RENEGAR'S CONDITION NUMBER, SHARPNESS AND COMPRESSED SENSING PERFORMANCE

VINCENT ROULET, NICOLAS BOUMAL, AND ALEXANDRE D'ASPREMONT

ABSTRACT. We show that several quantities controlling compressed sensing performance also directly control algorithmic complexity. We describe linearly convergent restart schemes solving a broad range of compressed sensing problems using first-order methods. The key term controlling convergence measures the sharpness of the optimum and can be interpreted as a condition number, computed as the ratio between the true signal sparsity and the maximum signal size that can be recovered by the observation matrix. In a similar vein, Renegar's condition number is a data-driven computational complexity measure for convex programs, generalizing classical condition numbers for linear systems. We provide evidence that for a broad class of compressed sensing problems, the worst case value of this algorithmic complexity measure taken over all signals matches the restricted eigenvalue of the observation matrix, which controls compressed sensing performance. This condition number also measures the robustness of the recovered solution with respect to a misspecification of the observation matrix A, a point rarely addressed by classical recovery results. Overall, this means that, in compressed sensing problems, a single parameter directly controls computational complexity and recovery performance.

1. INTRODUCTION

Several recent results have highlighted a clear tradeoff between computational complexity on one side, and statistical performance on the other (i.e., the number of samples required to recover the signal). We focus on sparse recovery problems written

$$\begin{array}{ll} \text{minimize} & \|x\| \\ \text{subject to} & Ax = b \end{array} \tag{1}$$

in the variable $x \in \mathbb{R}^p$, where $A \in \mathbb{R}^{n \times p}$ is a sensing matrix and $b \in \mathbb{R}^n$ is the vector of observations. Here, $\|\cdot\|$ is a sparsity inducing norm (e.g., ℓ_1) whose properties will be specified below. Donoho and Tanner [2005] and Candès and Tao [2006] have shown that, for certain matrices A, when the observations y are generated by a sparse signal, i.e., when $b = Ax^*$ and $\mathbf{Card}(x^*) = k$ so the signal is sparse, $O(k \log p)$ observations suffice for stable recovery of x^* by solving problem (1) with the ℓ_1 norm, in which case (1) is a linear program. These results have been generalized to many other recovery problems with various assumptions on the signal structure (e.g., where x is a block-sparse vector, a low-rank matrix, etc.) and a library of corresponding convex relaxations has been developed to recover these more complex structures.

Many algorithms have also been developed to solve compress sensing problems at scale. Besides specialized methods such as LARS [Efron et al., 2004], the classical FISTA [Beck and Teboulle, 2009] and NESTA [Becker et al., 2011a] solvers use accelerated gradient methods to solve LASSO problems, with efficient and flexible implementations covering a much wider range of compressed sensing instances developed in e.g. [Becker et al., 2011b]. Several authors have also studied restart schemes and ODE interpretations [O'Donoghue and Candes, 2015; Su et al., 2014; Giselsson and Boyd, 2014] to speed up convergence in this context. More recently, linear convergence results have been obtained for recovery problems, with [Agarwal

Date: February 16, 2017.
2010 *Mathematics Subject Classification.* 90C25, 94A12.
Key words and phrases. Renegar's condition number, distance to infeasibility, restart, sparse recovery, linear convergence.

et al., 2011; Yen et al., 2014; Zhou et al., 2015] showing linear convergence of some first-order methods using variants of the strong convexity assumption.

In sparse recovery problems, statistical performance is usually measured in terms of the number of samples required to guarantee stable recovery, while computational performance is usually measured in terms of classical bounds on the computational cost of the corresponding convex optimization problems or M-estimators. Early on, it was noticed, for example in [Donoho and Tsaig, 2008], that recovery problems which are easier to solve from a statistical point of view (i.e., where more samples are available), are also easier to solve numerically. The results in [Donoho and Tsaig, 2008] focused on homotopy methods and were essentially empirical. More recently, the authors of [Chandrasekaran and Jordan, 2013; Amelunxen et al., 2014] studied computational and statistical tradeoffs for increasingly tight convex relaxations of shrinkage estimators. They show that recovery performance is directly linked to the Gaussian squared-complexity of the tangent cone with respect to the constraint set and study the complexity of several convex relaxations. In [Chandrasekaran and Jordan, 2013; Amelunxen et al., 2014] the structure of the convex relaxation is varying and affecting both complexity and recovery performance, while in [Donoho and Tsaig, 2008] and in what follows, the structure of the relaxation is fixed, but the data (i.e., the observation matrix A) varies.

Here, following results in [Roulet and d'Aspremont, 2017], we first describe linearly convergent restart schemes solving a broad range of compressed sensing problems using first-order methods. The key term controlling convergence measures the sharpness of the optimum can be interpreted as a condition number, computed as the ratio between the true signal sparsity and the maximum signal size that can be recovered by the observation matrix.

In a similar spirit, we show that the *cone restricted eigenvalues* introduced in [Bickel et al., 2009] correspond to the worst-case value of Renegar's condition number for problem (1) taken over a class of signals x_0. This means that a single quantity drives both the complexity of solving problem (1) and its recovery performance, i.e., the number of samples required for exact recovery, and the solution's robustness when the observations y are noisy. This same condition number also controls the impact of misspecification in A on the optimal solution to (1). From a compressed sensing perspective, this confirms that obtaining more samples also makes the reconstructed solution more robust to experimental uncertainty in A.

2. Sharpness & Lower Bounds

In what follows, we show that Nullspace Property conditions (see e.g. [Cohen et al., 2009]) produce sharpness results on the optimum. In particular, in the ℓ_1 setting, we show that if $\hat{x}$ solves the sparse recovery problem (1), then

$$\frac{2-C}{C}\|x-\hat{x}\|_1 \leq \|x\|_1 - \|\hat{x}\|_1$$

for any x such that $Ax = b$. This *sharpness* bound on the optimum will allow us to produce restart schemes accelerating the performance of classical optimization algorithms. Furthermore, the constant C controlling acceleration depends explicitly on both the sparsity of the solution and on the recovery threshold of the observation matrix A. This directly links quantities controlling sparse recovery performance with measures of computational complexity. For simplicity, we start by describing the ℓ_1 setting, we then generalize our results to other sparsity inducing norms.

2.1. The ℓ_1 setting.

We first briefly recall key results on sparse recovery using the ℓ_1 norm, then use these results to produce sharpness bounds on the recovery problem.

Definition 2.1. (Nullspace Property [Cohen et al., 2009]) *The matrix A satisfies the Nullspace Property (NSP) of order k with constant $C > 1$ iff*

$$\|x\|_1 \leq C\|x_{T^c}\|_1, \tag{NSP}$$

for any $x \in \mathcal{N}(A)$ and $T \subset [1,p]$ with $\mathbf{Card}(T) \leq k$.

Given a matrix $A \in \mathbb{R}^{n\times p}$ and observations $b = Ax^*$ on a signal $x^* \in \mathbb{R}^p$, recovery is performed by solving the following ℓ_1 minimization program

$$\begin{array}{ll}\mbox{minimize} & \|x\|_1\\ \mbox{subject to} & Ax = b\end{array} \tag{2}$$

in the variable $x \in \mathbb{R}^n$. We call $\hat{x}$ the optimal solution of this problem. The nullspace property means this convex program recovers all signals up to some sparsity threshold on these sparse recovery problems.

Proposition 2.2. *Given a coding matrix $A \in \mathbb{R}^{n\times p}$ satisfying the Nullspace Property* (NSP) *at order k with constant $1 < C < 2$, then sparse recovery $\hat{x} = x^*$ is guaranteed if $\mathbf{Card}(x^*) \leq k$, and*

$$\|\hat{x} - x^*\|_1 \leq \frac{2C}{2-C} \inf_{\{\mathbf{Card}\, u \leq k\}} \|u - x^*\|_1$$

where $\hat{x}$ solves the ℓ_1-minimization LP and x^* is the true signal.

Proof. (see e.g. Cohen et al. [2009] Th. 4.3). If A satisfies the NSP at order $2k$ with constant C, then

$$\|z\|_1 \leq C\|z_{T^c}\|_1$$

for any $z \in \mathcal{N}(A)$ and $T \subset [1, p]$ with $\mathbf{Card}(T) \leq 2k$, means

$$\|z\|_1 \geq \frac{C}{C-1}\|z_T\|_1.$$

Now let $T = \mathrm{supp}(x^*)$ and let $x \neq x^*$ such that $Ax = b$, so $z = x - x^*$ satisfies $Az = 0$, then

$$\begin{array}{rcl}\|x\|_1 & = & \|x^*_T + z_T\|_1 + \|z_{T^c}\|_1\\ & \geq & \|x^*_T\|_1 - \|z_T\|_1 + \|z_{T^c}\|_1\\ & = & \|x^*\|_1 + \|z\|_1 - 2\|z_T\|_1\end{array}$$

and $C < 2$ means

$$\|z\|_1 - 2\|z_T\|_1 > \|z\|_1 - \frac{C}{C-1}\|z_T\|_1 \geq 0$$

hence $\|x\|_1 > \|x^*\|_1$, so $\hat{x} = x^*$. The error bound follows from similar arguments. ■

We can use these last results to bound suboptimality using the distance to the optimal set. We get the following proposition bounding the sharpness of the optimum of problem (1).

Proposition 2.3. *Given a coding matrix $A \in \mathbb{R}^{n\times p}$ satisfying the Nullspace Property* (NSP) *at order k with constant $1 < C < 2$. Let $\hat{x}$ be the solution of program* (2) *for $b = Ax^*$ with $\mathbf{Card}(x^*) \leq k$. Let $x \in \mathbb{R}^p$ satisfy $Ax = b$, we have*

$$\|x\|_1 - \|\hat{x}\|_1 \geq \frac{2-C}{C}\|x - \hat{x}\|_1 \tag{Sharp}$$

with $\hat{x} = x^*$.

Proof. The hypotheses of Proposition (2.2) are satisfied so $\hat{x} = x^*$, $z = x - \hat{x} \in \mathcal{N}(A)$ and following the proof of that proposition we get

$$\|x\|_1 - \|\hat{x}\|_1 \geq \|z\|_1 - 2\|z_T\|_1$$

and

$$\|z\|_1 \geq \frac{C}{C-1}\|z_T\|_1$$

yields the desired result. ■

The nullspace property (NSP) ensures that there are no (approximately) sparse vectors in the nullspace of the observation matrix A. We can give a more concrete geometric meaning to the constant C in (NSP) by connecting it with the diameter of a section of the ℓ_1 ball by the nullspace of the matrix A (see e.g. Kashin and Temlyakov [2007] for more details).

Lemma 2.4. *Suppose $A \in \mathbb{R}^{n\times p}$ satisfies*

$$\frac{1}{2}\,\mathbf{diam}(B_1^p \cap \mathcal{N}(A)) = \sup_{\substack{Ax=0\\ \|x\|_1\leq 1}} \|x\|_2 \leq k_D^{-1/2}$$

then A satisfies (NSP) at order k_T with constant

$$C = \frac{1}{1-\sqrt{k_T/k_D}} \tag{3}$$

provided $k_T < k_D$.

Proof. For any $x \in \mathcal{N}(A)$ and support set T with $\mathbf{Card}(T) \leq k_T$,

$$\|x_T\|_1 \leq \sqrt{k_T}\,\|x\|_2 \leq \sqrt{k_T/k_D}\,\|x\|_1,$$

which means

$$\|x_{T^c}\|_1 \geq (1-\sqrt{k_T/k_D})\|x\|_1$$

hence the desired result. ■

Precise estimates of the diameter of random sections of norm balls can be computed using classical results in geometric functional analysis. The low M^* estimates in [Pajor and Tomczak-Jaegermann, 1986] for example show that when E is random subspace of codimension k (e.g. the nullspace of a random matrix A), then

$$\mathbf{diam}(B_1^p \cap E) \leq c\sqrt{\frac{\log(n/k)}{k}}$$

with high probability, where $c > 0$ is an absolute constant. In this case, Proposition 2.3 means that if $\hat{x}$ solves the recovery problem in (1), then (Sharp) reads

$$\|x\|_1 - \|\hat{x}\|_1 \geq \left(1 - c\sqrt{\frac{\mathbf{Card}(T)\log(n/k)}{k}}\right)\|x-\hat{x}\|_1 \tag{4}$$

where T is the support of the true signal x^*. This means that the sharpness of the optimum of (1) is essentially controlled by the ratio of the true signal size $\mathbf{Card}(T)$ with the maximum signal size k that can be recovered w.h.p. by the observation matrix A.

2.2. **General sparsity inducing norms.** We now generalize the results above using the notion of *sparsity structures* introduced by Juditsky et al. [2014], which allows a common treatment of popular norms such as the ℓ_1 norm, group-ℓ_1 norms and the nuclear norm. Sparsity structures define sparsity (or simplicity) of signals through projectors.

We begin by briefly recalling the setting in [Juditsky et al., 2014]. Consider $\mathcal{X}$ and $\mathcal{E}$, two Euclidean spaces, and a map $B\colon \mathcal{X} \to \mathcal{E}$. In most cases, notably including the ℓ_1 and nuclear norm, $\mathcal{X} = \mathcal{E}$ and we may think of B as the identity map, but it is useful to consider more general B to model group norms as well. In this setting, the problem under consideration is that of finding a sparse representation Bx of a signal, given noisy observations $y = Ax$.

Definition 2.5. (Sparsity structure [Juditsky et al., 2014]) *A sparsity structure on $\mathcal{E}$ is defined as a norm $\|\cdot\|$ on $\mathcal{E}$, together with a family $\mathcal{P}$ of linear maps of $\mathcal{E}$ into itself, satisfying three assumptions:*

(1) Every $P \in \mathcal{P}$ is a projector, $P^2 = P$,
(2) Every $P \in \mathcal{P}$ is assigned a weight $\nu(P) \geq 0$ and a linear map $\bar{P}$ on $\mathcal{E}$ such that $P\bar{P} = 0$,

(3) For any $P \in \mathcal{P}$ and $f, g \in \mathcal{E}$, one has

$$\|P^* f + \bar{P}^* g\|_* \leq \max(\|f\|_*, \|g\|_*),$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$ and P^* is the conjugate mapping of the linear map P.

The last condition in Definition 2.5 is arguably the least intuitive and Lemma 4.1 connects it, in some cases, with the more intuitive notion of decomposable norm. For $k \geq 0$, let

$$\mathcal{P}_k = \{P \in \mathcal{P} \,:\, \nu(P) \leq k\}.$$

The notion of sparsity is defined as follows: a vector w is said to be k-sparse if there exists $P \in \mathcal{P}_k$ such that $Pw = w$. A signal x is said to be k-sparse if its representation Bx is k-sparse.

ℓ_1 **norm.** In the the ℓ_1 norm case, the sparsity structure is defined over $\mathcal{E} = \mathcal{X} = \mathbb{R}^p$, the map B reduces to the identity, and $\mathcal{P}$ is the set of projectors on coordinate subspaces of $\mathbb{R}^p$, that is, $\mathcal{P}$ contains all projectors which zero out all coordinates of a vector except for a subset of them, which are left unaffected. The companion maps are the complementary projectors: $\bar{P} = I - P$. Naturally, the complexity level corresponds to the number of coordinates preserved by P, i.e., $\nu(P) = \mathbf{Rank}(P)$. These definitions recover the usual notion of sparsity.

Nuclear norm. The nuclear norm is defined for matrices $X \in \mathbb{R}^{p\times q}$ with singular values $\sigma_i(X)$ as $\|X\| = \sum_{k=1}^{\min(p,q)} \sigma_k(X)$. It can be cast as a sparsity system by setting $\mathcal{X} = \mathcal{E} = \mathbb{R}^{p\times q}$, $B = I$. Its associated family of linear maps is

$$P\colon X \mapsto P_{\text{left}} X P_{\text{right}},$$

and

$$\bar{P}\colon X \mapsto (I - P_{\text{left}}) X (I - P_{\text{right}}),$$

where $P_{\text{left}} \in \mathbb{R}^{p\times q}$ and $P_{\text{right}} \in \mathbb{R}^{p\times q}$ are orthogonal projectors. Their weights are defined as $\nu(P) = \max\left(\mathbf{Rank}(P_{\text{left}}), \mathbf{Rank}(P_{\text{right}})\right)$ defining therefore k-sparse matrices as matrices of rank at most k.

Definition 2.5 allows us to revisit all the results of Section 2.1. This is essentially a direct generalization, with the caveat that since we do not assume $P + \bar{P} = \mathbf{I}$ (which roughly corresponds to decomposable norms), the recovery conditions in [Juditsky et al., 2014] that we recall below differs slightly from (NSP). In the setting discussed above, let $\hat{x} \in \mathcal{X}$ solve the following optimization problem

$$\begin{array}{ll} \text{minimize} & \|Bx\| \\ \text{subject to} & Ax = b \end{array} \tag{5}$$

in the variable $x \in \mathcal{X}$. [Juditsky et al., 2014] then show a slightly more general version of the following stable recovery result.

Proposition 2.6. *Suppose $\hat{x} \in \mathcal{X}$ solves the recovery problem (5) with observations $b = Ax^*$ where x^* is the true signal, up to a precision ϵ, and that the true signal x is nearly sparse, i.e. there exists $P \in \mathcal{P}_k$ such that*

$$\|(I - P)Bx\| \leq \delta_x$$

Assume also that the following condition holds

$$\|PBz\| - \|\bar{P}Bz\| + \|Bz\| \leq \|Bz\|_{\mathcal{P}}, \tag{6}$$

for any $z \in \mathcal{X},\ P \in \mathcal{P}_k$, then

$$\|B\hat{x} - Bx^*\| \leq \frac{\epsilon + 2\delta_x}{1 - \gamma} \tag{7}$$

where $\|\cdot\|$ is the norm defined in Def. 2.5.

Proof. See [Juditsky et al., 2014, Prop. 3.1]. ■

The condition in (6) is slightly stronger than the classical nullspace property but ensures stable recovery when problem (5) is only solved approximately. This result also allows us to produce a direct generalization of Proposition 2.3, as follows.

Proposition 2.7. *Given a sparsity system $\mathcal{P}$ satisfying assumption (6) in Proposition 2.6, let $\hat{x}$ be the solution of program (5) for $b = Ax^*$ with $\mathbf{Card}(x^*) \leq k$. If $x \in \mathcal{X}$ satisfies $Ax = b$, then*

$$\|x\|_1 - \|\hat{x}\|_1 \geq (1-\gamma)\|x - \hat{x}\|_1 \tag{Sharp-Gen}$$

with $\hat{x} = x^*$.

Proof. If $\mathcal{P}$ satisfies assumption (6) in Proposition 2.6, then [Juditsky et al., 2014, Prop. 3.1] shows that $\hat{x} = x^*$. For $z \in \mathcal{N}(A)$ we have

$$\|\bar{P}Bz\| - \|PBz\| \geq (1-\gamma)\|Bz\|$$

which combined with [Juditsky et al., 2014, Lem. 3.1] and the fact that $Ax = b$ yields the desired result. ■

This last bound is a direct generalization of the sharpness result in Proposition 2.3, with (Sharp-Gen) extending the inequality in (Sharp). While the coefficient $(1-\gamma)$ in (Sharp-Gen) is perhaps less intuitive than the condition number in (Sharp), we observe that here too this same coefficient controls both recovery stability in the error bound (7) of Prop. 2.6, and sharpness (hence computational complexity as we will see below) in bound (Sharp-Gen).

2.3. **Restarting First-Order Methods.** In this section we seek to solve the recovery problem (1) assuming that the sharpness bounds (Sharp) hold. The NESTA method detailed in [Becker et al., 2011a] uses the smoothing argument in [Nesterov, 2005] to solve (1). In practice, this means using the optimal algorithm in [Nesterov, 1983] to minimize

$$f_\mu(x) \triangleq \sup_{\|u\|_\infty \leq 1} u^T x - \mu\|u\|_2^2/2$$

for some $\mu > 0$, which approximates the ℓ_1 norm uniformly up to $\mu p/2$. This is the classical Huber function, which has a Lipschitz continuous gradient with constant equal to $1/\mu$. Starting at a point x_0, t iterations of the optimal algorithm in [Nesterov, 1983] will then yield a point x_t satisfying

$$\|x_t\|_1 - \|\hat{x}\|_1 \leq \frac{2\|x_0 - x^*\|_2^2}{\mu t^2} + \frac{\mu p^2}{2}$$

and the optimal bound is reached for $\mu = \sqrt{2}\|x_0 - x^*\|_2/(t\sqrt{p})$ and reads

$$\|x_t\|_1 - \|\hat{x}\|_1 \leq \frac{3\sqrt{p}\|x_0 - x^*\|_2}{t} \tag{8}$$

As in [Roulet and d'Aspremont, 2017], we write $\mathcal{A}(x_0, t) \triangleq x_t$ the output of this algorithm and describe a restart scheme exploiting the sharpness result in (Sharp) to improve the computational complexity of solving problem (1). In fact, when Proposition 2.3 holds, combining (9) and (Sharp) yields

$$\begin{aligned} \|x_t\|_1 - \|\hat{x}\|_1 &\leq \frac{3\sqrt{p}\|x_0 - x^*\|_2}{t} \\ &\leq \frac{3\sqrt{p}C}{t(2-C)}(\|x_0\|_1 - \|\hat{x}\|_1) \end{aligned}$$

hence if we pick t such that

$$\frac{3\sqrt{p}C}{t(2-C)} < 1$$

the simple, constant restart scheme running summarized in Algorithm 1. After τ outer iterations, hence a total of $N = t\tau$ inner iterations, the algorithm will produce a point y_τ satisfying

$$\|y\|_1 - \|\hat{x}\|_1 \leq \left(\frac{3\sqrt{p}C}{t(2-C)}\right)^{N/t}$$

minimizing this last bound in t yields the following proposition.

Proposition 2.8. *Given a coding matrix $A \in \mathbb{R}^{n\times p}$ satisfying the Nullspace Property (NSP) at order $2k$ with constant $C < 2$. Let $\hat{x}$ be the solution of program (2) for $b = Ax^*$ with $\mathbf{Card}(x^*) \leq k$. After running a total of N inner iterations in Algorithm 1, we get a point $y \in \mathbb{R}^p$ such that*

$$\|y\|_1 - \|\hat{x}\|_1 \leq \exp\left(-\frac{N(2-C)}{3e\sqrt{p}C}\right) \tag{9}$$

for $t = 3e\sqrt{p}C/(2-C)$, hence N/t restarts.

Recall from (3) that for random observations, we have

$$\frac{2-C}{C} = 1 - 2\sqrt{k_T/k_D}$$

where k_T is roughly the largest signal size that can be recovered by the observations A and k_T is the size of the true signal. This means that the complexity of the optimization problem (1) decreases with the complexity of the statistical recovery problem, with both quantities being controlled by the oversampling ratio k_D/k_T.

Algorithm 1 Restart Scheme

Input: Initial point $y_0 \in \mathbb{R}^p$

For $i = 1\ldots,\tau$ compute

$$y_i = \mathcal{A}(y_{i-1}, t) \tag{Restart}$$

Output: A point y_τ approximately solving (1).

---

3. Renegar's condition number & restricted eigenvalues

Renegar's condition number is a data-driven computational complexity measure for convex programs, generalizing classical condition numbers for linear systems. In what follows, we show that for a broad class of compressed sensing problems, the worst case value of this algorithmic complexity measure taken over all signals matches the restricted eigenvalue of the observation matrix, which controls compressed sensing performance.

3.1. **Computational complexity.** We begin by addressing computational complexity aspects of problem (1). Computational complexity for convex optimization problems is often described in terms of polynomial functions of the problem size. This produces a clear link between problem structure and computational complexity but fails to account for the nature of the data. If we use linear systems as a basic example, unstructured linear systems of dimension n can be solved with complexity $O(n^3)$ regardless of the matrix values, but iterative solvers will converge much faster on systems that are better conditioned. The seminal work of [Renegar, 1995a, 2001] extends this notion of conditioning to optimization problems, producing data-driven bounds on the complexity of solving conic programs, and showing that the number of outer iterations of interior point algorithms increases as the distance to ill-posedness decreases.

In what follows, we study the complexity of the oracle certifying optimality of a candidate solution x^* to (1) as a proxy for the problem of computing an optimal solution to this problem. As we will see below, certifying optimality means solving a pair of alternative conic linear systems of the form

$$Ax = 0,\ x \in C \tag{10}$$

and

$$-A^T b \in C^* \tag{11}$$

for a given cone $C \subset \mathbb{R}^p$. Several references have connected Renegar's condition number $\mathcal{C}(A)$ (which will be defined more precisely below) and the complexity of solving conic linear systems using various algorithms [Renegar, 1995a; Freund and Vera, 1999b; Epelman and Freund, 2000; Renegar, 2001; Vera et al., 2007; Belloni et al., 2009]. In particular, Vera et al. [Vera et al., 2007] link $\mathcal{C}(A)$ to the complexity of solving the primal dual pair (10)–(11) using a barrier method. They show that the number of outer barrier method iterations grows as

$$O\left(\sqrt{\nu_C}\log\left(\nu_C\,\mathcal{C}(A)\right)\right),$$

where ν_C is the barrier parameter, while the conditioning (hence the complexity) of the linear systems arising at each interior point iteration is controlled by $\mathcal{C}(A)^2$. This link was also tested empirically on linear programs using the NETLIB library of problems in [Ordóñez and Freund, 2003], where computing times and number of iterations were regressed against estimates of the condition number computed using the approximations for $\mathcal{C}(A)$ detailed in [Freund and Vera, 2003].

Studying the complexity of computing an optimality certificate in (10) gives insights on the performance of oracle based optimization techniques such as the ellipsoid method. Of course, these abstract methods are very different from those used to solve problem (1) in pratice. However, we will observe in the numerical experiments of Section 4 that the condition number is strongly correlated with the empirical performance of efficient recovery algorithms such as LARS [Efron et al., 2004] and Homotopy [Donoho and Tsaig, 2008; Asif and Romberg, 2014].

We now briefly recall optimality conditions for problem (1) and two equivalent constructions for the condition number of a conic linear system. Define the *tangent cone* at point x with respect to the norm $\|\cdot\|$, that is, the set of descent directions for the norm $\|\cdot\|$ at x, as

$$\mathcal{T}(x) = \operatorname{cone}\{z : \|x+z\| \leq \|x\|\}. \tag{12}$$

The simple lemma below characterizes unique optimal solutions to problem (1) in terms of $\mathcal{T}(x)$.

Lemma 3.1. *The point x^* is the unique minimizer of* (1) *if and only if $Null(A) \cap \mathcal{T}(x^*) = \{0\}$.*

Proof. This follows from standard KKT conditions (see for example [Chandrasekaran et al., 2012, Prop 2.1]). ■

In other words, x^* is the unique optimizer if and only if the following problem is *infeasible*

$$\begin{array}{ll} \text{find} & z \\ \text{s.t.} & Az = 0 \\ & z \in \mathcal{T}(x^*),\ z \neq 0, \end{array} \tag{P}$$

in the variable $z \in \mathbb{R}^p$. To certify feasibility, it is sufficient to exhibit a solution. One way of certifying infeasibility of (P) is to solve the dual problem

$$\begin{array}{ll} \text{find} & u \\ \text{s.t.} & -A^T u \in \mathcal{T}(x^*)^\circ,\ u \neq 0, \end{array} \tag{D}$$

in the variable $u \in \mathbb{R}^n$, where $\mathcal{T}(x^*)^\circ$ is the polar cone of $\mathcal{T}(x^*)$. Renegar's condition number [Renegar, 1995a,b; Peña, 2000] provides a data-driven measure of the complexity of this task. It is rooted in the sensible idea that certifying infeasibility is easiest if the problem is far from being feasible. Formally, the distance to feasibility $\rho_{x^*}(A)$ is defined as follows. Let $\mathcal{M}_{x^*}^P = \{A \in \mathbb{R}^{n\times p} : \text{(P) is infeasible}\}$. Then, using the spectral norm as matrix norm,

$$\rho_{x^*}^P(A) \triangleq \inf_{\Delta A} \{\|\Delta A\|_2 : A + \Delta A \notin \mathcal{M}_{x^*}^P\}. \tag{13}$$

Renegar's *condition number* for problem (P) with respect to x^* is then defined as the scale-invariant reciprocal of this distance

$$\mathcal{C}_{x^*}(A) \triangleq \frac{\|A\|_2}{\rho_{x^*}^P(A)}. \tag{14}$$

We can also define conically restricted minimal singular value of A as follows

$$\mu_{x^*}(A) = \inf_{z\in\mathcal{T}(x^*)} \frac{\|Az\|_2}{\|z\|_2}. \tag{15}$$

Interestingly, this last quantity turns out to be equal to the distance to infeasibility and we have the following result.

Lemma 3.2. *Distance to feasibility and cone restricted eigenvalues match, i.e.* $\rho_{x^*}^P(A) = \mu_{x^*}(A)$.

Proof. When (P) is feasible, both vanish. Otherwise, see [Freund and Vera, 1999a, Th. 2], or simplified versions in [Belloni and Freund, 2009, Lem. 3.2] and [Amelunxen and Lotz, 2014]. ■

Notice that, if $\mathcal{T}(x^*)$ were the whole space $\mathbb{R}^p$, and if A^TA were full-rank (never the case if $n < p$), then $\mu(A)$ would be the smallest singular value of A. As a result, $\mathcal{C}(A)$ would reduce to the classical condition number of A (and to ∞ when A^TA is rank-deficient). Renegar's condition number is necessarily smaller (better) than the latter, as it further incorporates the notion that A need only be well-conditioned along those directions that matter with respect to the norm $\|\cdot\|$ at x^*. Later, we will remove the dependence on x^* by considering a worst-case condition number over classes of "simple" signals.

When the primal problem (P) is feasible, so that $\mu_{x^*}(A) = 0$, the condition number as defined here is infinite. While this correctly captures the fact that, in that regime, statistical recovery does not hold, it does not properly capture the fact that, when (P) is "comfortably" feasible, certifying so is easy, and algorithms terminate quickly (although they return a useless estimator). From both a statistical and a computational point of view, the truly delicate cases correspond to problem instances for which both (P) and (D) are only barely feasible or infeasible. This is illustrated in simple numerical example in [Boyd and Vandenberghe, 2004, §11.4.3] and in our numerical experiments, corresponding to the peaks in the CPU time plots of the right column in Figure 3: problems where sparse recovery barely holds/fails are relatively harder. For simplicity, we only focused here on distance to feasibility for problem (P). However, it is possible to symmetrize the condition numbers used here as described in [Amelunxen and Lotz, 2014, §1.3], where a symmetric version of the condition number is defined as

$$\mathcal{R}(A) = \min\left\{\frac{\|A\|}{\rho_{x^*}^P(A)}, \frac{\|A\|}{\rho_{x^*}^D(A)}\right\}$$

This quantity peaks for programs that are nearly feasible/infeasible. Naturally, the condition number also controls the sensitivity of the solution to changes in the matrix A, with [Renegar, 1994, 1995b] for example directly bounding changes in the solution to (10) in terms of $\mathcal{C}(A)$ and changes ΔA in the system matrix. This means that $\mathcal{C}(A)$ also measures the robustness of the solution to the recovery with respect to misspecification *of the observation matrix* A, a point rarely addressed by classical recovery results.

3.2. **Statistical performance.** We now focus on the link between condition number and the statistical performance of the solution of problem (1). To this end, assume now that the observations y are affected by noise and that we solve a robust version of problem (1), written

$$\begin{array}{ll} \text{minimize} & \|x\| \\ \text{subject to} & \|Ax - b\|_2 \le \delta\|A\|_2, \end{array} \tag{16}$$

in the variable $x \in \mathbb{R}^p$, with the same design matrix $A \in \mathbb{R}^{n\times p}$, observations $y \in \mathbb{R}^n$ and noise level $\delta > 0$.

3.2.1. *Recovery bounds using $\mathcal{C}(A)$.* The following classical result bounds the reconstruction error in terms of $\mathcal{C}(A)$.

Lemma 3.3. *Suppose we observe $y = Ax_0 + w$ where $\|w\|_2 \leq \delta\|A\|_2$ and let x^* be an optimal solution of problem* (16)*. We get the following error bound:*

$$\|x^* - x_0\|_2 \leq 2\frac{\delta\|A\|_2}{\mu_{x_0}(A)} = 2\delta \cdot \mathcal{C}_{x_0}(A). \tag{17}$$

Proof. We recall the short proof of [Chandrasekaran et al., 2012, Prop. 2.2]. Both x^* and x_0 are feasible for (16) and x^* is optimal, so that $\|x^*\| \leq \|x_0\|$. Thus, the error vector $x^* - x_0$ is in the tangent cone $\mathcal{T}(x_0)$ (12). By the triangle inequality,

$$\|A(x^* - x_0)\|_2 \leq \|Ax^* - y\|_2 + \|Ax_0 - y\|_2 \leq 2\delta\|A\|_2.$$

Furthermore, by definition of μ_{x_0} (15),

$$\|A(x^* - x_0)\|_2 \geq \mu_{x_0}(A)\, \|x^* - x_0\|_2.$$

Combining the two concludes the proof. ■

Notice that, in the above lemma, the condition number is evaluated at x_0 (the true signal) rather than at x^* (the estimator). This will be convenient when considering worst cases over classes of target signals.

This means that Renegar's condition number defined in (14) also controls the statistical performance of estimators built on solving the approximate recovery problem (16). This at least partially explains the common empirical observation (see, e.g., [Donoho and Tsaig, 2008]) that problem instances where statistical estimation succeeds are computationally easy to solve. In fact, we will see in what follows that the worst-case value of the distance to infeasibility coincides with classical measures of recovery performance, such as restricted eigenvalues [Bickel et al., 2009] in the ℓ_1 case.

On paper, the computational complexities of (1) and (16) are very similar (in fact, infeasible start primal-dual algorithms designed for solving (1) actually solve problem (16) with δ small). In our experiments, we did observe sometimes significant differences in behavior between the noisy and noiseless case.

3.2.2. *Generalized restricted eigenvalues.* We now further specify the sparsity inducing norms in order to study Renegar's condition number on a class of signals that share the same sparsity properties. We use again the framework of sparsity structure introduced by Juditsky et al. [Juditsky et al., 2014], recalled in the previous section. Given a sparsity structure and $k \geq 0$, Lemma 3.2 shows that the worst-case distance to infeasibility on the class of k-sparse signals can be written as

$$\mu_k(A) \triangleq \inf_{\substack{P\in\mathcal{P},\, \nu(P)=k,\\ x\in\mathcal{X},\, PBx=Bx,}} \inf_{z\in\mathcal{T}(x)} \frac{\|Az\|_2}{\|z\|_2},$$

(the first infimum covers all signals x of sparsity k), where the tangent cone is defined here as

$$\mathcal{T}(x) = \operatorname{cone}\{z \in \mathcal{X} :\ \|Bx + Bz\| \leq \|Bx\|\}.$$

In the following lemma, we show that this worst-case distance to infeasibility μ_k is directly related to generalized restricted eigenvalues.

Lemma 3.4. *Given a sparsity structure $(\|\cdot\|, \mathcal{P})$, for $P \in \mathcal{P}$, let*

$$\mathcal{C}_P = \bigcup_{\{x\in\mathcal{X}\,:\,PBx=Bx\}} \mathcal{T}(x), \qquad \textit{and} \qquad \mathcal{D}_P = \{z \in \mathcal{X},\ \|\bar{P}Bz\| \leq \|PBz\|\}. \tag{18}$$

Then, $\mathcal{C}_P \subseteq \mathcal{D}_P$. Hence, for any $k \geq 0$,

$$\mu_k(A) \geq \sigma_k(A) = \inf_{\substack{P \in \mathcal{P}_k,\, z \in \mathcal{X},\\ \|\bar{P}Bz\| \leq \|PBz\|}} \frac{\|Az\|_2}{\|z\|_2}.$$

Proof. Let $P \in \mathcal{P}$ and $z \in \mathcal{T}(x)$ for $x \in \mathcal{X}$ such that $PBx = Bx$. Using [Juditsky et al., 2014, Lem. 3.1], we have

$$\|Bx\| + \|\bar{P}Bz\| - \|PBz\| \leq \|Bx + Bz\| \leq \|Bx\|.$$

So $\|\bar{P}Bz\| \leq \|PBz\|$ and $z \in \mathcal{D}_P$. ■

The inverse of the generalized restricted eigenvalue therefore also bounds the worst case computational complexity through the condition number. We remark that, in general, one does not have $\mu_k(A) = \sigma_k(A)$. A simple counterexample can be derived for the nuclear norm. Indeed, let $\mathcal{E} = \mathcal{X} = \mathbb{R}^{2\times 2}$, B be the identity and $(\|\cdot\|, \mathcal{P})$ be the nuclear norm and its associated family of linear maps. Let

$$Q = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} \qquad \text{and} \qquad U = \begin{pmatrix} 0 & u \\ u & 0 \end{pmatrix}$$

with $u \neq 0$. Setting $P\colon \mathcal{X} \to QXQ$, so that $P \in \mathcal{P}$, we have $\|PU\| = \|\bar{P}U\| = 0$, hence $U \in \mathcal{D}_P$. Now let

$$\mathcal{X}_P = \{X \in \mathcal{X} : PX = X\} = \left\{ \begin{pmatrix} x & 0 \\ 0 & 0 \end{pmatrix} : x \in \mathbb{R} \right\}.$$

For any $X \in \mathcal{X}_P$, $\|X+U\| = \sqrt{x^2 + 4u^2} > |x| = \|X\|$, hence $U \notin \cup_{X \in \mathcal{X}_P} \mathcal{T}(X)$, showing that $\mathcal{D}_P \nsubseteq \mathcal{C}_P$.

As shown below however, with the additional assumption that the norm is *strictly decomposable*, that is, $\bar{P} = I - P$ and that B is bijective (non-overlapping groups) the bound in Lemma 3.4 is tight and $\mu_k(A) = \sigma_k(A)$.

Lemma 3.5. *Given a sparsity structure $(\|\cdot\|, \mathcal{P})$, assume that for any $P \in \mathcal{P}$, $\bar{P} = I - P$. Then $\mathcal{C}_P = \mathcal{D}_P$ and, for any $k \geq 0$, we have*

$$\mu_k(A) = \sigma_k(A).$$

Proof. Let $P \in \mathcal{P}$ and $z \in \mathcal{X}$, $\|\bar{P}Bz\| \leq \|PBz\|$. Let $w = -PBz$. We have $Pw = w$ and

$$\|w + Bz\| = \|(I-P)Bz\| = \|\bar{P}Bz\| \leq \|PBz\| = \|w\|.$$

Thus, $z \in \mathcal{T}(x)$ for x such that $w = Bx$, and $PBx = Bx$ implies $z \in \mathcal{C}_P$. ■

In the ℓ_1 case, our definition for $\mu_k(A) = \sigma_k(A)$ matches the definition of restricted eigenvalue in [Bickel et al., 2009], with

$$\sigma_k(A) = \inf_{\substack{S \subset [1,p] \colon \mathbf{Card}(S) = k \\ z \in \mathbb{R}^p \colon \|z_{S^c}\|_1 \leq \|z_S\|_1}} \frac{\|Az\|_2}{\|z\|_2}.$$

This, we believe, makes for an interesting link between the statistical notion of restricted eigenvalue, and the computational notion of Renegar condition number.

4. Numerical Results

4.1. **Sharpness & restart.** We test the restart scheme in Algorithm 1 on ℓ_1-recovery problems with random design matrices. Throughout the experiments, we use the NESTA code described in [Becker et al., 2011a]. We generate a random design matrix $A \in \mathbb{R}^{n \times p}$ with i.i.d Gaussian coefficients. We then normalize A so that $AA^T = \mathbf{I}$ (to fit NESTA's format) and generate observations $b = Ax^*$ where $x^* \in \mathbb{R}^p$ is a k-sparse vector whose nonzero coefficients are all ones. In Figure 1 we compare the performance of running NESTA with and without restart, for various values of the number of inner iteration t and outer iterations τ. We observe that restart can improve performance but that this improvement can be neutralized if the number of outer iterations is set much too high. Here, we have set $p = 500$, $m = 300$ and $k = 50$. To minimize the number of moving parameters, we do not use continuation in [Becker et al., 2011a] hence directly implement the method in [Nesterov, 2005].

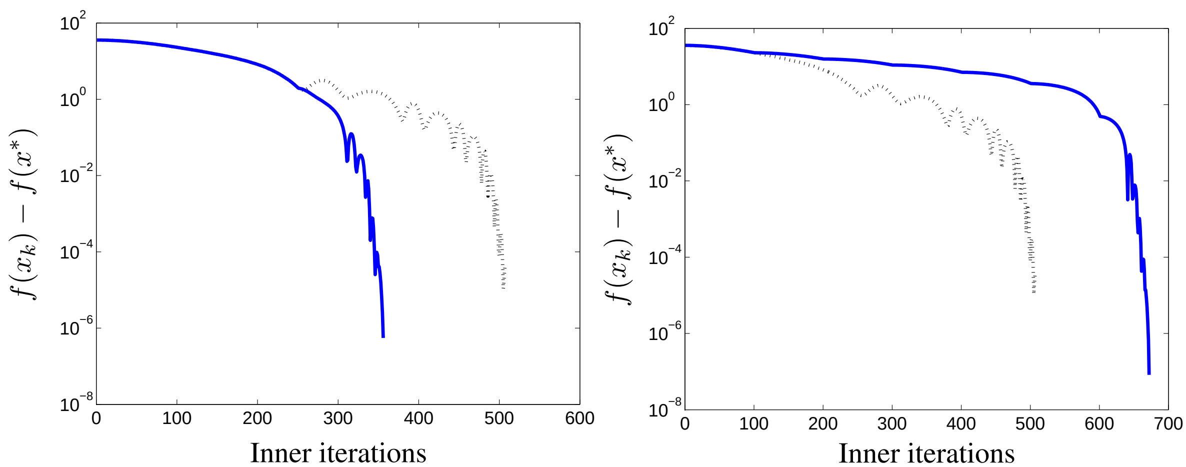

FIGURE 1. Restarted NESTA (solid blue line), versus NESTA (dotted black line). *Left:* Using 5 restart and 250 max inner iterations. *Right:* Using 10 restart and 100 max inner iterations.

We also check the result in Proposition 2.8 on toy problems with $p = 100$, $k = 5$ and increasing values of m, with $m = \{25, 30, 50\}$. As m grows, the matrix A satisfies the nullspace property (NSP) with diminishing values of C. The complexity bound in (9) improves as m increases and the recovery problem becomes less ill-posed, which is confirmed by the numerical experiments reported in Figure 2.

4.2. **Renegar's condition number and compressed sensing performance.** We first describe how we approximate the value of $\mathcal{C}_{x_0}(A)$, we then detail numerical experiments on synthetic data sets.

4.2.1. *Computing* $\mathcal{C}_{x_0}(A)$. The condition number $\mathcal{C}_{x_0}(A)$ appears here in upper bounds on computational complexities and statistical performances. In order to test numerically whether this quantity truly explains those features (as opposed to merely appearing in a wildly pessimistic bound), we explicitly compute it in numerical experiments.

We focus on the ℓ_1 norm. To compute $\mathcal{C}_{x_0}(A)$, we propose a heuristic which computes $\mu_{x_0}(A)$ in (15), the value of a nonconvex minimization problem over the cone of descent directions $\mathcal{T}(x_0)$. The closure of the latter is the polar of the cone generated by the subdifferential to the ℓ_1-norm ball at x_0 [Chandrasekaran et al., 2012, §2.3]. Let $S \subset \{1, \ldots, p\}$ denote the support of x_0, $\bar{S}$ denote its complement, and $|\bar{S}|$ denote the cardinality of $\bar{S}$. Then, with $s = \operatorname{sign}(x_0)$,

$$\mathcal{T}(x_0) = \operatorname{cone}\left\{z \in \mathbb{R}^p : z_S = s_S, z_{\bar{S}} \in [-1,1]^{|\bar{S}|})\right\}^\circ = \left\{z \in \mathbb{R}^p : \|z_{\bar{S}}\|_1 \leq -s_S^T z_S = -s^T z\right\}.$$

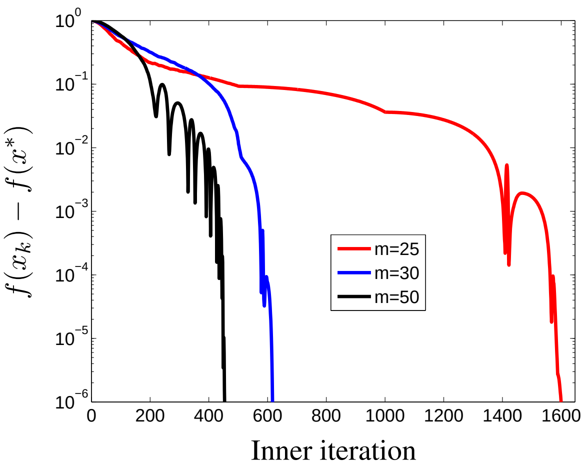

FIGURE 2. Restarted NESTA versus number of iterations for $m = \{25, 30, 50\}$.

Thus, $\mu_{x_0}(A)$ is the square root of

$$\min_{z \in \mathbb{R}^p} z^T A^T A z \quad \text{s.t.} \quad \|z\|_2 = 1 \quad \text{and} \quad \|z_{\bar{S}}\|_1 \leq -s^T z. \tag{19}$$

Let λ denote the largest eigenvalue of $A^T A$. If it were not for the cone constraint, solutions of this problem would be the dominant eigenvectors of $\lambda I - A^T A$, which suggests a *projected power method* [Deshpande et al., 2014] as follows. Given an initial guess $z_{(0)} \in \mathbb{R}^p$, $\|z_{(0)}\|_2 = 1$, iterate

$$\hat{z}_{(k+1)} = \text{Proj}_{x_0}\left((\lambda I - A^T A) z_{(k)}\right), \qquad z_{(k+1)} = \hat{z}_{(k+1)} / \|\hat{z}_{(k+1)}\|_2, \tag{20}$$

where we used the orthogonal projector to $\mathcal{T}(x_0)$,

$$\text{Proj}_{x_0}(\tilde{z}) = \arg\min_{z \in \mathbb{R}^p} \|z - \tilde{z}\|_2^2 \quad \text{s.t.} \quad \|z_{\bar{S}}\|_1 \leq -s^T z. \tag{21}$$

This convex, linearly constrained quadratic program is easily solved with CVX Grant et al. [2001]. As can be seen from KKT conditions, this iteration is a generalized power iteration Journée et al. [2008]

$$z_{(k+1)} \in \arg\max_{z \in \mathbb{R}^p} z^T (\lambda I - A^T A) z_{(k)} \quad \text{s.t.} \quad \|z\|_2 \leq 1 \quad \text{and} \quad \|z_{\bar{S}}\|_1 \leq -s^T z.$$

From the latter, it follows that $\|Az_{(k)}\|_2$ decreases monotonically with k. Indeed, owing to convexity of $f(z) = \frac{1}{2} z^T (\lambda I - A^T A) z$, we have $f(z) - f(z_{(k)}) \geq (z - z_{(k)})^T (\lambda I - A^T A) z_{(k)}$. The next iterate $z = z_{(k+1)}$ maximizes this lower bound on the improvement. Since $z = z_{(k)}$ is admissible, the improvement is nonnegative and $f(z_{(k)})$ increases monotonically.

Thus, the sequence $\|Az_{(k)}\|_2$ converges, but it may do so slowly, and the value it converges to may depend on the initial iterate $z_{(0)}$. On both accounts, it helps greatly to choose $z_{(0)}$ well. To obtain one, we modify (19) by smoothly penalizing the inequality constraint in the cost function, which results in a smooth optimization problem on the ℓ_2 sphere. Specifically, for small $\varepsilon_1, \varepsilon_2 > 0$, we use smooth proxies $h(x) = \sqrt{x^2 + \varepsilon_1^2} - \varepsilon_1 \approx |x|$ and $q(x) = \varepsilon_2 \log(1 + \exp(x/\varepsilon_2)) \approx \max(0, x)$. Then, with $\gamma > 0$ as Lagrange multiplier, we consider

$$\min_{\|z\|_2 = 1} \|Az\|_2^2 + \gamma \cdot q\left(s^T z + \sum\nolimits_{i \in \bar{S}} h(z_i)\right).$$

We solve the latter locally with Manopt [Boumal et al., 2014], itself with a uniformly random initial guess on the sphere, to obtain $z_{(0)}$. Then, we iterate the projected power method. The value $\|Az\|_2$ is an upper bound on $\mu_{x_0}(A)$, so that we obtain a lower bound on $\mathcal{C}_{x_0}(A)$. Empirically, this procedure, which is random only through the initial guess on the sphere, consistently returns the same value, up to five digits of accuracy, which suggests the proposed heuristic computes a good approximation of the condition number. Similarly positive results have been reported on other cones in Deshpande et al. [2014], where the special structure of the cone even made it possible to certify that this procedure indeed attains a global optimum in proposed experiments. Similarly, a generalized power method was recently shown to converge to global optimizers for the phase synchronization problem (in a certain noise regime) Boumal [2016]. This gives us confidence in the estimates produced here.

4.2.2. *Sparse recovery performance.* We conduct numerical experiments in the ℓ_1 case to illustrate the connection between the condition number $\mathcal{C}_{x_0}(A)$, the computational complexity of solving (1), and the statistical efficiency of the estimator (16). Importantly, throughout the experiments, the classical condition number of A will remain essentially constant, so that the main variations cannot be attributed to the latter.

We follow a standard setup, similar to some of the experiments in Donoho and Tsaig [2008]. Fixing the ambient dimension $p = 300$ and sparsity $k = \mathbf{Card}(x_0) = 15$, we let the number of linear measurements n vary from 1 to 150. For each value of n, we generate a random signal $x_0 \in \mathbb{R}^p$ (uniformly random support, i.i.d. Gaussian entries, unit ℓ_2-norm) and a random sensing matrix $A \in \mathbb{R}^{n \times p}$ with i.i.d. standard Gaussian entries. Furthermore, for a fixed value $\delta = 10^{-2}$, we generate a random noise vector $w \in \mathbb{R}^n$ with i.i.d. standard Gaussian entries, normalized such that $\|w\|_2 = \delta \|A\|_2$, and we let $y = Ax_0 + w$. This is repeated 100 times for each value of n.

For each triplet (A, x_0, y), we first solve the noisy problem (16) with the L1-Homotopy algorithm ($\tau = 10^{-7}$) Asif and Romberg [2014], and report the estimation error $\|x^* - x_0\|_2$. Then, we solve the noiseless problem (1) with L1-Homotopy and the TFOCS routine for basis pursuit ($\mu = 1$) Becker et al. [2011b]. Exact recovery is declared when the error is less than 10^{-5}, and we report the empirical probability of exact recovery, together with the number of iterations required by each of the solvers. The number of iterations of LARS Efron et al. [2004] is also reported, for comparison. For L1-Homotopy, we report the computation time, normalized by the computation time required for one least-squares solve in A, as in [Donoho and Tsaig, 2008, Fig. 3], which accounts for the growth in n. Finally, we compute the classical condition number of A, $\kappa(A)$, as well as (a lower bound on) the cone restricted condition number $\mathcal{C}_{x_0}(A)$, as per the previous section. As it is the computational bottleneck of the experiment, it is only computed for 20 of the 100 repetitions.

The results of Figure 3 show that the cone-restricted condition number explains both the computational complexity of (1) and the statistical complexity of (16): fewer samples mean bad conditioning which in turn implies high computational complexity. We caution that our estimate of $\mathcal{C}_{x_0}(A)$ is only a lower bound. Indeed, for small n, the third plot on the left shows that, even in the absence of noise, recovery of x_0 is not achieved by (16). Lemma 3.3 then requires $\mathcal{C}_{x_0}(A)$ to be infinite. But the computational complexity of solving (1) is visibly favorable for small n, where far from the phase transition, problem (P) is far from infeasibility, which is just as easy to verify as it is to certify that (P) is infeasible when n is comfortably larger than needed. This phenomenon is best explained using a symmetric version of the condition number Amelunxen and Lotz [2014] (omitted here to simplify computations).

We also solved problem (1) with interior point methods (IPM) via CVX. The number of iterations appeared mostly constant throughout the experiments, suggesting that the practical implementation of such solvers renders their complexity mostly data agnostic in the present setting. Likewise, the computation time required by L1-Homotopy on the noisy problem (16), normalized by the time of a least-squares solve, is mostly constant (at about 150). This hints that the link between computational complexity of (1) and (16) remains to be fully explained.

Appendix

The last condition in Definition 2.5 is arguably the least intuitive. Lemma 4.1 below connects it, in some cases, with the more intuitive notion of decomposable norm.

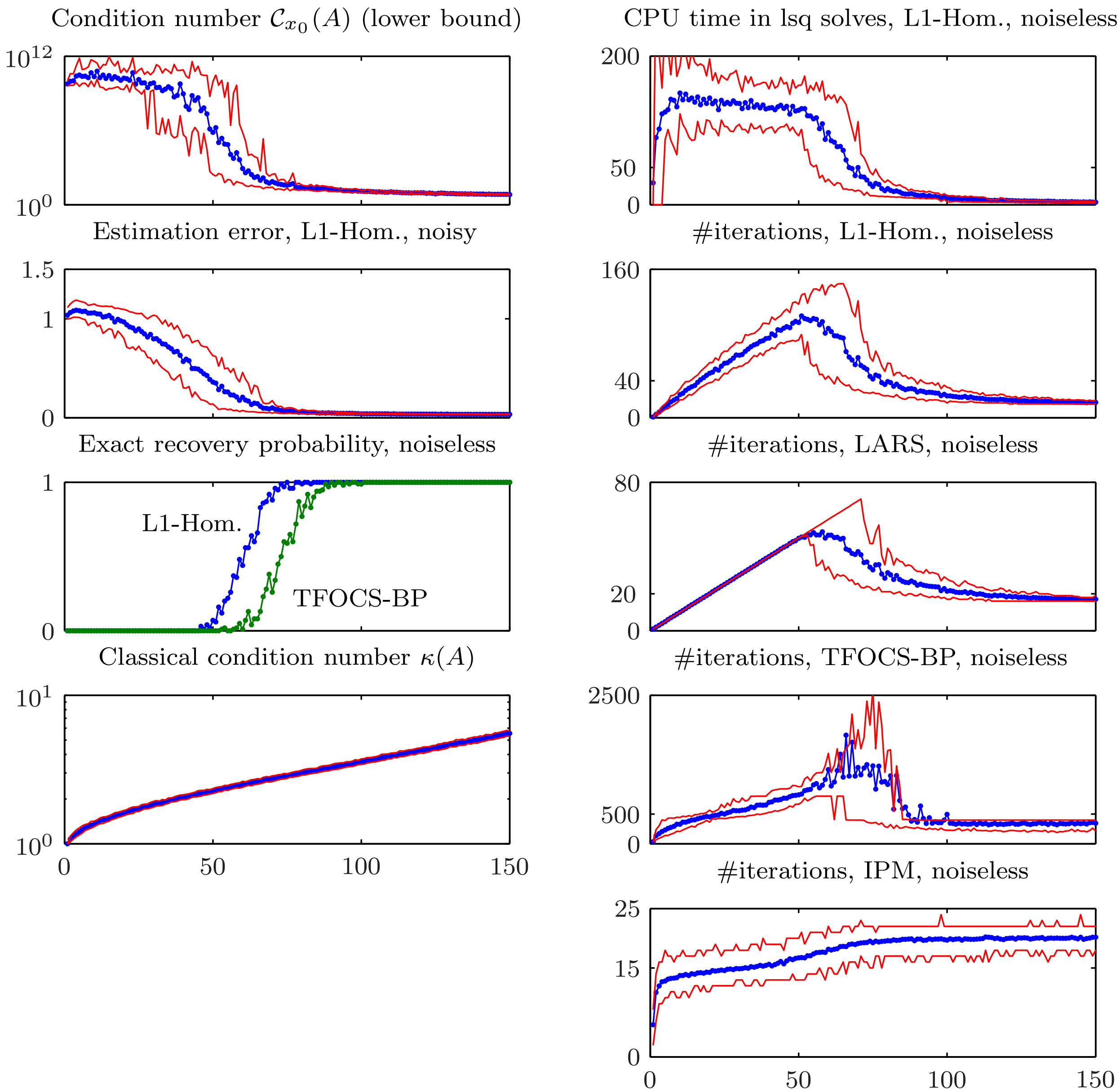

FIGURE 3. We plot the cone-restricted condition number of A (upper left), explaining both the computational complexity of problem (1) (right column) and the statistical complexity of problem (16) (second on the left). Central curves represent the mean (geometric mean in log-scale plots), red curves correspond to 10th and 90th percentile. We observe that high computing times (peaks in the right column) are directly aligned with instances where sparse recovery barely holds/fails (left), i.e. near the phase transition around $n = 70$, where the distance to feasibility for problem (P) also follows a phase transition.

Lemma 4.1. *Assume $B = \mathbf{I}$, condition iii) above, which reads*

$$\|P^* f + \bar{P}^* g\|_* \leq \max(\|f\|_*, \|g\|_*),$$

for any $f, g \in \mathcal{E}$, implies

$$\|u\| \geq \|Pu\| + \|\bar{P}u\|.$$

for any $u \in \mathcal{E}$.

Proof. We combine the conjugacy result for squared norm [Boyd and Vandenberghe, 2004, Example 3.27] showing that the conjugate of a squared norm $\|x\|^2/2$ is the squared conjugate norm $\|x\|_*^2/2$, with the result in [Rockafellar, 1970, Th. 16.3], to show

$$(\|P^* f + \bar{P}^* g\|_*^2/2)_* = \inf_u \{\|u\|^2/2 : Pu = y, \bar{P}u = z\}$$

Also, the dual of the norm $\max(\|f\|_*, \|g\|_*)$ is the norm $\|y\| + \|z\|$, hence taking the conjugate of condition iii) implies

$$\inf_u \{\|u\| : Pu = y, \bar{P}u = z\} \geq \|y\| + \|z\|$$

or again

$$\|u\| \geq \|Pu\| + \|\bar{P}u\|.$$

which is the desired result. ■

This means in particular that $\|u\| = \|Pu\| + \|\bar{P}u\|$ when $\bar{P} = I - P$, in which case sparsity systems match the *decomposable norms* setting in [Negahban et al., 2009]. The condition $\bar{P} = I - P$ holds for the ℓ_1 norm and some group sparsity problems detailed below, but not for the nuclear norm.

Acknowledgements. AA is at CNRS, at the Département d'Informatique at École Normale Supérieure, 2 rue Simone Iff, 75012 Paris, France. INRIA, Sierra project-team, PSL Research University. The authors would like to acknowledge support from a starting grant from the European Research Council (ERC project SIPA), an AMX fellowship, as well as support from the chaire *Économie des nouvelles données*, the *data science* joint research initiative with the *fonds AXA pour la recherche*.

References

A. Agarwal, S.N. Negahban, and M.J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Arxiv preprint arXiv:1104.4824*, 2011.

Dennis Amelunxen and Martin Lotz. Gordon's inequality and condition numbers in conic optimization. *arXiv preprint arXiv:1408.3016*, 2014.

Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference*, page iau005, 2014.

M.S. Asif and J. Romberg. Sparse recovery of streaming signals using ℓ_1-homotopy. *Signal Processing, IEEE Transactions on*, 62(16):4209–4223, 2014. doi: 10.1109/TSP.2014.2328981.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Stephen Becker, Jérôme Bobin, and Emmanuel J Candès. Nesta: a fast and accurate first-order method for sparse recovery. *SIAM Journal on Imaging Sciences*, 4(1):1–39, 2011a.

Stephen R Becker, Emmanuel J Candès, and Michael C Grant. Templates for convex cone problems with applications to sparse signal recovery. *Mathematical Programming Computation*, 3(3):165–218, 2011b.

Alexandre Belloni and Robert M Freund. A geometric analysis of renegar's condition number, and its interplay with conic curvature. *Mathematical programming*, 119(1):95–107, 2009.

Alexandre Belloni, Robert M Freund, and Santosh Vempala. An efficient rescaled perceptron algorithm for conic systems. *Mathematics of Operations Research*, 34(3):621–641, 2009.

P.J. Bickel, Y. Ritov, and A.B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

N. Boumal, B. Mishra, P.-A. Absil, and R. Sepulchre. Manopt, a Matlab toolbox for optimization on manifolds. *Journal of Machine Learning Research*, 15:1455–1459, 2014.

Nicolas Boumal. Nonconvex phase synchronization. *arXiv preprint arXiv:1601.06114*, 2016.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

V. Chandrasekaran, B. Recht, P. Parrilo, and A.S. Willsky. The convex geometry of linear inverse problems. *Foundations of Computational Mathematics*, 12(6):805–849, 2012.

Venkat Chandrasekaran and Michael I Jordan. Computational and statistical tradeoffs via convex relaxation. *Proceedings of the National Academy of Sciences*, 110(13):1181–1190, 2013.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the AMS*, 22 (1):211–231, 2009.

Y. Deshpande, A. Montanari, and E. Richard. Cone-constrained principal component analysis. In *NIPS 27*, pages 2717–2725. 2014.

D. L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. *Proc. of the National Academy of Sciences*, 102(27):9446–9451, 2005.

David L Donoho and Yaakov Tsaig. Fast solution of ℓ_1-norm minimization problems when the solution may be sparse-norm minimization problems when the solution may be sparse. *Information Theory, IEEE Transactions on*, 54(11): 4789–4812, 2008.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

Marina Epelman and Robert M Freund. Condition number complexity of an elementary algorithm for computing a reliable solution of a conic linear system. *Mathematical Programming*, 88(3):451–485, 2000.

Robert M Freund and Jorge R Vera. Some characterizations and properties of the "distance to ill-posedness" and the condition measure of a conic linear system. *Mathematical Programming*, 86(2):225–260, 1999a.

Robert M Freund and Jorge R Vera. Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm. *SIAM Journal on Optimization*, 10(1):155–176, 1999b.

Robert M Freund and Jorge R Vera. On the complexity of computing estimates of condition measures of a conic linear system. *Mathematics of Operations Research*, 28(4):625–648, 2003.

Pontus Giselsson and Stephen Boyd. Monotonicity and restart in fast gradient methods. In *53rd IEEE Conference on Decision and Control*, pages 5058–5063. IEEE, 2014.

M. Grant, S. Boyd, and Y. Ye. CVX: Matlab software for disciplined convex programming. 2001.

M. Journée, Y. Nesterov, P. Richtárik, and R. Sepulchre. Generalized power method for sparse principal component analysis. *arXiv:0811.4724*, 2008.

Anatoli Juditsky, Fatma Kılınç Karzan, and Arkadi Nemirovski. On a unified view of nullspace-type conditions for recoveries associated with general sparsity structures. *Linear Algebra and its Applications*, 441:124–151, 2014.

B.S. Kashin and V.N. Temlyakov. A remark on compressed sensing. *Mathematical notes*, 82(5):748–755, 2007.

Sahand Negahban, Bin Yu, Martin J Wainwright, and Pradeep K Ravikumar. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *Advances in Neural Information Processing Systems*, pages 1348–1356, 2009.

Y. Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

Y. Nesterov. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Brendan O'Donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

Fernando Ordóñez and Robert M Freund. Computational experience and the explanatory value of condition measures for linear optimization. *SIAM Journal on Optimization*, 14(2):307–333, 2003.

A. Pajor and N. Tomczak-Jaegermann. Subspaces of small codimension of finite-dimensional banach spaces. *Proceedings of the American Mathematical Society*, 97(4):637–642, 1986.

Javier Peña. Understanding the geometry of infeasible perturbations of a conic linear system. *SIAM Journal on Optimization*, 10(2):534–550, 2000.

James Renegar. Some perturbation theory for linear programming. *Mathematical Programming*, 65(1):73–91, 1994.

James Renegar. Linear programming, complexity theory and elementary functional analysis. *Mathematical Programming*, 70(1-3):279–351, 1995a.

James Renegar. Incorporating condition measures into the complexity theory of linear programming. *SIAM Journal on Optimization*, 5(3):506–524, 1995b.

James Renegar. *A mathematical view of interior-point methods in convex optimization*, volume 3. Siam, 2001.

R. T. Rockafellar. *Convex Analysis*. Princeton University Press., Princeton., 1970.

Vincent Roulet and Alexandre d'Aspremont. Sharpness, restart and acceleration. *ArXiv preprint arXiv:1702.03828*, 2017.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

Juan Carlos Vera, Juan Carlos Rivera, Javier Pena, and Yao Hui. A primal–dual symmetric relaxation for homogeneous conic systems. *Journal of Complexity*, 23(2):245–261, 2007.

Ian En-Hsu Yen, Cho-Jui Hsieh, Pradeep K Ravikumar, and Inderjit S Dhillon. Constant nullspace strong convexity and fast convergence of proximal methods under high-dimensional settings. In *Advances in Neural Information Processing Systems*, pages 1008–1016, 2014.

Zirui Zhou, Qi Zhang, and Anthony Man-Cho So. l1, p-norm regularization: Error bounds and convergence rate analysis of first-order methods. In *Proceedings of the 32nd International Conference on Machine Learning,(ICML)*, pages 1501–1510, 2015.

D.I., ÉCOLE NORMALE SUPÉRIEURE, PARIS, FRANCE.
E-mail address: `vincent.roulet@inria.fr`

MATHEMATICS DEPARTMENT,
PRINCETON UNIVERSITY, PRINCETON NJ 08544, USA.
E-mail address: `nboumal@math.princeton.edu`

CNRS & D.I., UMR 8548,
ÉCOLE NORMALE SUPÉRIEURE, PARIS, FRANCE.
E-mail address: `aspremon@di.ens.fr`