Lemma 7. Notice that here we have two bases involved, one for the signal (e.g., wavelet) and one for the corruption (e.g., noiselet). Let U, V denote the design matrices corresponding to these two bases. Then it is easy to see that calculating the SU-incoherence constant μ requires us to bound

$$\max_{\substack{\mathbf{u} \in S^{n-1}, \|\mathbf{u}\|_0 \leq s \\ \mathbf{v} \in S^{n-1}, \|\mathbf{v}\|_0 \leq k}} \left(\mathbf{u}^\top U^\top V \mathbf{v}\right)^2$$

Going as before, we can see that since $\mathbf{u}, \mathbf{v}$ are sparse vectors, we have $\mathbf{u}^\top U^\top V \mathbf{v} \leq \left\| U_S^\top V_K \right\|_2$ where $S = \operatorname{supp}(\mathbf{u})$ and $K = \operatorname{supp}(\mathbf{v})$ are the supports of $\mathbf{u}, \mathbf{v}$ respectively. Now, as $U_S^\top V_K$ is effectively an $s \times k$ matrix since all its other rows and columns are zeroed out, we have $\left\| U_S^\top V_K \right\|_2 \leq \sqrt{sk} \cdot \nu$ where $\nu := \left\| U_S^\top V_K \right\|_\infty$. Now, results from Candes and Wakin (2008), Foucart and Rauhut (2013) show us that $\nu \leq 3$ for the (wavelet-noiselet) and (Fourier-noiselet) systems where wavelet could either be the Haar or Daubechies D4/D8 variants. Proceeding similarly as in Lemma 7 and then Corollary 1 finishes the proof.

E Handling unmodelled errors with APIS

Recall that in this case, we modify Eq. (1) to include an unmodelled error term.

$$\mathbf{y} = \tilde{\mathbf{a}} + \mathbf{b}^* + \mathbf{e}^*,$$

where $\tilde{\mathbf{a}} \in \mathscr{A}, \mathbf{b}^* \in \mathscr{B}$ and $\mathbf{a}^* = \tilde{\mathbf{a}} + \mathbf{e}^*$. We make no assumptions on $\mathbf{e}^*$ such as requiring it to belong to any union of subspaces etc. $\mathbf{e}^*$ can be completely arbitrary; in particular it can be dense $\|\mathbf{e}^*\|_0 = n$ and need not have a sparse representation in any particular basis. A useful case is when $\tilde{\mathbf{a}}$ can be taken to be the best approximation of $\mathbf{a}^*$ in the union of subspaces $\mathscr{A}$. Below, we offer a recovery guarantee for APIS in this case. As in Appendix A, we will first present the main proof ideas with the special case of $P = 1$ (the so-called *known signal support* case (Chen and De 2020)) where the union $\mathscr{A}$ consists of a single subspace A. We will then extend the proof to the general case where both $P, Q \geq 1$. Recall that we denote using P (resp. Q), the number of subspaces in the union $\mathscr{A} = \bigcup_{i=1}^P A_i$ in which the signal $\tilde{\mathbf{a}}$ resides (resp the union $\mathscr{B} = \bigcup_{j=1}^Q B_j$ in which the corruption $\mathbf{b}^*$ resides).

E.1 Convergence analysis for $P = 1$ i.e. $\mathscr{A} = A$ but still $Q \geq 1$

We now present the proof in the case of known signal support.

Lemma 9 *Suppose we obtain data as described in Eq. (2) where the two unions $\mathscr{A}, \mathscr{B}$ are μ-SU incoherent with $\mu < \frac{1}{3}$ and in addition, the union $\mathscr{A}$ contains a single subspace (the known signal support model). Then, for any $\epsilon > 0$ within $T = \mathscr{O}\left(\log \frac{\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2}{\epsilon}\right)$ iterations,* APIS *offers* $\left\|\mathbf{a}^T - \tilde{\mathbf{a}}\right\|_2 \leq \epsilon + \frac{4\sqrt{\mu}}{1-3\mu} \cdot \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{e}^*)\right\|_2 + 2 \cdot \left\|\Pi_A(\mathbf{e}^*)\right\|_2$.

Proof As in the proof of Lemma 1, denote $\mathbf{p} = \Pi_A(\mathbf{b}^* - \mathbf{b})$ and $\mathbf{p}^+ = \Pi_A(\mathbf{b}^* - \mathbf{b}^+)$. Let $\mathfrak{Q} := B^+ \cap B^*$ denote the meet of the two subspaces, as well as denote the symmetric difference subspaces $\mathfrak{P} := B^+ \cap (B^*)^\perp$ and $\mathfrak{R} = B^* \cap (B^+)^\perp$ (recall that $A \ni \mathbf{a}^*, B^* \ni \mathbf{b}^*$). We also use the shorthand $\mathbf{r} := \mathbf{e}^* - \Pi_A(\mathbf{e}^*)$. In this case, we have $\mathbf{a}^+ = \Pi_A(\tilde{\mathbf{a}} + \mathbf{b}^* + \mathbf{e}^* - \mathbf{b}^+)$. Since $\Pi_A(\tilde{\mathbf{a}}) = \tilde{\mathbf{a}}$, we get $\mathbf{a}^+ - \tilde{\mathbf{a}} = \Pi_A(\mathbf{b}^* - \mathbf{b}^+ + \mathbf{e}^*)$. Applying the triangle inequality gives us $\|\mathbf{a}^+ - \tilde{\mathbf{a}}\|_2 \leq \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2 + \left\|\Pi_A(\mathbf{e}^*)\right\|_2 = \|\mathbf{p}^+\|_2 + \left\|\Pi_A(\mathbf{e}^*)\right\|_2$. Now, we have

$$\mathbf{b}^+ = \Pi_{B^+}(\tilde{\mathbf{a}} + \mathbf{b}^* + \mathbf{e}^* - \mathbf{a}) = \Pi_{B^+}(\mathbf{b}^* + \mathbf{e}^* - \Pi_A(\mathbf{b}^* + \mathbf{e}^* - \mathbf{b})) = \Pi_{B^+}(\mathbf{b}^* - \mathbf{p} + \mathbf{r}),$$

and thus $\mathbf{b}^* - \mathbf{b}^+ = \Pi_{\mathfrak{R}}(\mathbf{b}^*) + \Pi_{B^+}(\mathbf{p} - \mathbf{r})$. The triangle inequality then gives us

$$\left\|\mathbf{p}^+\right\|_2 = \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2 \leq \left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 + \left\|\Pi_A(\Pi_{B^+}(\mathbf{p} - \mathbf{r}))\right\|_2$$

Now, if we denote $\mathbf{z} = \mathbf{b}^* - \mathbf{p} + \mathbf{r}$, the projection step assures us, as before, that,

$$\left\|\Pi_{\mathfrak{R}}(\mathbf{z})\right\|_2^2 \leq \left\|\Pi_{\mathfrak{P}}(\mathbf{z})\right\|_2^2 = \left\|\Pi_{\mathfrak{P}}(\mathbf{p} - \mathbf{r})\right\|_2^2$$

since $\Pi^\perp_{B^*}(\mathbf{b}^*) = \mathbf{0}$. Going as before gives us

$$\left\|\Pi_{\mathfrak{R}}(\mathbf{b}^*)\right\|_2 \leq \left\|\Pi_{B^+}(\mathbf{p} - \mathbf{r})\right\|_2 + \left\|\Pi^*_B(\mathbf{p} - \mathbf{r})\right\|_2$$

Applying incoherence results now tells us that

$$\begin{aligned}\left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 \leq \sqrt{\mu} \cdot \left\|\Pi_{\mathfrak{R}}(\mathbf{b}^*)\right\|_2 &= \sqrt{\mu}(\left\|\Pi_{B^+}(\mathbf{p} - \mathbf{r})\right\|_2 + \left\|\Pi^*_B(\mathbf{p} - \mathbf{r})\right\|_2) \\ &\leq 2\mu\|\mathbf{p}\|_2 + 2\sqrt{\mu} \cdot \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{r})\right\|_2\end{aligned}$$

Putting things together gives us

$$\begin{aligned}\left\|\mathbf{p}^+\right\|_2 &\leq \left\|\Pi_A(\Pi_{\mathfrak{R}}(\mathbf{b}^*))\right\|_2 + \left\|\Pi_A(\Pi_{B^+}(\mathbf{p} - \mathbf{r}))\right\|_2 \\ &\leq 3\mu\|\mathbf{p}\|_2 + 3\sqrt{\mu} \cdot \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{r})\right\|_2 \\ &\leq 3\mu\|\mathbf{p}\|_2 + 3\sqrt{\mu} \cdot \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{e}^*)\right\|_2,\end{aligned}$$

where the last step follows since $\mathbf{r} = \Pi^\perp_A(\mathbf{e}^*)$ and projections are always non-expansive. Now, APISinitializes $\mathbf{a}^0 = \mathbf{0}$ which means that initially, we have (using $\mathbf{a}^* = \tilde{\mathbf{a}} + \mathbf{e}^*$)

$$\mathbf{p}^1 = \Pi_A(\mathbf{b}^* - \mathbf{b}) = \Pi_A(\mathbf{b}^* - \Pi_{B^+}(\mathbf{a}^* + \mathbf{b}^*))$$

and thus, $\left\|\mathbf{p}^1\right\|_2 \leq \|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2$ since projections are always non-expansive. Thus, if $\mu < \frac{1}{3}$, then the linear rate of convergence implies that within $T = \mathscr{O}\left(\log \frac{\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2}{\epsilon}\right)$ iterations, we will have

$$\left\|\mathbf{p}^T\right\|_2 \leq \epsilon + \frac{4\sqrt{\mu}}{1 - 3\mu} \cdot \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{e}^*)\right\|_2 + \left\|\Pi_A(\mathbf{e}^*)\right\|_2.$$

Using our earlier observation $\left\|\mathbf{a}^T - \tilde{\mathbf{a}}\right\|_2 = \left\|\mathbf{p}^T\right\|_2 + \left\|\Pi_A(\mathbf{e}^*)\right\|_2$ then finishes the proof.

E.2 Application to simultaneous sparse corruptions and dense Gaussian noise case

Consider the robust linear regression problem with the true linear model being $\mathbf{w}^* \in \mathbb{R}^d$ where, apart from k adversarially corrupted points, all n points get Gaussian noise i.e. $\mathbf{y} = X^\top \mathbf{w}^* + \mathbf{b}^* + \mathbf{e}^*$, where $\mathbf{e}^* \sim \mathscr{N}(\mathbf{0}, \sigma^2 \cdot I_n)$. It is easy to see that for any fixed r-dimensional subspace S, we have $\left\|\Pi_S(\mathbf{e}^*)\right\|_2 \leq \mathscr{O}\left(\sqrt{r}\right)$. Thus, $\left\|\Pi_A(\mathbf{e}^*)\right\|_2 \leq \mathscr{O}\left(\sqrt{d}\right)$ and taking a union bound over all $\begin{pmatrix} n \\ k \end{pmatrix}$ subspaces of k-sparse vectors tells us that $\max_{B\in\mathscr{B}} \left\|\Pi_B(\mathbf{e}^*)\right\|_2 \leq \mathcal{O}\left(\sqrt{k\log n}\right)$.

Lemma 9 shows that within $T = \mathcal{O}(\log n)$ iterations, APIS guarantees a model vector $\mathbf{w}^T$ such that $\|X\mathbf{w}^t - X\mathbf{w}^*\|_2 \leq \mathcal{O}\left(\sqrt{d} + \sqrt{k\log n}\right)$. Using $\mathbf{w}^T - \mathbf{w}^* = X^\dagger(X\mathbf{w}^t - X\mathbf{w}^*)$ and the lower bounds on the eigenvalues of $XX^\top$ from the proof of Lemma 3 tell us that $\left\|\mathbf{w}^T - \mathbf{w}^*\right\|_2 = \mathcal{O}\left(\frac{\sqrt{d}+\sqrt{k\log n}}{\sqrt{n}}\right)$. Squaring both sides tells us that $\left\|\mathbf{w}^T - \mathbf{w}^*\right\|_2^2 \leq \mathcal{O}\left(\sigma^2\left(\frac{(d+k)\ln n}{n}\right)\right)$.

Note that as $n \to \infty$, the above model recovery error behaves as $\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \mathcal{O}(k\log n/n)$. This guarantees consistent recovery if $k\log n/n \to 0$ as $n \to \infty$. This is a sharper result than previous works (Bhatia et al. 2015; Mukhoty et al. 2019) that do not offer consistent estimation even if $k\log n/n \to 0$.

E.3 Applicability to robust non-parametric kernel ridge regression

The above results are also useful when applying APIS to robust kernel ridge regression. In several cases, the function (signal) we are trying to approximate need not be exactly represented in terms of the top s eigenvectors of the Gram matrix (see Sect. 3). However, Lemma 9 shows that APIS still offers recovery of the s-sparse representation of the signal in terms of the top-s eigenvectors. As discussed in Sect. 6.5, this still constitutes a universal model in the limit, and experiments in Sect. 7 show that APIS offers excellent reconstruction even under adversarial corruptions for sinusoids, polynomials, and their combinations.

E.4 Convergence analysis for general case i.e both $P, Q \geq 1$

We now present the proof in the general case.

Lemma 10 *Suppose we obtain data as described in Eq. (2) where the two unions $\mathscr{A}, \mathscr{B}$ are μ-SU incoherent with $\mu < \frac{1}{9}$. Then, for any $\epsilon > 0$ within $T = \mathcal{O}\left(\log \frac{\|\mathbf{a}^*\|_2+\|\mathbf{b}^*\|_2}{\epsilon}\right)$ iterations,* APIS *offers* $\left\|\mathbf{a}^T - \tilde{\mathbf{a}}\right\|_2 \leq \epsilon + \mathcal{O}\left(\max_{B\in\mathscr{B}} \left\|\Pi_B(\mathbf{e}^*)\right\|_2 + \left\|\Pi_A(\mathbf{e}^*)\right\|_2\right)$.

Proof The analysis in the general case proceeds by extending the proof of Lemma 9 in a manner similar to how Lemma 2 extended the proof of Lemma 1. We define the quantities $p := \max_{A\in\mathscr{A}} \left\|\Pi_A(\mathbf{b}^* - \mathbf{b})\right\|_2, p^+ := \max_{A\in\mathscr{A}} \left\|\Pi_A(\mathbf{b}^* - \mathbf{b}^+)\right\|_2$ and correspondingly $q := \max_{B\in\mathscr{B}} \left\|\Pi_B(\tilde{\mathbf{a}} - \mathbf{a})\right\|_2, q^+ := \max_{B\in\mathscr{B}} \left\|\Pi_B(\tilde{\mathbf{a}} - \mathbf{a}^+)\right\|_2$ as in the proof of Lemma 2 and introduce two new notations $u = \max_{A\in\mathscr{A}} \left\|\Pi_A(\mathbf{e}^*)\right\|_2$ and $v = \max_{B\in\mathscr{B}} \left\|\Pi_B(\mathbf{e}^*)\right\|_2$. We get the following results

$$\begin{aligned} \left\|\mathbf{a}^+ - \tilde{\mathbf{a}}\right\|_2 &\leq 3p^+ + 3u \\ p &\leq 3\sqrt{\mu}(q + u) \\ q^+ &\leq 3\sqrt{\mu}(p + v) \end{aligned}$$

The second result above can be rewritten as $p^+ \leq 3\sqrt{\mu}(q^+ + u)$ which gives us $p^+ \leq 9\mu \cdot p + (9\mu \cdot v + 3\sqrt{\mu} \cdot u)$. Thus, we continue to get a linear rate of convergence whenever $\mu < \frac{1}{9}$ and, since $p^1 \leq \mathcal{O}\big(\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2\big)$, after $T = \mathcal{O}\Big(\log \frac{\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2}{\epsilon}\Big)$ iterations get $p^T \leq \frac{\epsilon}{3}$. Since $\left\|\mathbf{a}^T - \tilde{\mathbf{a}}\right\|_2 \leq 3p^T + 3u$ from above, we get

$$\left\|\mathbf{a}^+ - \mathbf{a}^*\right\|_2 \leq \epsilon + 10\mu \cdot v + 4\sqrt{\mu} \cdot u + 3u \leq \epsilon + 5 \cdot \max_{A \in \mathscr{A}} \left\|\Pi_A(\mathbf{e}^*)\right\|_2 + 2 \cdot \max_{B \in \mathscr{B}} \left\|\Pi_B(\mathbf{e}^*)\right\|_2,$$

which finishes the proof.

E.5 Application to recovery of compressible signals

The above result has applications in, for example, the sparse signal transform example, where $\mathbf{e}^*$ may model components of the signal not captured in the low-rank model. For instance, the signal may not come entirely from any single rank-s subspace $A \in \mathscr{A}$ but merely have *most* of its weight concentrated on a single rank-s subspace $A^* \in \mathscr{A}$. $\mathbf{e}^*$ would then model the component of the signal orthogonal to A^*.

Consider an image $\mathbf{a}^*$ that is not wavelet-sparse, but (s, ϵ)-approximately wavelet sparse meaning that there exists an image $\tilde{\mathbf{a}}$ that is s wavelet-sparse, and $\|\mathbf{a}^* - \tilde{\mathbf{a}}\|_2 \leq \epsilon \cdot \|\mathbf{a}^*\|_2$. In particular, $\tilde{\mathbf{a}}$ can be taken to be the best s wavelet-sparse approximation of $\mathbf{a}^*$. This means that $\|\mathbf{e}^*\|_2 \leq \epsilon \cdot \|\mathbf{a}^*\|_2$. Lemma 10 shows that APIS offers a recovery of $\tilde{\mathbf{a}}$ to within $\mathcal{O}\big(\epsilon \cdot \|\mathbf{a}^*\|_2\big)$ error within $T = \mathcal{O}\Big(\log \frac{\|\mathbf{a}^*\|_2 + \|\mathbf{b}^*\|_2}{\epsilon \cdot \|\mathbf{a}^*\|_2}\Big) = \mathcal{O}\Big(\log \frac{1}{\epsilon} + \log \frac{\|\mathbf{b}^*\|_2}{\epsilon \|\mathbf{a}^*\|_2}\Big)$ iterations.

F Handling lack of incoherence with APIS

The results of Theorem 1 and Lemmas 7 and 8 rely on the notion of incoherence described in Sect. 6.1. However, in certain situations, the bases in question are not incoherent. For example, when the signal is s-sparse in the Haar wavelet basis and the corruption is k-sparse in the Fourier basis, it precludes the recovery guarantee offered by APIS.

In the following, we sketch an argument, taking the (Haar-Fourier) case as an example, to show, when the signal offers more structure, *local* incoherence can still be guaranteed and APIS, with suitable modifications made to the signal projection step $\Pi_{\mathscr{A}}(\cdot)$ to exploit this additional structure (discussed later), can offer exact recovery at a linear rate.

As before, let U, V denote the design matrices corresponding to the signal and corruption bases. As before, calculating the SU-incoherence constant μ requires us to bound

$$\max_{\substack{\mathbf{u} \in S^{n-1}, \|\mathbf{u}\|_0 \leq s \\ \mathbf{v} \in S^{n-1}, \|\mathbf{v}\|_0 \leq k}} \left(\mathbf{u}^\top U^\top V \mathbf{v}\right)^2$$

Since $\mathbf{u}, \mathbf{v}$ are sparse vectors, we have $\mathbf{u}^\top U^\top V \mathbf{v} \leq \left\|U_S^\top V_K\right\|_2$ where $S = \operatorname{supp}(\mathbf{u})$ and $K = \operatorname{supp}(\mathbf{v})$ are the supports of $\mathbf{u}, \mathbf{v}$. Now, the proof strategy in Lemma 8 fails here since for the Haar-wavelet pair, we get $\left\|U_S^\top V_K\right\|_\infty =: \nu = 1$ where $S = \operatorname{supp}(\mathbf{u})$ and $K = \operatorname{supp}(\mathbf{v})$ are the supports of $\mathbf{u}, \mathbf{v}$ respectively (see, for example (Zhou et al. 2016) for this lack of incoherence result).

This happens because there are individual basis vectors in the Haar and Fourier bases, say $\mathbf{m}, \mathbf{n}$ whose inner product is unity i.e. $|\langle \mathbf{m}, \mathbf{n} \rangle| = 1$ i.e. $\mathbf{m} = \pm \mathbf{n}$. This allows a situation where there is a signal that is just 1-sparse in the Haar basis, specifically $\mathbf{a}^* = c \cdot \mathbf{m}$ for some $c \in \mathbb{R}$, and the signal then gets corrupted by a corruption vector that is again just 1-sparse in the Fourier basis, specifically $\mathbf{b}^* = d \cdot \mathbf{n}$ for some $d \in \mathbb{R}$. Exact recovery is information theoretically impossible since the algorithm would essentially receive $\mathbf{y} = \mathbf{a}^* + \mathbf{b}^* = (c \pm d) \cdot \mathbf{m} = (c \pm d) \cdot \mathbf{n}$ with no way of separating c and d (we use ± since $\mathbf{m}, \mathbf{n}$ could be parallel or anti-parallel depending on convention).

F.1 Structured anti-concentrated signals

It turns out that one way to avoid the above problem is to ensure that our signal does not concentrate its mass on just a few coordinates (this prevents the signal from being 1-sparse). Although several ways may exist to enforce the above, in the following definitions, we present the notions of anti-concentrated signal with *stratified* sparsity. Specifically, suppose the uncorrupted signal is $\mathbf{a}^* = U\mathbf{u}$ with U being the design matrix of the Haar wavelet transformation.

Definition 3 A signal $\mathbf{a}^* = U\mathbf{u} \in \mathbb{R}^n$ is said to be (γ, s) anti-concentrated if it is s-sparse i.e. $\|\mathbf{u}\|_0 \leq s$, and there exists some $\gamma > 0$, such that $\|\mathbf{u}\|_\infty \leq \frac{\gamma}{\sqrt{s}} \cdot \|\mathbf{u}\|_2$.

Note that, in general, all s-sparse vectors are at least $(\sqrt{s}, s)$-anti-concentrated. However, a $(\sqrt{s}, s)$-anti concentrated signal is allowed to put almost all its weight on a single coordinate. In contrast, the most anti-concentrated s-sparse vector, for which all s coordinates have equal magnitude, would be $(1, s)$-anti-concentrated. Before presenting the notion of stratified sparsity, we need to introduce the notion of strata for the Haar basis. The Haar basis elements can be arranged into $\log n$-many *strata* with the i^{th} stratum containing $n_i = 2^i$ basis elements (see the proof of Lemma 11 below for details).

Definition 4 A signal $\mathbf{a}^* = U\mathbf{u} \in \mathbb{R}^n$ is said to be α-*stratified sparse* if for some $\alpha \in (0, 1)$, the support of $\mathbf{u}$ is such that the i^{th} stratum of the Haar basis contains at most $(n_i)^\alpha$ support elements of $\mathbf{u}$. Note that this implies that the vector $\mathbf{u}$ is s-sparse with $s \leq n^\alpha$ as well (although it need not necessarily be anti-concentrated as required by Def. 3).

F.2 Local incoherence with structured anti-concentrated signals

Given signals with additional structure as described above in Defs 3 and 4, the following result shows how local incoherence still continues to hold. Note that the following result starts giving vacuous results ($\mu \to 1$) for (γ, s)-anti concentrated vectors, as $\gamma \to \sqrt{s}$. This is as expected since $\gamma \approx \sqrt{s}$ allows signals that are very concentrated e.g. being close to being 1-sparse.

Lemma 11 *Suppose the set of signals $\mathscr{A}$ is the set of s-sparse (w.r.t Haar basis), α-stratified and (γ, s)-anti-concentrated signals with $\alpha \in (0, 1), s = n^\alpha, \gamma \in [1, \sqrt{s}]$. Then, with respect to $\mathscr{B}$ being the set of k-sparse corruption vectors (no further assumptions being imposed* *on corruption vectors), for some small universal constant* $c > 0$*, the following (local) incoherence bound continues to hold.*

$$\mu \leq c \cdot \gamma^2 \cdot \begin{cases} \frac{k^{2+4\alpha}}{s} & \alpha < \frac{1}{2} \\ \frac{k^{2+4\alpha}}{s} + \frac{k^2}{s} \log^2 \frac{n}{k^2} & \alpha = \frac{1}{2} \\ \frac{k^{2+4\alpha}}{s} + \frac{sk^2}{n} & \alpha > \frac{1}{2} \end{cases}$$

Proof Calculating th SU-incoherence constant μ now requires us to bound

$$\max_{\substack{\mathbf{u} \in S^{n-1}, \|\mathbf{u}\|_0 \leq s \\ \mathbf{u} \text{ is } \alpha-\text{strat.}, (\gamma, s)-\text{anti-conc.} \\ \mathbf{v} \in S^{n-1}, \|\mathbf{v}\|_0 \leq k}} \left(\mathbf{u}^\top U^\top V \mathbf{v}\right)^2$$

Then, applying the $L_1 - L_\infty$ Hölder's inequality gives us

$$\left|\mathbf{u}^\top U^\top V \mathbf{v}\right| = \left|\sum_{i \in S} \sum_{j \in K} \mathbf{u}_i \mathbf{v}_j (U^\top V)_{ij}\right| \leq \max_{i \in S, j \in K} \left|\mathbf{u}_i \mathbf{v}_j\right| \cdot \sum_{i \in S} \sum_{j \in K} \left|(U^\top V)_{ij}\right| \leq \gamma k \sqrt{s} \cdot \bar{v}_{s,k},$$

where $(U^\top V)_{ij} = \mathbf{u}_i^\top \mathbf{v}_j$ and $\bar{v}_{s,k}$ is the largest average value of entries in the matrix $U_S^\top V_K$ for any choice of sets S, K of size s, k respectively i.e.

$$\bar{v}_{s,k} = \max_{\substack{S, K \subset [n] \\ |S| = s, |K| = k}} \frac{1}{sk} \sum_{i \in S} \sum_{j \in K} \left|(U^\top V)_{ij}\right|$$

The above step is perhaps the most crucial in the proof since it shows that the incoherence constant μ *depends on the average of the* $\left|(U^\top V)_{ij}\right|$ *values rather than the largest values, which are always* $\Omega(1)$ *for the Haar-Fourier system*.

The result of Krahmer and Ward (2014, Lemma 6.1) shows that upon indexing the Haar basis elements by $i \in [1, \log n - 1]$ into the $\log n$ strata and further indexing the 2^i basis elements within the i^{th} stratum using $l \in [0, 2^i - 1]$, as well as indexing the Fourier basis elements by $j \in \left[-\frac{n}{2} + 1, \frac{n}{2}\right] \backslash \{0\}$, we get a *local* incoherence bound

$$\left|\mathbf{u}_{i,l}^\top \mathbf{v}_j\right| \leq \min\left\{\frac{6 \cdot 2^{\frac{i}{2}}}{|j|}, 3\pi \cdot 2^{-\frac{i}{2}}\right\} \leq \mathcal{O}\left(\min\left\{\frac{2^{\frac{i}{2}}}{|j|}, 2^{-\frac{i}{2}}\right\}\right)$$

Noting that $2^{-\frac{i}{2}} \leq \frac{2^{\frac{i}{2}}}{j}$ iff $i \geq 2 \log j$, elementary calculations show that

$$\bar{v}_{s,k} \leq \mathcal{O}\left(\frac{1}{sk} \sum_{j=1}^{k} \left(\sum_{i=1}^{2 \log j} 2^{i\alpha} \cdot \frac{2^{\frac{i}{2}}}{j} + \sum_{i=2 \log j}^{\log n} 2^{i\alpha} \cdot 2^{-\frac{i}{2}}\right)\right)$$

If $\alpha < \frac{1}{2}$, the second summation is that of a decreasing series. Thus, the second summation can be upper bounded in this case as

$$\sum_{i=2 \log j}^{\log n} 2^{i\left(\alpha - \frac{1}{2}\right)} \leq \mathcal{O}\left(2^{2 \log j \left(\alpha - \frac{1}{2}\right)}\right) \leq \mathcal{O}\left(j^{2\alpha - 1}\right)$$

If $\alpha = \frac{1}{2}$ then we have a much simpler summation

$$\sum_{i=2\log j}^{\log n} 2^0 \le \mathcal{O}\left(\log \frac{n}{j^2}\right)$$

If $\alpha > \frac{1}{2}$, the second summation is that of an increasing series. Thus, the second summation can be upper bounded in this case as

$$\sum_{i=2\log j}^{\log n} 2^{i\left(\alpha-\frac{1}{2}\right)} \le \mathcal{O}\left(2^{\log n\left(\alpha-\frac{1}{2}\right)}\right) \le \mathcal{O}\left(n^{\alpha-\frac{1}{2}}\right) = \mathcal{O}\left(\frac{s}{\sqrt{n}}\right)$$

Thus, ignoring constant factors, we get

$$\sum_{i=2\log j}^{\log n} 2^{i\left(\alpha-\frac{1}{2}\right)} \le \begin{cases} j^{2\alpha-1} & \alpha < \frac{1}{2} \\ \log \frac{n}{j^2} & \alpha = \frac{1}{2} \\ \frac{s}{\sqrt{n}} & \alpha > \frac{1}{2} \end{cases}$$

This gives us

$$\sum_{j=1}^{k} \sum_{i=2\log j}^{\log n} 2^{i\left(\alpha-\frac{1}{2}\right)} \le \begin{cases} k^{2\alpha} & \alpha < \frac{1}{2} \\ k \log \frac{n}{k^2} & \alpha = \frac{1}{2} \\ \frac{sk}{\sqrt{n}} & \alpha > \frac{1}{2} \end{cases}$$

Similarly, the first summation can be bounded, ignoring constant factors, as

$$\sum_{j=1}^{k} \left(\frac{1}{j} \cdot \sum_{i=1}^{2\log j} 2^{i\left(\alpha+\frac{1}{2}\right)}\right) \le \sum_{j=1}^{k} \left(\frac{1}{j}\left(2^{2\log j\left(\alpha+\frac{1}{2}\right)}\right)\right) \le \sum_{j=1}^{k} \left(\frac{1}{j} \cdot j^{2\alpha+1}\right) \le k^{2\alpha+1}$$

Absorbing all constant factors into a single constant $c > 0$ gives us

$$\bar{v}_{s,k} \le \frac{c}{sk}\left(k^{2\alpha+1} + \begin{cases} k^{2\alpha} & \alpha < \frac{1}{2} \\ k \log \frac{n}{k^2} & \alpha = \frac{1}{2} \\ \frac{sk}{\sqrt{n}} & \alpha > \frac{1}{2} \end{cases}\right) = \frac{1}{s} \cdot \begin{cases} k^{2\alpha} & \alpha < \frac{1}{2} \\ k^{2\alpha} + \log \frac{n}{k^2} & \alpha = \frac{1}{2} \\ k^{2\alpha} + \frac{s}{\sqrt{n}} & \alpha > \frac{1}{2} \end{cases}$$

which in turn gives us (renaming $c^2 =: c$),

$$\mu \le \left(\gamma k \sqrt{s} \cdot \bar{v}_{s,k}\right)^2 \le c \cdot \gamma^2 \cdot \begin{cases} \frac{k^{2+4\alpha}}{s} & \alpha < \frac{1}{2} \\ \frac{k^{2+4\alpha}}{s} + \frac{k^2}{s} \log^2 \frac{n}{k^2} & \alpha = \frac{1}{2} \\ \frac{k^{2+4\alpha}}{s} + \frac{sk^2}{n} & \alpha > \frac{1}{2} \end{cases} .$$

Thus, we do have incoherence when $k \ll s$ as well as $sk \ll n$. We can get a stronger result if the corruption is also assured to be anti concentrated. Specifically $\mathbf{b}^* = V\mathbf{v}$ where $\mathbf{v}$ is k-sparse as well as (δ, k) anti-concentrated for some $\delta \in [1, \sqrt{k}]$. We present this improved result below and note that it offers superior dependence on k due to the additional structure in the corruption vector.

$$\mu \leq c \cdot \gamma^2 \delta^2 \cdot \begin{cases} \frac{k^{1+4\alpha}}{s} & \alpha < \frac{1}{2} \\ \frac{k^{1+4\alpha}}{s} + \frac{k}{s} \log^2 \frac{n}{k^2} & \alpha = \frac{1}{2} \\ \frac{k^{1+4\alpha}}{s} + \frac{sk}{n} & \alpha > \frac{1}{2} \end{cases}$$

This finishes the proof.

F.3 Algorithmic modifications to APIS

Since signals now have additional structure, specifically, anti-concentration and stratified sparsity, we need to modify the projection step $\Pi_{\mathscr{A}}(\cdot)$ appropriately to handle both properties. Fortunately, simple modifications to the hard-thresholding operator address both.

Algorithm 3 Projection onto bounded, stratified, Haar-sparse vectors

```
Input: A vector $\mathbf{z}$, Haar basis $U$, bounds for sparsity $s$, stratification $\alpha$, and sup-norm $M$
Output: A vector $\hat{\mathbf{z}}$ that is $s$ Haar-sparse, $\alpha$-stratified, has $\|U^\top \hat{\mathbf{z}}\|_\infty \leq M$ and minimizes $\|\mathbf{z} - \hat{\mathbf{z}}\|_2^2$
$\mathbf{v} \leftarrow U^\top \mathbf{z}$                                   //Inverse Haar transform
Break up $\mathbf{v}$ into $\log n$ vectors $\mathbf{v}^1, \ldots, \mathbf{v}^{\log n}$, corresponding the $\log n$ strata of the Haar basis
for $i = 1, 2, \ldots, \log n$ do
    $\hat{\mathbf{v}}^i \leftarrow \mathrm{BHT}(\mathbf{v}^i, 2^{\alpha i}, M)$                 //Apply Bounded Hard Thresholding
end for
Concatenate $\left\{\hat{\mathbf{v}}^i : i \in [1, \log n]\right\}$ back into a single vector $\hat{\mathbf{v}}$
return $U\hat{\mathbf{v}}$
```

Algorithm 4 Bounded Hard Thresholding BHT

```
Input: A vector $\mathbf{r} \in \mathbb{R}^n$, sparsity $t$, sup-norm bound $M$
Output: A vector $\hat{\mathbf{r}} \in \mathbb{R}^n$ that is $t$-sparse, has $\|\hat{\mathbf{r}}\|_\infty \leq M$ and minimizes $\|\mathbf{r} - \hat{\mathbf{r}}\|_2^2$
Create a new vector $\mathbf{m} \in \mathbb{R}^n$ with $\mathbf{m}_i = |\mathbf{r}_i| - \min\{|\mathbf{r}_i|, M\}$ for all $i \in [n]$
Create a new vector $\mathbf{d} \in \mathbb{R}^n$ with $\mathbf{d}_i = \sqrt{\mathbf{r}_i^2 - \mathbf{m}_i^2}$          //Discounted magnitudes
Let $S \subset [n]$ denote the set of $t$ coordinates with largest values of $\mathbf{d}_i$
Create a new vector $\hat{\mathbf{r}}$ with $\hat{\mathbf{r}}_i = \min\{|\mathbf{r}_i|, M\} \cdot \mathrm{sign}(\mathbf{r}_i)$ for $i \in S$ and $\hat{\mathbf{r}}_j = 0$ for $j \notin S$
return $\hat{\mathbf{r}}$
```

Algorithm 3 gives the recipe to perform projections onto vectors that are s-sparse in the Haar basis, as well as stratified and sup-norm bounded (to ensure anti-concentration). The sup-norm bound M is a new hyperparameter in the algorithm and can be tuned according to the hyperparameter tuning procedure outlined in Sect. 7. After performing an inverse Haar transform, Algorithm 3 breaks up the resulting vector into the $\log n$ strata offered by the Haar basis and performs *Bounded Hard Thresholding* (BHT) on each stratum separately.

Algorithm 4 presents BHT, a modified hard thresholding operation that admits the sup-norm restriction in addition to the sparsity restriction. Instead of the traditional hard-thresholding operator HT (see Sect. 4) which simply selects the top t coordinates according to magnitude, BHT instead uses the *discounted* magnitude of each coordinate to do so. The discounted magnitude d of a value $v \in \mathbb{R}$ given a sup-norm bound $M > 0$ is defined as

$$d = \sqrt{v^2 - (|v| - \min\{|v|, M\})^2}$$

Note that if there is no sup-norm bound (equivalently if $M = \infty$), then the discounted magnitude is simply the magnitude i.e. $d = |v|$. Thus, in the absence of an sup-norm bound, BHT becomes simply HT.

To prove the optimality of Algorithm 3 it is sufficient to prove the optimality of the BHT procedure since Algorithm 3 simply applies it in a stratum-wise manner. We prove the optimality of BHT below.

Theorem 2 *For any vector* $\mathbf{r} \in \mathbb{R}^n, t \in [n], M > 0$*, let* $\mathbf{p} = \mathrm{BHT}(\mathbf{r}, t, M)$ *(see Algorithm 4). Then* $\mathbf{p}$ *is t-sparse and satisfies* $\|\mathbf{p}\|_\infty \leq M$*. Moreover, let* $\mathbf{q} \in \mathbb{R}^n$ *be any vector that is also t-sparse and satisfies* $\|\mathbf{q}\|_\infty \leq M$*. Then we must have* $\|\mathbf{r} - \mathbf{p}\|_2^2 \leq \|\mathbf{r} - \mathbf{q}\|_2^2$ *i.e. BHT does provide the optimal projection onto sup-norm bounded sparse vectors.*

Proof That $\mathbf{p}$ is t-sparse and satisfies $\|\mathbf{p}\|_\infty \leq M$ is immediate from the steps taken by Algorithm 4. To prove the second part, let $S = \mathrm{supp}\,(\mathbf{p}), T = \mathrm{supp}\,(\mathbf{q})$ be the support of the two vectors. Assume w.l.o.g. that $|S| = t = |T|$. Now, we create a third vector $\mathbf{k}$ with the same support as $\mathbf{q}$ but with possibly different values. Specifically, set $\mathbf{k}_j = \min\left\{\left|\mathbf{r}_j\right|, M\right\} \cdot \mathrm{sign}\left\{\mathbf{r}_j\right\}$ for $j \in T$ and $\mathbf{k}_j = 0$ for $j \notin T$. Notice that $\mathbf{k}$ is also t-sparse, $\mathrm{supp}\,(\mathbf{k}) = T$, and it satisfies $\|\mathbf{k}\|_\infty \leq M$ as well.

It is easy to see that $\|\mathbf{r} - \mathbf{k}\|_2^2 \leq \|\mathbf{r} - \mathbf{q}\|_2^2$ which captures our intuition that once we have chosen a t-sized support for our vector, the ideal thing to do is to fill coordinates in the support with the value $\min\left\{\left|\mathbf{r}_j\right|, M\right\} \cdot \mathrm{sign}\,(\mathbf{r}_j)$ (with the absolute value and sign operations being applied component-wise) which maximally preserves the vector in that coordinate subject to the sup-norm bound.

Now we prove that the choice of support made by BHT is optimal by showing that $\|\mathbf{r} - \mathbf{p}\|_2^2 \leq \|\mathbf{r} - \mathbf{k}\|_2^2$. To see this, we consider the following sequence of inequalities. We will find the shorthand $\mathbf{m}_i := |\mathbf{r}_i| - \min\left\{|\mathbf{r}_i|, M\right\}$ very useful in the following. This is because

$$\mathrm{sign}\,(\mathbf{r}_j) \cdot \mathbf{m}_j = \mathbf{r}_j - \min\left\{\left|\mathbf{r}_j\right|, M\right\} \cdot \mathrm{sign}\,(\mathbf{r}_j)$$

is simply the residual error at any coordinate that is in the support of either $\mathbf{p}$ or $\mathbf{k}$. Note that we have

$$\|\mathbf{r} - \mathbf{k}\|_2^2 = \sum_{i \in T} \mathbf{m}_i^2 + \sum_{j \in S \setminus T} \mathbf{r}_j^2 + \sum_{l \notin S \cup T} \mathbf{r}_l^2$$
$$\|\mathbf{r} - \mathbf{p}\|_2^2 = \sum_{i \in S} \mathbf{m}_i^2 + \sum_{j \in T \setminus S} \mathbf{r}_j^2 + \sum_{l \notin S \cup T} \mathbf{r}_l^2$$

This gives us

$$\|\mathbf{r} - \mathbf{k}\|_2^2 - \|\mathbf{r} - \mathbf{p}\|_2^2 = \sum_{i \in S \setminus T} (\mathbf{r}_i^2 - \mathbf{m}_i^2) - \sum_{j \in T \setminus S} (\mathbf{r}_j^2 - \mathbf{m}_j^2) = \sum_{i \in S \setminus T} \mathbf{d}_i^2 - \sum_{j \in T \setminus S} \mathbf{d}_j^2,$$

where $\mathbf{d}_i = \sqrt{\mathbf{r}_i^2 - \mathbf{m}_i^2}$ is the discounted magnitude of the i^{th} coordinate as defined above. However, since BHT always chooses the t coordinates with highest discounted magnitude, we must have $\sum_{i \in S \setminus T} \mathbf{d}_i^2 \geq \sum_{j \in T \setminus S} \mathbf{d}_j^2$ since $|S| = t = |T|$. Thus, we get $\|\mathbf{r} - \mathbf{k}\|_2^2 \geq \|\mathbf{r} - \mathbf{p}\|_2^2$

and since we have $\|\mathbf{r}-\mathbf{k}\|_2^2 \leq \|\mathbf{r}-\mathbf{q}\|_2^2$ from the construction of $\mathbf{k}$ as we saw earlier, this finishes the proof.

Acknowledgements The authors thank the anonymous reviewers for comments that helped improve the presentation of the paper. B.M. thanks the Research-I Foundation for financial support. The work of S.D. was partially supported by the DST-SERB grant ECR/2017/000374. P.K. thanks Microsoft Research India and Tower Research for research grants.

Declarations

Conflicts of interest The authors declare that they have no conflict of interest.

References

Agarwal, A., Negahban, S. N., & Wainwright, M. J. (2012). Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics, 40*(5), 2452–2482.

Bafna, M., Murtagh, J., & Vyas, N. (2018). Thwarting adversarial examples: an L_0-robust sparse Fourier transform. In *Proceedings of the 32nd annual conference on neural information processing systems (NIPS)*.

Baraniuk, R. G., Cevher, V., Duarte, M. F., & Hegde, C. (2010). Model-based compressive sensing. *IEEE Transactions on Information Theory, 56*(4), 1982–2001. https://doi.org/10.1109/TIT.2010.2040894

Barchiesi, D., & Plumbley, M. D. (2013) Learning incoherent subspaces for classification via supervised iterative projections and rotations. In *IEEE international workshop on machine learning for signal processing (MLSP)*. IEEE, pp 1–6.

Barchiesi, D., & Plumbley, M. D. (2015). Learning incoherent subspaces: classification via incoherent dictionary learning. *Journal of Signal Processing Systems, 79*(2), 189–199.

Bhatia, K., Jain, P., & Kar, P. (2015). Robust regression via hard thresholding. In *Proceedings of the 29th annual conference on neural information processing systems (NIPS)*.

Bouwmans, T., Javed, S., Zhang, H., Lin, Z., & Otazo, R. (2018). On the applications of robust PCA in image and video processing. *Proceedings of the IEEE, 106*(8), 1427–1457. https://doi.org/10.1109/JPROC.2018.2853589

Candes, E. J., & Wakin, M. B. (2008). An introduction to compressive sampling. *IEEE Signal Processing Magazine, 25*(2), 21–30.

Chen, X., & De, A. (2020). Reconstruction under outliers for Fourier-sparse functions. In *Proceedings of the ACM-SIAM symposium on discrete algorithms (SODA)*.

Chen, Y. (2015). Incoherence-optimal matrix completion. *IEEE Transactions on Information Theory, 61*(5), 2909–2923.

Chen Y, Bhojanapalli S, Sanghavi S, Ward R (2014) Coherent matrix completion. In: *Proceedings of the 31 st international conference on machine learning (ICML)*.

Cizek, P., & Sadikoglu, S. (2020). Robust nonparametric regression: A review. *WIREs Computational Statistics, 12*(3), e1492.

Coifman, R., Geshwind, F., & Meyer, Y. (2001). Noiselets. *Applied and Computational Harmonic Analysis, 10*(1), 27–44.

Dabov, K., Foi, A., Katkovnik, V., & Egiazarian, K. (2007). Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Transactions on Image Processing, 16*(8), 2080–2095. https://doi.org/10.1109/TIP.2007.901238

Diakonikolas, I., Kamath, G., Kane, D., Li, J., Steinhardt, J., Stewart, A. (2019). Sever: A robust meta-algorithm for stochastic optimization. In *36th international conference on machine learning (ICML)*.

Du SS, Wang Y, Balakrishnan S, Ravikumar P, Singh A (2018) Robust nonparametric regression under Huber's ϵ-contamination model. arXiv:1805.10406 [math.ST]

Fan, J., Hu, T. C., & Truong, Y. K. (1994). Robust non-parametric function estimation. *Scandinavian Journal of Statistics, 21*(4), 433–446.

Fan, L., Zhang, F., Fan, H., & Zhang, C. (2019). Brief review of image denoising techniques. *Visual Computing for Industry, Biomedicine, and Art, 2*(1), 7.

Fasshauer, G. E. (2011). Positive definite kernels: past, present and future. *Dolomites Research Notes on Approximation, 4*, 21–63.

Foucart, S., & Rauhut, H. (2013). *A mathematical introduction to compressive sensing*. Birkhäuser: Applied and Numerical Harmonic Analysis.

Getreuer, P. (2012). Rudin–Osher–Fatemi total variation denoising using split Bregman. *Image Processing on Line, 2*, 74–95.

Gu, S., Zhang, L., Zuo, W., & Feng, X. (2014) Weighted nuclear norm minimization with application to image denoising. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)* (pp. 2862–2869).

Guruswami, V., & Zuckerman, D. (2016) Robust Fourier and polynomial curve fitting. In *Proceedings of the 57th IEEE annual symposium on foundations of computer science (FOCS)*.

Hegde, C., & Baraniuk, R. G. (2012). Signal recovery on incoherent manifolds. *IEEE Transactions on Information Theory, 58*(12), 7204–7214. https://doi.org/10.1109/TIT.2012.2210860

Krahmer, F., & Ward, R. (2014). Stable and robust sampling strategies for compressive imaging. *IEEE Transactions on Image Processing, 23*(2), 612–622.

McCoy, M. B., & Tropp, J. A. (2014). Sharp recovery bounds for convex demixing, with applications. *Foundations of Computational Mathematics, 14*(3), 503–567.

Micchelli, C. A., Xu, Y., & Zhang, H. (2006). Universal kernels. *Journal of Machine Learning Research, 7*, 2651–2667.

Minh, H. Q., Niyogi, P., & Yao, Y. (2006). Merce's theorem, feature maps, and smoothing. In *Proceedings of the international conference on computational learning theory (COLT)*.

Mukhoty, B., Gopakumar, G., Jain, P., & Kar, P. (2019). Globally-convergent iteratively reweighted least squares for robust regression problems. In *Proceedings of the 22nd international conference on artificial intelligence and statistics (AISTATS)*.

Prasad, A., Suggala, A. S., Balakrishnan, S., & Ravikumar, P. (2018). *Robust estimation via robust gradient estimation*. arXiv:1802.06485 [stat.ML].

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Rosasco, L., Belkin, M., & Vito, E. D. (2010). On learning with integral operators. *Journal of Machine Learning Research, 11*, 905–934.

Schnass, K., & Vandergheynst, P. (2010). *Classification via incoherent subspaces*. arXiv:1005.1471 [cs.CV].

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing, 13*(4), 600–612.

Zhang, H., Zhou, Y., & Liang, Y. (2015). Analysis of robust PCA via local incoherence. In *Proceedings of the 29th annual conference on neural information processing systems (NIPS)*.

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing, 26*(7), 3142–3155.

Zhou, Y., Zhang, H., & Liang, Y. (2016). On compressive orthonormal sensing. In *54th annual allerton conference on communication, control, and computing (Allerton)*.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.